BOLYAI-KÖNYVEK SOROZAT

A csaknem 40 éve indult, igen sikeres Bolyai-könyvek példatár sorozat kötetét tartja kezében az Olvasó.

A sorozat könyveiben a szerzők a középiskolai tanulóknak, továbbá főiskolai és egyetemi hallgatóknak adnak szerencsésen választott példákat, kidolgozott feladatokat.

E könyv első részében a szerző tömören ismerteti a valószínűségszámítási alapokat, majd a második részben a matematikai statisztikát, ezen belül a statisztikai minta jellemzőit, a statisztikai becslések problémáját és azok megoldásait, a statisztikai próbákat, az illeszkedés-, homogenitás- és függetlenségvizsgálat, valamint a korreláció, és regresszióelemzés módszereit.

Befejezésként a matematikai statisztika gyakorlati alkalmazásával kapcsolatos példák kaptak helyet.

ISBN 963-16-3036-6

9 789631 630367

MŰSZAKI KÖNYVKIADÓ

MATEMATIKAI STATISZTIKA

BOLYAI-KÖNYVEK

LUKÁCS OTTÓ

MATEMATIKAI STATISZTIKA

$$f(x,y) = F''_{xy}(x,y)$$

LUKÁCS OTTÓ

MATEMATIKAI STATISZTIKA

A BOLYAI-SOROZAT KÖTETEI:

Bárczy Barnabás: Differenciálszámítás
Solt György: Valószinűségszámítás
Lukács Ottó: Matematikai statisztika
Scharnitzky Viktor: Differenciálegyenletek
Bárczy Barnabás: Integrálszámítás
Scharnitzky Viktor: Mátrixszámítás
Urbán János: Matematikai logika
Urbán János: Határérték-számítás
Fekete Zoltán–Zalay Miklós: Többváltozós függvények analízise

LUKÁCS OTTÓ

MATEMATIKAI STATISZTIKA

PÉLDATÁR

4. változatlan utánnyomás

MŰSZAKI KÖNYVKIADÓ, BUDAPEST

Lektorálta:

MESZÉNA GYÖRGY
matematikus

ISBN 963 10 6848 X (első kiadás)
ISBN 963 16 3036 6
ISSN 1216-5344

TARTALOM

ELŐSZÓ

A matematikai statisztika egyike a matematika azon ágainak, amelyet a gyakorlatban nagyon sok területen alkalmaznak. Használják a műszaki tudományokban, a minőségellenőrzésben, az orvostudományban, a mezőgazdaságban és még számos helyen.

Elterjedtségének oka kettős: 1. A világ jelenségei igen gyakran nem determinisztikusak, hanem *sztochasztikusok (véletlentől függőek)*. Pl. egy árucikknek a hosszmérete, élettartama véletlentől függő változó. 2. Gyakori eset, hogy pl. egy árucikkel kapcsolatban nem ismerjük az egész sokaságot, csak mintát veszünk belőle. A minta paraméterei alapján adunk *becslést* az egész sokaság megfelelő paramétereire (ez a matematikai statisztika egyik feladata).

A matematikai statisztika a valószínűségszámítással együtt fejlődött ki. Mindkettőről azt szokták mondani, hogy a „véletlen tudománya". Ezért könyvünket is valószínűségszámítási alapokkal kezdjük.

A könyv *felépítése* a sorozat felépítéséhez hasonlít: az elméleti alapok összefoglalása után gyakorló feladatok következnek. Aki a matematikai statisztikában otthon akar lenni, annak háromtípusú feladattal érdemes ismerkednie:

- a statisztika elméleti problémáival;
- a gyakorlati feladatokkal;
- a számítógépes megoldás lehetőségeivel.

Könyvünkben mindhárom problématípust megtalálja az Olvasó. A *számítógépi programokhoz* folyamatábrát készítettünk, az elsőknél még bővebb magyarázattal (hogy az e téren kezdő Olvasó is be tudjon kapcsolódni). A programok BASIC nyelvűek, és COMMODORE személyi számítógépen voltak futtatva. Mindegyiknél található próbafeladat és futtatási eredmény is. Ezenkívül közlünk néhány HT–PTK–1050 zsebszámológépre írt programot is (ezek is kipróbált programok). Megjegyezzük, hogy a gép a TEXAS INSTRUMENTS 57 zsebszámológép gyártási licence alapján készült, a programok erre a gépre is jók. Röviden ismertetjük az IBM 370/145 gépen futtatható, STEPWISE MULTIPLE REGRESSION programcsomagot is, valamint néhány, ezzel készült futtatást.

Nem haszontalan időtöltés, ha valaki a matematikai statisztikával ismerkedik. Tisztában lesz azzal, hogy milyen következtetéseket, becsléseket tudunk tenni a minta alapján az egész sokaságra. A matematikai statisztikai próbákkal a minőségellenőrzéshez kap megfelelő alapokat. A korreláció (regresszióanalízis) tulajdonképpen a mérési eredmények egyfajta értékelésére ad lehetőséget. Még tovább is sorolhatnánk, hogy mennyi területe van a matematikai statisztikának.

A numerikus jellegű gyakorló feladatok a *legkülönbözőbb iparágak területéről származnak* (könnyűipar, gépészet, elektromos szakterületek, vegyészet, építészet, hidrológia stb.). Érdemes ezeket tanulmányozni, hátha az Olvasó indíttatást kap a saját munkaterületében való alkalmazásra is.

A könyvet várhatóan haszonnal forgathatják mérnökök, közgazdászok, matematikusok. Hasznosítani tudják egyetemi, főiskolai hallgatók és oktatók, a műszaki területekről és a tudományegyetemekről. A számítástechnikai részek érdekesek lehetnek programozók, számítástechnikai szakemberek részére.

Nem könnyű és nem kis anyag ez, amibe az Olvasó belefog. Munkájának megkönnyítése érdekében a megírás során törekedtem a lehető legnagyobb egyszerűségre, a legvilágosabb fogalmazásra, az eredmények szemléletes megvilágítására.

Köszönetet mondok értékes megjegyzéseiért lektoromnak, Meszéna Györgynek, a kézirat kimunkálásában való közreműködéséért és a gondos gépelésért Váry Péternének, valamint a kézirat átnézéséért Román Ferencnének.

Lukács Ottó

I. VALÓSZÍNŰSÉG-SZÁMÍTÁS

Ebben a részben csak összefoglaljuk a szükséges tudnivalókat, képleteket. A feladatok iránt érdeklődő Olvasónak az irodalomjegyzék megfelelő könyveit ajánljuk.

1. KOMBINATORIKA

a) n különböző elem különböző sorrendjének *(az n elem permutációinak) a száma:*

$$P_n = 1 \cdot 2 \cdot 3 \cdots n = n!$$

Ha az n elem között k_1, k_2, ... k_l darab egyező van, az *n elem ismétléses permutációinak a száma:*

$$P_n^{k_1, k_2, \dots k_l} = \frac{n!}{k_1!\, k_2! \dots k_l!}.$$

b) n különböző elemből k különbözőt kiválasztunk ($k \leqq n$) és minden lehetséges sorrendbe állítjuk. Az így keletkező *variációk száma:*

$$V_{n,k} = n(n-1)\dots(n-k+1).$$

Ha megengedünk a kiválasztásnál ismétlődést is, az *ismétléses variációk* száma:

$$V_{n,k}^{\text{ism}} = n^k.$$

c) Ha az n különböző elemből k-t kiválasztunk ($k \leqq n$), de a kiválasztottakat nem rakjuk különböző sorrendbe, a keletkező *kombinációk száma:*

$$C_{n,k} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{1\cdot 2\cdot 3\cdots k} = \binom{n}{k}.$$

Megegyezés szerint $\binom{n}{0} = 1$. Ha n-ből k-t kiválasztunk, de a kiválasztott elem megismétlődhet, a kapható *ismétléses kombinációk száma:*

$$C_{n,k}^{\text{ism}} = \binom{n+k-1}{k}.$$

2. ESEMÉNYALGEBRA

Egy véletlen jelenségre vonatkozó kísérlet lehetséges kimeneteleit *elemi eseményeknek* nevezzük. Az elemi események halmaza az *I eseménytér. Véletlen esemény* (röviden: esemény) az eseménytér egy részhalmaza. Az *eseményalgebra* tehát egyenértékű egy *halmazalgebrával.* Az események *jelölése* általában nagybetűkkel történik: pl. A esemény; B_1, B_2, B_3 események stb.

A maga után vonja B-t, ha az A-nak megfelelő halmaz részhalmaza B-nek. Jelölése: $A \subset B$.

Az összes elemi eseményt tartalmazó halmaznak az *I biztos esemény* felel meg.

Egy A esemény *ellentéte* az az $\bar{A}$ esemény, amely akkor és csak akkor következett be, ha A nem következett be. A biztos esemény ellentéte az *O lehetetlen esemény.*

$A+B$ esemény (más jelöléssel $A \cup B$, A uniója, egyesítése B-vel) jelenti *vagy az A, vagy a B, vagy mindkettő bekövetkezését* (legalább az egyik bekövetkezését).

$A \cdot B = AB$ (halmazelméleti jelöléssel $A \cap B$, A metszete B-vel) jelenti *mind az A, mind a B bekövetkezését* (a két esemény együttes bekövetkezését).

Az *elemi eseményeknek* az eseménytér egyelemű részhalmazai felelnek meg, az *összetett eseményeknek* a többelemű részhalmazok.

Az A_1, A_2, A_3, ..., A_n *teljes eseményrendszert alkotnak,* ha
a) egyikük biztosan bekövetkezik:

$$\sum_{i=1}^{n} A_i = I \text{ és ha}$$

b) egymást páronként kizárják:

$$A_iA_j = O \qquad (i \neq j).$$

Az eseménytér összes lehetséges elemi eseménye teljes eseményrendszert alkot.

Az A és B *események különbsége:*

$$A - B = A\bar{B}.$$

Az eseményalgebrában, ill. a vele egyenértékű halmazalgebrában

- két művelet van: összeadás és szorzás;
- minden eseménynek van ellentéte;
- egységelemnek az I-t, nullelemnek az O-t tekinthetjük.

Az így definiált algebrára érvényesek a *Boole-algebra* következő tulajdonságai:

Összeg	Szorzat	Tulajdonság
$A+B = B+A$	$AB = BA$	kommutativitás
$(A+B)+C = A+(B+C)$	$(AB)C = A(BC)$	asszociativitás
	$A(B+C) = AB+AC$	disztributivitás I.
	$A+BC = (A+B)(A+C)$	disztributivitás II.
$A+A = A$	$AA = A$	idempotencia
$A+\bar{A} = I$	$A\bar{A} = O$	
$A+I = I$	$AI = A$	
$A+O = A$	$AO = O$	

3. KLASSZIKUS VALÓSZÍNŰSÉG

Egy A esemény *relatív gyakorisága* az A esemény bekövetkezéseinek száma, arányítva az összes kísérlet számához $\left(\frac{k_A}{n}\right)$. Ha egyre több kísérletet végzünk, a relatív gyakoriság egy számérték körül ingadozik, ez a számérték az A esemény *valószínűsége*, $P(A)$ (a probability = valószínűség angol szó kezdőbetűjéből). $P(A)$ tehát azt mutatja, hogy A esemény az összes kísérletnek *várhatóan* (átlagosan) *hányad részében* következik be. A relatív gyakoriságra érvényes összefüggések alapján állította fel *Kolmogorov a $P(A)$ valószínűségre érvényes három alapkövetelményt (axiómát):*

I. $0 \leqq P(A) \leqq 1$;

II. $P(I) = 1$;

III. ha $AB = \emptyset$, akkor $P(A+B) = P(A) + P(B)$.

A III. axióma általánosítható véges (vagy megszámlálhatóan végtelen sok), egymást páronként kizáró A_i eseményekre:

III. *a)* Ha $A_i A_j = O$ $(i \neq j)$, akkor

$$P(\sum_i A_i) = \sum_i P(A_i).$$

Az axiómákból levezethetők a valószínűségszámítás tételei. Pl.

a) $P(\emptyset) = 0$ (a lehetetlen esemény valószínűsége 0);

b) $P(\bar{A}) = 1 - P(A)$ (egy esemény ellentétének a valószínűsége = 1 – az esemény valószínűsége);

c) ha $A \subset B$, akkor $P(A) \leqq P(B)$;

d) az „A és B közül legalább az egyik” esemény valószínűsége (ha nem kötjük ki, hogy A és B kizárják egymást):

$$P(A+B) = P(A) + P(B) - P(AB).$$

Ennek általánosítása 3 eseményre:

$$P(A+B+C) = P(A) + P(B) + P(C) - \\ - P(AB) - P(AC) - P(BC) + P(ABC).$$

De általánosítható n tetszőleges eseményre is *(Poincaré tétele):*

$$P(A_1 + A_2 + \ldots + A_n) = \sum_{1 \leqq i} P(A_i) - \sum_{1 \leqq i_1 < i_2} P(A_{i_1} A_{i_2}) + \\ + \sum_{1 \leqq i_1 < i_2 < i_3} P(A_{i_1} A_{i_2} A_{i_3}) \pm \ldots + \\ + (-1)^{n+1} P(A_1 A_2 \ldots A_n).$$

A több évszázados *klasszikus valószínűségszámítás* kezdetben *olyan eseményterekkel foglalkozott, amelyeknél az $A_1, A_2, \ldots, A_n$ elemi események egyenlő valószínűségűek:*

$$P(A_i) = \frac{1}{n} \quad (i = 1, 2, 3, \ldots, n).$$

Bizonyítható az axiómákból, hogy ilyen esetben a k-féleképpen bekövetkezhető A esemény valószínűsége:

$$P(A) = \frac{k}{n} = \frac{\text{kedvező esetek száma}}{\text{összes lehetséges esetek száma}}$$

(klasszikus valószínűségszámítási képlet).

A lehetetlen esemény valószínűsége 0, ez fordítva nem áll fenn: a 0 valószínűségű esemény nem biztos, hogy lehetetlen (csak nagyon ritkán következik be).

A biztos esemény valószínűsége 1. Fordítva itt sem igaz: az 1 valószínűségű esemény nem feltétlenül biztos esemény (csak „majdnem mindig” bekövetkezik).

4. FELTÉTELES VALÓSZÍNŰSÉG. FÜGGETLENSÉG

Ha egy A esemény valószínűségét nem a teljes I eseménytéren, hanem csak azok közül az esetek közül vizsgáljuk, amikor B is bekövetkezett, az így kapott $P(A/B)$ *feltételes valószínűség:*

$$P(A/B) = \frac{P(AB)}{P(B)}$$

(feltéve, hogy $P(B) \neq 0$). Ez a szám azt mutatja, hogy A várhatóan hányad részében következik be azon esetek közül, amelyekben B is bekövetkezett. Az előző képletből:

$$P(AB) = P(A/B)P(B).$$

Ez általánosítható n esemény szorzatára:

$$P(A_1A_2\ldots A_n) = P(A_n/A_1A_2\ldots A_{n-1})P(A_{n-1}/A_1A_2\ldots A_{n-2})\ldots$$
$$\ldots P(A_2/A_1)P(A_1).$$

Ha A és B *függetlenek*, akkor

$$P(AB) = P(A)P(B).$$

Az $A_1, A_2, \ldots, A_n$ események *teljesen függetlenek*, ha belőlük tetszőleges kettőt, hármat, ..., n-et kiválasztva, a kiválasztott események mind függetlenek.

5. A TELJES VALÓSZÍNŰSÉG TÉTELE. A BAYES-TÉTEL

Az A esemény valószínűségét ismerjük a $B_1, B_2, \ldots, B_n$ feltételek mellett, ahol $B_1, B_2, \ldots, B_n$ teljes eseményrendszer. Ekkor A teljes valószínűsége így számítható:

$$P(A) = \sum_{i=1}^{n} P(A/B_i)P(B_i).$$

Itt a $P(B_i)$ értékek nem lehetnek nullák. Ez a *teljes valószínűség tétele.*

A *Bayes-tétel* a $P(B_i/A)$ feltételes valószínűségek kiszámítására ad módot, ha ismertek a $P(A/B_i)$ és $P(B_i) \neq 0$ valószínűségek:

$$P(B_i/A) = \frac{P(A/B_i)P(B_i)}{\sum_{i=1}^{n} P(A/B_i)P(B_i)}.$$

Itt $B_1, B_2, \ldots, B_n$ teljes eseményrendszer.

6. A VALÓSZÍNŰSÉGI VÁLTOZÓ ÉS JELLEMZŐI

Ha az eseménytér elemeihez egy-egy számértéket rendelünk, az így kapott véletlentől (véletlen elemi eseményektől) függő változót *valószínűségi változónak* (véletlen változónak, sztochasztikus változónak) nevezzük. A könyvben többnyire latin nagybetűkkel ($X, Y, \ldots$) jelöljük őket, szokásos még a görög betűkkel ($\xi, \eta, \ldots$) való jelölés is.

Ha X felvett értékei a számegyenes mentén diszkrét értékek (véges vagy megszámlálhatóan végtelen halmazt alkotnak), akkor a valószínűségi változó *diszkrét.* Az egyes értékek felvételi valószínűségei a

$$p_k = P(X = x_k) \quad (k = 1, 2, 3, \ldots)$$

értékek adják az X *eloszlását:*

$$\sum_k p_k = 1.$$

Azt a számot, amely körül X megfigyelt értékeinek átlaga ingadozik, az X *várható értékének* nevezzük. Jele: $M(X)$ (az angol mean = középérték szóból), vagy újabban $E(X)$ (az expected value = várható érték szavakból).

Diszkrét valószínűségi változó várható értéke:

$$M(X) = \sum_i x_i p_i.$$

Ha itt i végtelenig megy, akkor fel kell tenni, hogy $\sum_{i=1}^{\infty} |x_i| p_i$ konvergens, különben a sor átrendezésével a várható értékre más számot kaphatnánk.

Az X véletlen változó várható értékétől való eltérését (körötte való szóródását) jellemzi a *szórás*, $D(X)$. A *szórásnégyzet*

$$D^2(X) = M\{[X-M(X)]^2\}.$$

Diszkrét véletlen változó szórásnégyzete:

$$D^2(X) = \sum_i (x_i - M)^2 p_i.$$

Az olyan valószínűségi változót, amelynek értékei a számegyenes egy teljes intervallumát (általános esetben a teljes számegyenest) kitöltik, *folytonos valószínűségi változónak* nevezzük.

Mind diszkrét, mind folytonos eloszlásnál megadható egy olyan $F(x)$ *eloszlásfüggvény*, amelynek x helyen felvett értéke megadja, hogy hányad része esik a változónak x határ alá:

$$P(X<x) = F(x).$$

Diszkrét esetben:

$$F(x) = \sum_{x_i<x} p_i.$$

Ez egy x tengellyel párhuzamos szakaszokból álló lépcsős függvény, amelynek minden felvehető x_i értéknél p_i ugrása van (1. ábra).

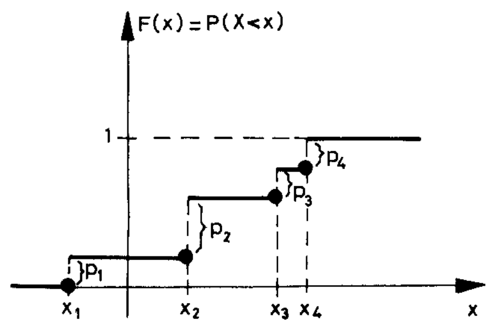
A véletlen változó a és b közé esésének valószínűsége:

$$P(a \leqq X < b) = F(b) - F(a),$$

az a „fölé" esés valószínűsége pedig:

$$P(X \geqq a) = 1 - F(a).$$

Az eloszlásfüggvény határértéke mínusz végtelenben 0, plusz végtelenben 1.



1. ábra

Folytonos eloszlású változó $f(x)$ sűrűségfüggvénye a $F(x)$ eloszlásfüggvény deriváltja:

$$\frac{dF(x)}{dx} = f(x).$$

Ebből következik, hogy

$$P(X<x) = F(x) = \int_{-\infty}^{x} f(u)\,du,$$

$$P(a \leqq X < b) = F(b) - F(a) = \int_a^b f(x)\,dx,$$

$$P(X \geqq a) = 1 - F(a) = \int_a^{\infty} f(x)\,dx,$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Egy h hosszúságú intervallumon a sűrűségfüggvény közelítő értéke:

$$f(x) \approx \frac{P(x \leqq X < x+h)}{h},$$

vagyis a sűrűségfüggvény az egységnyi intervallumhosszra eső valószínűséget (a valószínűség-sűrűséget) jellemzi.

Folytonos valószínűségi változó várható értéke:

$$M(X) = \int_{-\infty}^{\infty} xf(x)\,dx,$$

szórásnégyzete:

$$D^2(X) = \int_{-\infty}^{\infty} (x-M)^2 f(x)\,dx.$$

Igazolható, hogy

$$D^2(X) = M(X^2) - M^3(X),$$

vagyis diszkrét eloszlásnál:

$$D^2(X) = \sum_i x_i^2 p_i - \left(\sum_i x_i p_i\right)^2,$$

folytonos eloszlásnál:

$$D^2(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \left(\int_{-\infty}^{\infty} xf(x)\,dx\right)^2.$$

Az X változóból képezett $aX+b$ véletlen változónál (ahol a és b állandók):

$$M(aX+b) = aM(X)+b,$$
$$D^2(aX+b) = a^2 D^2(X).$$

Ha X nemnegatív valószínűségi változó, és $a>0$ konstans, akkor

$$P(X \geqq a) \leqq \frac{M(X)}{a}$$

(Markov-egyenlőtlenség). A *Csebisev-egyenlőtlenség* a valószínűségi változó várható érték körüli szóródására ad felvilágosítást:

$$P(|X-M(X)| \geqq kD(X)) \leqq \frac{1}{k^2},$$

vagy $kD(x)=a$ helyettesítéssel:

$$P(|X-M(X)| \geqq a) \leqq \frac{D^2(X)}{a^2}.$$

Momentumok:

- k-adik momentum: $M(X^k)$;
- k-adik centrális momentum: $M[(X-M(X))^k]$;
- k-adik abszolút momentum: $M(|X|^k)$,
- k-adik centrális abszolút momentum: $M(|X-M(X)|^k)$.

A *medián* az a szám, amely alá és fölé 50–50% valószínűséggel esik a véletlen változó. Ha az

$$F(x) = \frac{1}{2}$$

egyenletnek egy megoldása van, ez a medián; ha a megoldások egy teljes intervallumot kitöltenek, az intervallum középpontjának abszcisszája adja meg a mediánt. Ha az egyenletnek nincs megoldása, akkor a medián az az abszcisszaérték, ahol a függvény átugorja az $\frac{1}{2}$-et.

X p-edrendű *kvantilise* az a szám, amely alá p, fölé $1-p$ valószínűséggel esik a változó. Itt is az előbbi három eset fordul elő.

Folytonos eloszlásnál minden olyan abszcissza, ahol a sűrűségfüggvénynek helyi maximuma van, a *módusz*. Diszkrét eloszlásnál a módusz azon x_i érték, ahol a p_i valószínűség a környezetében levő valószínűségekhez képest a legnagyobb.

$$\textit{Ferdeségi együttható} = \frac{M[(X-M(X))^3]}{D^3(X)},$$

$$\textit{lapultsági együttható} = \frac{M[(X-M(X))^4]}{D^4(X)} - 3.$$

7. TÖBBDIMENZIÓS ELOSZLÁSOK. A KORRELÁCIÓ

Csak kétdimenziós eloszlásokat tárgyalunk. A mondottak kettőnél több változóra is kiterjeszthetők.

1. Az (X, Y) kétdimenziós valószínűségi változót *diszkrét valószínűségi vektorváltozónak nevezzük,* ha az általa felvehető (x_i, y_k) számpárok (az XY síkon levő pontok) véges vagy megszámlálhatóan végtelen halmazt alkotnak. Eloszlásukat a

$$p_{ik} = P(X=x_i,\ Y=y_k)$$

számok összessége (a kétdimenziós véletlen változó *együttes valószínűség-eloszlása*) jellemzi. Itt $i = 1, 2, 3, \ldots$, $k = 1, 2, 3, \ldots$. Természetesen fennáll, hogy

$$\sum_i \sum_k p_{ik} = 1.$$

Vizsgálhatjuk, hogy *önmagában X vagy önmagában Y eloszlása* milyen. A

$$P(X=x_i) = \sum_k p_{ik}$$

valószínűségek *X perem-* vagy *vetületeloszlását* adják meg, a

$$P(Y=y_k) = \sum_i p_{ik}$$

valószínűségek pedig az Y perem- (vetület-) eloszlását.

A két véletlen változó *együttes eloszlásfüggvénye:*

$$F(x, y) = P(X<x,\ Y<y) = \sum_{\substack{x_i<x \\ y_k<y}} p_{ik}.$$

A

$$P(X<x) = F_1(x) = P(X<x,\ Y<\infty) = F(x, \infty) = \sum_{x_i<x} P(X=x_i)$$

függvény az *X peremeloszlás-függvénye.* Ugyanígy Y peremeloszlás-függvénye:

$$F_2(y) = P(Y<y) = P(X<\infty,\ Y<y) = F(\infty, y) = \sum_{y_k<y} P(Y=y_k).$$

Mennyi annak a valószínűsége, hogy $X<x$, ha csak azon eseteket vizsgáljuk, amelyekben a második változóra valamilyen feltételt szabunk? *Négy különböző feltételes eloszlásfüggvényt* definiálhatunk:

$$F(x/Y=y_k) = P(X<x/Y=y_k) = \frac{P(X<x,\ Y=y_k)}{P(Y=y_k)},$$

$$F(x/Y<y) = P(X<x/Y<y) = \frac{P(X<x,\ Y<y)}{P(Y<y)} = \frac{F(x, y)}{F_2(y)},$$

$$F(x/y_1 \leqq Y<y_2) = P(X<x/y_1 \leqq Y<y_2) = \frac{P(X<x,\ y_1 \leqq Y<y_2)}{P(y_1 \leqq Y<y_2)},$$

$$F(x/Y \geqq y) = P(X<x/Y \geqq y) = \frac{P(X<x,\ Y \geqq y)}{P(Y \geqq y)}.$$

Y-ra hasonló feltételes eloszlásfüggvények adhatók meg, ha a feltétel X-re vonatkozik.

2. Ha az (X, Y) vektorváltozó által meghatározott pontok egy folytonos, síkbeli T tartományt alkotnak, akkor a *vektorváltozót folytonosnak* nevezzük. Az *együttes eloszlásfüggvény:*

$$F(x, y) = P(X<x,\ Y<y).$$

Annak a valószínűsége, hogy (X, Y) az

$$a_1 \leqq X < a_2, \quad b_1 \leqq Y < b_2$$

téglalapra esik, kifejezhető az együttes eloszlásfüggvénynek a téglalap csúcsaiban felvett négy értékével:

$$P(a_1 \leqq X < a_2, b_1 \leqq Y < b_2) = F(a_2, b_2) - F(a_1, b_2) - \\ - F(a_2, b_1) + F(a_1, b_1).$$

Ez folytonos és diszkrét véletlen vektorváltozóra egyaránt érvényes.

A sűrűségfüggvényt csak folytonos esetben értelmezzük. Az (X, Y) *együttes sűrűségfüggvénye:*

$$f(x, y) = F''_{xy}(x, y).$$

Ha h és k elég kicsi, akkor

$$f(x, y) \approx \frac{P(x \leqq X < x+h, y \leqq Y < y+k)}{hk},$$

vagyis a sűrűségfüggvény jó közelítésben a területegységre eső valószínűség (ún. valószínűség-sűrűség).

A kétszeres deriválás megfordításával:

$$P(X < x, Y < y) = F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v)\, du\, dv.$$

Ennek alapján

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1,$$

annak a valószínűsége pedig, hogy a vektorváltozó az XY sík egy T tartományába esik,

$$P((X, Y) \in T) = \iint_{(T)} f(x, y)\, dx\, dy.$$

3. Ha ismert X, Y együttes eloszlása, akkor az együttes eloszlásfüggvényből vagy sűrűségfüggvényből meghatározható külön-külön X, ill. Y eloszlása (az ún. *perem- vagy vetületi eloszlásfüggvények*):

$$F_1(x) = P(X < x) = F(x, \infty) = \int_{-\infty}^{\infty} \int_{-\infty}^{x} f(u, v)\, du\, dv,$$

$$F_2(y) = P(Y < y) = F(\infty, y) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f(u, v)\, du\, dv.$$

x, ill. y szerinti deriválással adódnak a *peremsűrűség-függvények:*

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy,$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

4. Az Y-ra tett különböző feltételek mellett X különböző *feltételes eloszlásfüggvényei:*

$$F(x/Y<y) = P(X<x/Y<y) = \frac{P(X<x,\ Y<y)}{P(Y<y)} = \frac{F(x, y)}{F_2(y)},$$

$$F(x/y_1 \leqq Y < y_2) = P(X<x/y_1 \leqq Y < y_2) = \\ = \frac{P(X<x,\ y_1 \leqq Y < y_2)}{P(y_1 \leqq Y < y_2)} = \frac{F(x, y_2) - F(x, y_1)}{F_2(y_2) - F_2(y_1)},$$

$$F(x/Y \geqq y) = P(X<x/Y \geqq y) = \frac{P(X<x,\ Y \geqq y)}{P(Y \geqq y)} = \\ = \frac{F_1(x) - F(x, y)}{1 - F_2(y)},$$

$$F(x/y) = F(x/Y=y) = \lim_{k \to 0} F(x/y \leqq Y < y+k) = \\ = \lim \frac{\dfrac{F(x, y+k) - F(x, y)}{k}}{\dfrac{F_2(y+k) - F_2(y)}{k}} = \frac{F'_y(x, y)}{f_2(y)}.$$

Ennek x szerinti parciális deriváltja adja meg a feltételes sűrűségfüggvényt:

$$f(x/y) = \frac{f(x, y)}{f_2(y)}.$$

Ez a képlet a

$$P(A/B) = \frac{P(AB)}{P(B)}$$

feltételes valószínűség kiterjesztése eseményekről folytonos változókra. Az $f(x, y) = f(x/y)f_2(y)$ mindkét oldalának y szerinti integrálásával:

$$f_1(x) = \int_{-\infty}^{\infty} f(x/y)f_2(y)\, dy,$$

ez a $P(A) = \sum_i P(A/B_i)P(B_i)$ *teljes valószínűség tételének az általánosítása* folytonos változókra. Hasonló képletek kaphatók $F(y/x)$ és $f(y/x)$-re. Ezekből

$$f(x/y) = \frac{f(y/x)f_1(x)}{\int_{-\infty}^{\infty} f(y/x)f_1(x)\, dx},$$

ez pedig a

$$P(B_i/A) = \frac{P(A/B_i)P(B_i)}{\sum_i P(A/B_i)P(B_i)}.$$

Bayes-tétel általánosítása eseményekről folytonos változókra.

5. X és Y valószínűségi változók *függetlenek*, ha

$$P(X<x,\ Y<y) = F(x, y) = P(X<x)P(Y<y) = F_1(x)F_2(y).$$

Ekkor egyúttal

$$P(a \leqq x < b,\ c \leqq Y < d) = P(a \leqq X < b)P(c \leqq Y < d).$$

Diszkrét vektorváltozónál:

$$P(X = x_i,\ Y = y_k) = P(X = x_i)P(Y = y_k).$$

Folytonos esetben: X és Y akkor és csak akkor függetlenek, ha sűrűségfüggvényeikre:

$$f(x, y) = f_1(x)f_2(y).$$

6. Két diszkrét, véletlen változó összegének, különbségének, szorzatának és hányadosának eloszlása:

Z	$P(Z = z_j)$ $(j = 1, 2, 3, \ldots)$
$X+Y$	$\sum_{x_i + y_k = z_j} P(X = x_i,\ Y = y_k)$
$X-Y$	$\sum_{x_i - y_k = z_j} P(X = x_i,\ Y = y_k)$
XY	$\sum_{x_i y_k = z_j} P(X = x_i,\ Y = y_k)$
$\frac{X}{Y}$	$\sum_{\frac{x_i}{y_k} = z_j} P(X = x_i,\ Y = y_k)$

7. X_1 és X_2 együttes sűrűségfüggvénye az $f(x_1, x_2)$. Transzformáljuk őket az

$$y_1 = r_1(x_1, x_2) \quad \text{és} \quad y_2 = r_2(x_1, x_2)$$

transzformációval, amelyről feltesszük, hogy egyértelműen megfordítható, az inverz transzformációt az

$$x_1 = h_1(y_1, y_2) \quad \text{és} \quad x_2 = h_2(y_1, y_2)$$

egyenletek adják meg. Tegyük fel, hogy a h_1 és h_2 függvények parciális deriváltjai léteznek és folytonosak, a belőlük alkotott Jacobi-féle determináns nem nulla:

$$J = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} \end{vmatrix} \neq 0.$$

Ekkor az

$$Y_1 = r_1(X_1, X_2) \quad \text{és} \quad Y_2 = r_2(X_1, X_2)$$

transzformált változók együttes sűrűségfüggvénye az $f(x_1, x_2)$ sűrűségfüggvényből így keletkezik:

$$g(y_1, y_2) = f(h_1(y_1, y_2), h_2(y_1, y_2)) \cdot |J|\,.$$

8. Ennek alkalmazásával kiszámítható a folytonos esetben X és Y *összegének, különbségének, szorzatának és hányadosának az eloszlása* az X és Y változók $f(x, y)$ együttes sűrűségfüggvénye, függetlenség esetén X és Y $f_1(x)$, ill. $f_2(y)$ peremsűrűség-függvénye segítségével:

z	A $g(z)$ peremsűrűség-függvény	
	általánosan	X és Y függetlensége esetén
$X+Y$	$g(z) = \int_{-\infty}^{\infty} f(z-t, t)dt$	$g(z) = \int_{-\infty}^{\infty} f_1(z-t)f_2(t)dt$
$X-Y$	$g(z) = \int_{-\infty}^{\infty} f(t, t-z)dt$	$g(z) = \int_{-\infty}^{\infty} f_1(t)f_2(t-z)dt$
XY	$g(z) = \int_{-\infty}^{\infty} \frac{1}{\lvert t\rvert} f\left(\frac{z}{t}, t\right) dt$	$g(z) = \int_{-\infty}^{\infty} \frac{1}{\lvert t\rvert} f_1\left(\frac{z}{t}\right) f_2(t)dt$
$\frac{X}{Y}$	$g(z) = \int_{-\infty}^{\infty} \lvert t\rvert f(zt, t)dt$	$g(z) = \int_{-\infty}^{\infty} \lvert t\rvert f_1(zt)f_2(t)dt$

9. *Két valószínűségi változó összegének várható értéke* a változók várható értékének az összege:

$$M(X+Y) = M(X)+M(Y).$$

Két független valószínűségi változó szorzatának várható értéke a változók várható értékének a szorzata:

$$M(XY) = M(X)M(Y).$$

Két független változó összegének a szórásnégyzete az egyes valószínűségi változók szórásnégyzetének az összege:

$$D^2(X+Y) = D^2(X)+D^2(Y).$$

Mindhárom tétel véges sok valószínűségi változóra is érvényes.

10. Az X és Y valószínűségi változók *kovarianciája:*

$$\operatorname{cov}(X, Y) = M\{[X-M(X)][Y-M(Y)]\} = \\ = M(XY)-M(X)M(Y),$$

korrelációs együtthatója:

$$R(X, Y) = \frac{M\{[X-M(X)][Y-M(Y)]\}}{D(X)D(Y)} = \frac{M(XY)-M(X)M(Y)}{D(X)D(Y)},$$

$$-1 \leqq R(X, Y) \leqq 1.$$

Ha X és Y között lineáris kapcsolat áll fenn:

$$Y = aX+b,$$

akkor $a>0$ esetén $R=1$, $a<0$ esetén $R=-1$. Ha fordítva: $R=1$ vagy $R=-1$, akkor az

$$Y = aX+b$$

lineáris kapcsolat 1 valószínűséggel áll fenn.

Ha X és Y függetlenek, akkor $R=0$. Sajnos, fordítva nem áll fenn: $R=0$ esetén nem biztos X és Y függetlensége. $R=0$ esetén jobb, ha csak annyit mondunk: X és Y *korrelálatlanok*.

Ha $|R|$ közel van 1-hez, akkor a kapcsolatot *közel lineárisnak* tekintjük. Ha $|R|$ közel jár a 0-hoz, *a lineáris összefüggés (lineáris korreláció) laza.* Lehet, hogy ilyenkor X és Y függetlenek, de lehet, hogy más, nemlineáris függvénnyel írható le az összefüggésük. *A korrelációs együttható a lineáris kapcsolat szorosságát (jóságát) méri.*

Kétdimenziós normál eloszlásnál azonban $R(X, Y) = 0$ esetén X és Y függetlenek.

11. Ha az Y X-re vonatkoztatott regresszióját (lásd a következő, 12. pontban) az $y = f(x)$ függvénnyel jellemezzük, a kapcsolat szorosságát (a korreláció erősségét) az

$$I(X, Y) = \sqrt{1 - \frac{D^2(Y-f(X))}{D^2(Y)}}$$

korrelációs index (korrelációs hányados) méri.

$$0 \leqq I \leqq 1,$$

ha $Y=f(X)$, akkor $I=1$. Ha fordítva: $I=1$ fennáll, akkor 1 valószínűségű az az esemény, hogy $Y=f(X)$.

Minél közelebb van I az 1-hez, annál erősebben fennáll, hogy $Y \approx f(X)$.

Ha X és Y független, akkor $I=0$. Sajnos, fordítva: $I=0$ esetén lehet, hogy X és Y független, de lehet, hogy más összefüggés áll fenn köztük. Ilyenkor csak annyit mondunk: X és Y *korrelálatlan.* Ha I közel esik 0-hoz, a feltételezett $Y=f(X)$ kapcsolat laza. *A korrelációs index a feltételezett $Y=f(X)$ kapcsolat szorosságát méri.* Lineáris kapcsolat esetében $I = |R|$.

12. *Diszkrét változó feltételes várható értéke:*

$$M(Y/X=x_i) = \sum_k y_k P(Y=y_k/X=x_i) = \sum_k y_k \frac{p_{ik}}{\sum_k p_{ik}}.$$

Hasonlóképpen írható fel $M(X/Y=y_k)$ is.

Folytonos változó feltételes várható értéke:

$$M(Y/X=x) = \int_{-\infty}^{\infty} yf(y/x)\,dy = \frac{1}{f_1(x)} \int_{-\infty}^{\infty} yf(x, y)\,dy.$$

Hasonló összefüggés áll fenn $M(X/Y=y)$-ra is.

Van egy (X, Y) véletlen vektorváltozónk. *X adott értéke alapján Y értékét célszerű az $M(Y/X=x)$ feltételes várható értékkel becsülni.* Ezt a becslést minden x-re elvégezve, a kapott

$$y = m_2(x) = M(Y/X=x)$$

függvény Y-nak X-re vonatkozó *elsőfajú regressziója (regressziós függvénye).*

Hasonlóképpen beszélhetünk az

$$x = m_1(y) = M(X/Y=y)$$

elsőfajú regressziós függvényről is. Ha X és Y együttes eloszlása normális, akkor a két regressziós függvény egyenes. Ha X és Y függetlenek, a regressziós függvények állandók.

Ha Y-t X-nek bármely más $h(X)$ függvényével becsüljük, $Y-m_2(X)$ négyzetének várható értéke kisebb, mint $Y-h(X)$ négyzetének várható értéke, vagy legfeljebb egyenlő vele:

$$M\{[Y-h(X)]^2\} \geqq M\{[Y-m_2(X)]^2\}.$$

Hasonló minimumtulajdonsággal rendelkezik $m_1(y)$ is.

Ha tehát X és Y véletlen változók közül *az egyiknek adott értékéből a másik értékére kívánunk következtetni, akkor a legjobb becslést (a most mutatott legkisebb négyzetek elve alapján) az elsőfajú regressziós függvény adja.*

Kétváltozós normál eloszlás esetén a regressziós függvények egyenesek.

Nagyon sokszor a két véletlen változó együttes eloszlása nem ismeretes. Ilyenkor is szokták az egyik változó értékét a másiknak lineáris függvényével, másodfokú függvényével stb. becsülni. Ezek a függvények a *másodfajú regressziós függvények* (regressziós egyenes, regressziós parabola stb.).

Most azt az

$$y = ax+b$$

egyenest keressük, amelyre

$$H(a, b) = M[(Y-aX-b)^2]$$

minimális lesz *(legkisebb négyzetek elve)*. Megoldásként az

$$y = r\frac{\sigma_2}{\sigma_1}(x-m_1)+m_2$$

egyenletű *másodfajú regressziós egyenest kapjuk*. Itt

$$m_1=M(X),\ m_2=M(Y),\ \sigma_1=D(X),\ \sigma_2=D(Y),\ r=R(X,\ Y).$$

Hasonlóképpen, X-nek Y-ra vonatkoztatott, másodfajú regressziós egyenese:

$$x = r\frac{\sigma_1}{\sigma_2}(x-m_2)+m_1.$$

Ha az elsőfajú regressziós függvények egyenesek, akkor azok megegyeznek a másodfajúakkal.

Mindkét fajta regressziós függvénynél az adott x behelyettesítésével Y becsült értékét kapjuk meg, amit így jelölünk: $\hat{Y}$. Az $Y-\hat{Y}$ eltérés az ún. *maradék* (reziduum). A maradékok szórásnégyzete (ami egyben a minimális négyzetes eltérést adó négyzetösszeg is):

$$D^2(Y-\hat{Y}) = \sigma_2^2(1-r^2).$$

13. A regressziónál mondottakat kiterjesztjük kettőnél több változóra is. Ha X_1 véletlen változó az $X_2, X_3, \ldots, X_n$ véletlen változóktól függ, X_1 becslését a többi adott érték esetén az

$$x_1 = M(X_1/X_2 = x_2, \ldots, X_n = x_n) = m(x_2, x_3, \ldots, x_n)$$

elsőfajú regressziós függvénnyel (az elméleti regresszióval) kaphatjuk. Ennek is megvan az említett minimumtulajdonsága:

$$M\{[X_1-h(X_2, \ldots, X_n)]^2\} \geqq M\{[X_1-m(X_2, \ldots, X_n)]^2\}.$$

Együttes n-dimenziós normál eloszlásnál ez a legkisebb négyzetek elve szerint a „legjobban közelítő", regressziós függvény egy „hipersík".

Az együttes eloszlás nem mindig ismert, ezért gyakran keresnek X_1-et jól közelítő, egyszerűen kezelhető *másodfajú regressziós függvényt,* pl. regressziós síkot

$$x_1 = a_1+b_{12}x_2+\ldots+b_{1n}x_n$$

alakban. A legkisebb négyzetek elve alapján az a követelmény, hogy

$$M[(X_1-a_1-b_{12}X_2-\ldots-b_{1n}X_n)^2]$$

legyen minimális. Megoldásként a következő „hipersíkot" (regressziós síkot) kapjuk:

$$x_1 = m_1-\sum_{k=2}^{n}\frac{C_{1k}}{C_{11}}(x_k-m_k),$$

vagy más alakban:

$$x_1 = m_1-\sum_{k=2}^{n}\frac{\sigma_1}{\sigma_k}\frac{R_{1k}}{R_{11}}(x_k-m_k).$$

Az x_k változók együtthatóit regressziós együtthatóknak nevezik. Itt $m_k=M(X_k)$, $\sigma_k=D(X_k)$, C_{ik} és R_{ik} az $X_1, X_2, \ldots, X_k$

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ \hdashline c_{n1} & c_{n2} & \ldots & c_{nn} \end{bmatrix}$$

kovarianciamátrixának, ill.

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ \hdashline r_{n1} & r_{n2} & \ldots & r_{nn} \end{bmatrix}$$

korrelációmátrixának c_{ik}, ill. r_{ik} elemeihez tartozó előjeles aldeterminánsai, $c_{ik} = \text{cov}(X_1, X_k)$, $r_{ik} = R(X_i, X_k)$. A kovariancia-, ill. a korrelációs mátrix szimmetrikus és $c_{ik} = \sigma_i\sigma_k r_{ik}$, $c_{ii} = \sigma_i^2$, $r_{ii} = 1$.

Ugyanígy kapható X_i változó becslése a többi változó segítségével:

$$X_i = m_i - \sum_{\substack{k=1\\k\neq i}}^{n} \frac{C_{ik}}{C_{ii}}(X_k - m_k) = m_i - \sum_{\substack{k=1\\k\neq i}}^{n} \frac{\sigma_i}{\sigma_k}\frac{R_{ik}}{R_{ii}}(X_k - m_k).$$

Az $X_i - \hat{X}_i$ *maradék* korrelálatlan a többi $(X_1, \ldots, X_{i-1}, X_i, \ldots, X_n)$ változóval, szórásnégyzete pedig

$$D^2(X_i - \hat{X}_i) = \frac{|C|}{C_{ii}} = \sigma_i^2\frac{|R|}{R_{ii}},$$

ahol $|C|$ a kovarianciamátrix, $|R|$ a korrelációs mátrix determinánsa. A maradék szórásnégyzete a legkisebb négyzetek elvénél kitűzött minimális várható értéket adja meg, így a becslés hatásos voltát jellemzi.

Az r_{ik} *totális korrelációs együtthatók* az X_i és X_k közti *teljes kapcsolat* erősségének a mérőszámai. Ha ebből a kapcsolatból a többi változó hatását kiküszöböljük, a két változó *közvetlen kapcsolatának* erősségét a *parciális korrelációs* együttható mutatja:

$$\varrho_{ik} = -\frac{C_{ik}}{\sqrt{C_{ii}C_{kk}}} = -\frac{R_{ik}}{\sqrt{R_{ii}R_{kk}}}.$$

A többi változó hatásának a kiküszöbölése a következőképpen történik. A többi (vagyis az $(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)$ változók segítségével elkészítjük az X_i és X_k regressziós becsléseit, $\hat{X}_i$-t és $\hat{X}_k$-t. Az $X_i - \hat{X}_i$, ill. $X_k - \hat{X}_k$ maradékok korrelálatlanok a többi változóval, *ezen maradékok korrelációs együtthatója* a többi változó hatását csökkentve (ill. normál eloszlás esetén a többi változó hatásától függetlenül) méri X_i és X_k *közvetlen kapcsolatát* ($X_i - \hat{X}_i$ pozitívan korrelált X_i-vel, $X_k - \hat{X}_k$ pozitívan korrelált X_k-val).

X_i-nek készítsük el a többi változóval való regressziós becslését, $\hat{X}_i$-t! Az X_i és $\hat{X}_i$ korrelációs együtthatója *a regressziós becslés jóságát méri*, ez az ún. *többszörös korrelációs együttható:*

$$\varrho_i = \sqrt{1 - \frac{|C|}{\sigma_i^2 C_{ii}}} = \sqrt{1 - \frac{|R|}{R_{ii}}}, \qquad 0 \leqq \varrho_i \leqq 1,$$

ha $\varrho_i = 1$, akkor 1 valószínűséggel $X_i = \hat{X}_i$. Minél közelebb esik ϱ_i az 1-hez, annál jobb, minél közelebb esik ϱ_i a 0-hoz, annál lazább a kapcsolat X_i és $\hat{X}_i$ között.

Az X_1, X_2, X_3 véletlen változók totális, páronkénti r_{ik} korrelációs együtthatóiból a parciális korrelációs együtthatók kiszámíthatók. Az említett jelölésmód mellett a Yule-féle jelölésmódot használva:

$$\varrho_{12} = r_{12\cdot 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}.$$

Az 12 index mutatja, hogy X_1 és X_2 parciális korrelációs együtthatóját számoltuk, kiszűrve a pont utáni 3. változó hatását. Ugyanígy

$$\varrho_{13} = r_{13\cdot 2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}},$$

$$\varrho_{23} = r_{23\cdot 1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}}.$$

4 vagy több változó esetén is kifejezhetők a parciális korrelációs együtthatók a totális korrelációs együtthatókkal, de egyre bonyolultabban.

Két véletlen változó, X és Y összefüggését n-edfokú parabolával közelítve

$$y = M(Y/X = x) \approx a + b_1x + b_2x^2 + \ldots + b_nx^n,$$

a kapott probléma visszavezethető egy n-dimenziós lineáris függvénnyel való közelítésre. Ha ugyanis az $x = x_1$, $x^2 = x_2$, ..., $x^n = x_n$ helyettesítést végrehajtjuk, ezt kapjuk:

$$y = M(Y/X = x) \approx a + b_1x_1 + b_2x_2 + \ldots + b_nx_n.$$

8. FONTOSABB VALÓSZÍNŰSÉG-ELOSZLÁSOK

8.1. Diszkrét eloszlások

A diszkrét eloszlásokat az x_k érték felvételének valószínűségével, a várható értékkel és a szórásnégyzettel szokás jellemezni.

1. *Egyenletes eloszlás:*

$$P(X=x_k) = \frac{1}{n} \quad (k = 1, 2, 3, \ldots, n),$$

$$M = \frac{\sum_{i=1}^{n} x_i}{n}, \quad D^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - M^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2.$$

2. Annak a valószínűsége, hogy n független kísérlet során

– a p valószínűségű A esemény k-szor,
– az $1-p$ valószínűségű $\bar{A}$ esemény $(n-k)$-szor következett be (vagyis A bekövetkezéseinek a száma $X=k$),

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, 1, 2, \ldots, n).$$

Az ilyen eloszlású X változó *binomiális eloszlású.*

$$M(X) = np, \quad D^2(X) = np(1-p).$$

A binomiális eloszlásnál fellépő $P(X=k)$ valószínűségeket tartalmazza az 1. táblázat (l. a könyv végén), $p=0{,}05$ selejt-valószínűség esetén annak a valószínűsége, hogy az $n=10$-elemű visszatevéses, egymástól függetlenül vett mintában a selejtesek száma 0, 1, 2, a táblázatból így olvasható ki:

n	k	$p=0{,}05$
10	0	$0{,}5987 = P(X=0) = \binom{10}{0} 0{,}05^0 \cdot 0{,}95^{10}$
	1	$0{,}3151 = P(X=1) = \binom{10}{1} 0{,}05^1 \cdot 0{,}95^9$
	2	$0{,}0746 = P(X=2) = \binom{10}{2} 0{,}05^2 \cdot 0{,}95^8$

Annak a valószínűsége, hogy az előbbi mintában maximálisan 2 selejtes legyen, az előbbi valószínűségek összege:

$$P(X \leqq 2) = \sum_{k=0}^{2} P(X=k) = 0{,}9884.$$

Ugyanez rövidebben megtalálható a 2. táblázatban, amely a binomiális valószínűségek összegeit, a $\sum_{k=0}^{k_{max}} P(X=k)$ valószínűségeket tartalmazza:

p	k_{max}	$n=10$
0,05	2	0,9885

A binomiális eloszlás általánosítása a *polinomiális eloszlás.* Egy kísérletnek legyen r számú kimenetele, az $A_1, A_2, \ldots, A_r$ események. Annak a valószínűsége, hogy n független kísérlet során
a p_1 valószínűségű A_1 esemény k_1-szer,
a p_2 valószínűségű A_2 esemény k_3-szor,

. . . .

a p_r valószínűségű A_r esemény k_r-szer
következett be:

$$P(X_1=k_1, X_2=k_2, \ldots, X_r=k_r) = \frac{n!}{k_1!\,k_2!\ldots k_r!} p_1^{k_1} p_2^{k_2} \ldots p_r^{k_r},$$

ahol $k_1+k_2+\ldots+k_r = n$.

3. *Karakterisztikus eloszlás.* Ha egy kísérletsorozatnál a p valószínűségű A esemény következett be, legyen $X=1$, ha az $1-p$ valószínűségű $\bar{A}$ esemény, legyen $X=0$. Az X az A eseményhez rendelt karakterisztikus vagy indikátorváltozó, eloszlása a karakterisztikus eloszlás:

$$P(X=0) = 1-p, \qquad P(X=1) = p;$$
$$M(X)=p, \qquad D^2(X) = p(1-p).$$

4. Ha egy N darab alkatrészt tartalmazó halmazból, amelyben s a selejtesek száma, n elemű mintát veszünk *visszatevés nélkül,* akkor annak a valószínűsége, hogy a selejtes darabok száma $X=k$ lesz:

$$P(X=k) = \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}} \qquad (k=0, 1, 2, \ldots, n).$$

Ez az eloszlás a *hipergeometrikus eloszlás.* Ha a selejtarány $\frac{s}{N} = p$, akkor

$$M(X)=np, \qquad D^2 = np(1-p)\left(1-\frac{n-1}{N-1}\right).$$

Ha n és s elég nagy, a visszatevés nélküh mintavétel $P(X=k)$ valószínűsége jól közelíthető a binomiális eloszlás $P(X=k)$ valószínűségével, mert

$$\lim_{\substack{s\to\infty \\ N\to\infty}} \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}} = \binom{n}{k} p^k(1-p)^{n-k}.$$

Itt $\frac{s}{N} = p$ állandó.

Ennek általánosítása a *polihipergeometrikus eloszlás.* Legyen a független, visszatevés nélküli mintavételnél r-féle különböző kategória: $A_1, A_2, \ldots, A_r$, ahol egy-egy kategórián belül rendre $N_1, N_2 \ldots, N_r$ elem van. Annak a valószínűsége, hogy

az 1. kategóriába esők közül k_1-et,
a 2. kategóriába esők közül k_2-t,

. . . .

az r-edik kategóriába esők közül k_r-et tartalmaz az n-elemű minta:

$$P(X_1=k_1, X_2=k_2, \ldots, X_r=k_r) = \frac{\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_r}{k_r}}{\binom{N}{n}}.$$

Természetesen

$$N_1+N_2+\ldots+N_r = N,$$
$$k_1+k_2+\ldots+k_r = n.$$

5. Egy kísérlet két lehetséges kimenetele a p valószínűségű A esemény és az $1-p$ valószínűségű $\bar{A}$ esemény. A kísérletet tetszés szerinti sokszor egymástól függetlenül elvégezve, legyen X annak a kísérletnek a sorszáma, amelyben A először következett be. Ekkor annak a valószínűsége, hogy A először a k-adik kísérletben következett be:

$$P(X=k) = (1-p)^{k-1}p \qquad (k=1, 2, 3, \ldots).$$

Ez a valószínűség-eloszlás a *geometriai eloszlás:*

$$M(X) = \frac{1}{p}, \qquad D^2(X) = \frac{1-p}{p^2}.$$

6. Egy kísérlet két lehetséges kimenetele: A (p valószínűséggel), $\bar{A}$ ($1-p$ valószínűséggel).

Többször, egymástól függetlenül megismételve ezt a kísérletet, legyen X annak a kísérletnek a sorszáma, amelyben A éppen r-edszer következett be ($X=r$, $r+1$, $r+2$, $r+3$, ...). Annak a valószínűsége, hogy A r-edszer az $(r+k)$-adik esetben következett be:

$$P(X=r+k)=\binom{r+k-1}{r-1}p^r(1-p)^k \quad (k=0, 1, 2, 3, \ldots).$$

Ez a valószínűség-eloszlás a *negatív binomiális eloszlás.*

$$M(X)=\frac{r}{p}, \quad D^2(X)=\frac{r(1-p)}{p^2}.$$

7. A binomiális eloszlás $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ valószínűsége tart a $\frac{\lambda^k e^{-\lambda}}{k!}$ valószínűséghez, ha $n\to\infty$, $p\to 0$, de úgy, hogy közben $np=\lambda$ állandó. Ezért

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}\approx\frac{\lambda^k e^{-\lambda}}{k!},$$

ha n elég nagy, p elég kicsi és $np=\lambda$.

Ha az X változó a k értéket

$$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!} \quad (k=0, 1, 2, \ldots)$$

valószínűséggel veszi fel, akkor a véletlen változó

Poisson-eloszlású:

$$M(X)=D^2(x)=\lambda.$$

A Poisson-eloszlás valószínűségei kiolvashatók a 3. táblázatból, az eloszlás valószínűségeinek összege a 4. táblázatból. Pl. $\lambda=0{,}5$ esetén $X=0, 1, 2, 3$ valószínűsége:

k	$\lambda=0{,}5$
0	0,6065
1	0,3033
2	0,0758
3	0,0126

Ezen valószínűségek összege 0,9982:

$\lambda=np$	$k_{max}=3$
0,50	998

8.2. Folytonos eloszlások. A véletlen változó transzformációja

A folytonos eloszlások jellemzése általában az $f(x)$ sűrűségfüggvénnyel, az $F(x)$ eloszlásfüggvénnyel, az $M(X)$ várható értékkel és a $D^2(X)$ szórásnégyzettel történik.

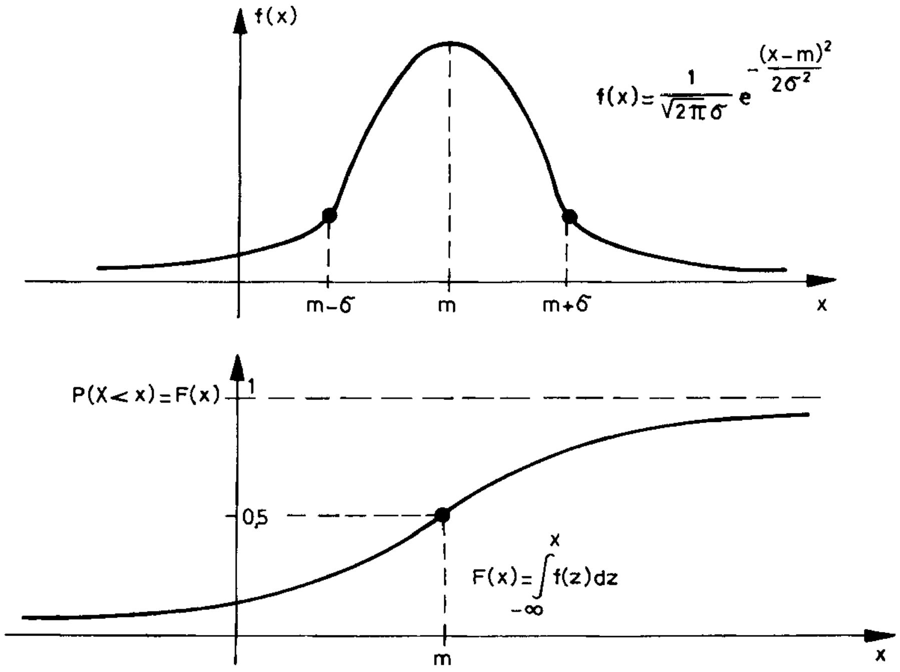
1. *Egydimenziós normál eloszlás* ($N(m, \sigma)$-eloszlás, 2. ábra):

$$f(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}} \quad \text{(Gauss-féle haranggörbe),}$$

$$F(x)=\int_{-\infty}^{x}f(z)\,dz, \quad M(X)=m, \quad D(X)=\sigma.$$

Az $m=0$ várható értékű, $\sigma=1$ szórású *standard normális eloszlás* ($N(0, 1)$-eloszlás) sűrűség- és eloszlásfüggvénye:

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \quad \Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{z^2}{2}}\,dz.$$



2. ábra

Ezek értékei az 5. táblázatban találhatók. Pl. $x=0{,}30$-nál $\varphi(0{,}30)=0{,}3814$, $\varnothing(0{,}30)=0{,}6179$:

x	$\varphi(x)$	$\varnothing(x)$
0,30	0,3814	0,6179

Negatív abszcisszánál a haranggörbe y tengelyre való szimmetriája miatt:

$$\varphi(-b)=\varphi(b),$$

ugyanakkor

$$\varnothing(-b) = 1-\varnothing(b).$$

Az általános $N(m, \sigma)$-eloszlás eloszlás- és sűrűségfüggvényének az értékeit $\varnothing(x)$ és $\varphi(x)$ segítségével így számíthatjuk:

$$F(x) = \varnothing\left(\frac{x-m}{\sigma}\right), \quad f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-m}{\sigma}\right).$$

Nagy n esetén a binomiális eloszlás valószínűségei megközelíthetően kiszámíthatók a $\varphi(x)$ függvény segítségével:

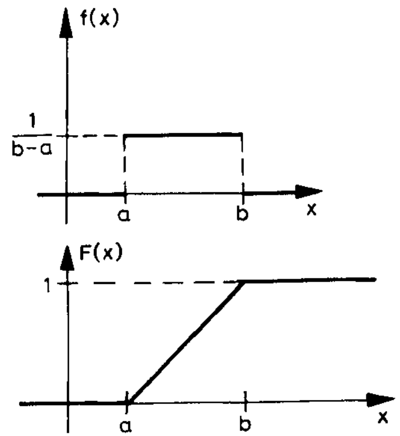
$$\binom{n}{k}p^k(1-p)^{n-k} \approx \frac{1}{\sqrt{np(1-p)}}\varphi\left(\frac{k-np}{\sqrt{np(1-p)}}\right),$$

$$\sum_{k<x}\binom{n}{k}p^k(1-p)^{n-k} \approx \varnothing\left(\frac{x-np}{\sqrt{np(1-p)}}\right).$$

A közelítés annál jobb, minél közelebb van p az $\frac{1}{2}$-hez.

Két vagy több független, normális eloszlású változó összege szintén normális eloszlású, $m = \sum_{i=1}^{n} m_i$, $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$ (a várható értékek, valamint a szórásnégyzetek összeadódnak).

2. Az (a, b) intervallumon egyenletes eloszlás sűrűségfüggvénye az (a, b) intervallumon állandó, eloszlásfüggvénye ennek integrálásával kapható (3. ábra):



3. ábra

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{ha} \quad a<x<b, \\ 0 & \text{különben,} \end{cases}$$

$$F(x) = \begin{cases} 0, & \text{ha} \quad x \leqq a, \\ \dfrac{x-a}{b-a}, & \text{ha} \quad a<x \leqq b, \\ 1, & \text{ha} \quad b<x, \end{cases}$$

$$M(X) = \frac{a+b}{2}, \quad D^2(X) = \frac{(b-a)^2}{12}.$$

Egy-egy intervallumba esés valószínűsége egyenesen arányos az intervallum hosszával.

A számítógépeken gyakran van *véletlenszám-generátor,* amely a (0, 1) intervallumon egyenletes eloszlású véletlenszám-sorozatot állít elő. A Commodore 64-es személyi számítógépen az RND(0) függvény egy-egy értéke ilyen véletlenszám, kinyomtatása

```
PRINT RND(Ø)
```

utasítással történik, $0 < \text{RND}(\emptyset) < 1$.

A gép által előállított, a (0, 1) intervallumon egyenletes eloszlású sorozatból tetszőleges F(y) eloszlásfüggvénynek megfelelő eloszlású véletlenszám-sorozat nyerhető. Ugyanis az

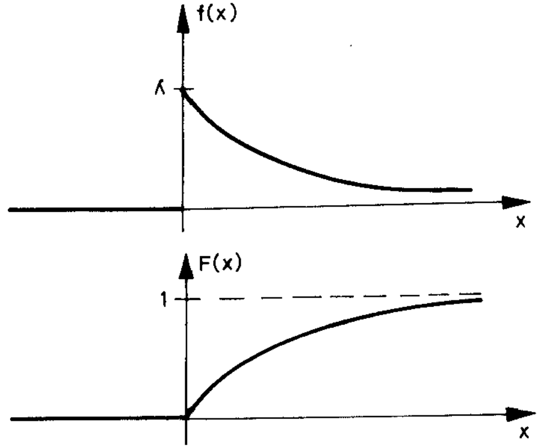
$$x = F(y)$$

értékek egyenletes eloszlásúak a (0, 1) intervallumon. Az egyenletből $F(y)$ inverz függvénye segítségével kapható

$$y = F^{-1}(x)$$

a kívánt eloszlású lesz.

3. Az *exponenciális eloszlású* véletlen változó sűrűség- és eloszlásfüggvénye (4. ábra):



4. ábra

$$f(x) = \begin{cases} 0, & \text{ha} \quad x \leqq 0, \\ \lambda e^{-\lambda x}, & \text{ha} \quad x>0, \end{cases}$$

$$F(x) = \begin{cases} 0, & \text{ha} \quad x \leqq 0, \\ 1-e^{-\lambda x}, & \text{ha} \quad x>0; \end{cases} \qquad M(X) = D(X) = \frac{1}{\lambda}.$$

Egyes élettartamokra az ún. *„örökifjú tulajdonság”* érvényes: bármely x időpontot választva, ha az élettartam x ideig nem ért véget, akkor úgy tekinthető, mintha a folyamat az x időpontban kezdődött volna:

$$P(X \geqq x+y / X \geqq x) = P(X \geqq y), \qquad (x>0,\ y>0).$$

Igazolható, hogy ez a tulajdonság azt jelenti, hogy X exponenciális eloszlású.

Az exponenciális eloszláshoz szükséges e^x, e^{-x} értékek megtalálhatók a 6. táblázatban. Pl.

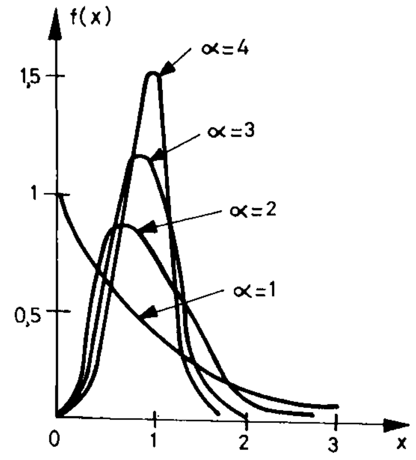
$$e^{0,50} = 1{,}6487, \quad e^{-0,50} = 0{,}6065:$$

x	e^x	e^{-x}
0,50	1,6487	0,6065

Az exponenciális eloszlás általánosítása a *Weibull-eloszlás,* amelynek eloszlásfüggvénye:

$$F(x) = \begin{cases} 1-e^{-cx^{\alpha}}, & \text{ha} \quad x \geqq 0, \\ 0, & \text{ha} \quad x < 0. \end{cases}$$

Itt $c>0$, $\alpha>0$ állandók. Sűrűségfüggvénye (5. ábra):



5. ábra

$$f(x) = \begin{cases} c\alpha x^{\alpha-1}e^{-cx^{\alpha}}, & \text{ha} \quad x \geqq 0, \\ 0, & \text{ha} \quad x < 0. \end{cases}$$

4. A

$$\Gamma(p) = \int_0^{\infty} t^{p-1}e^{-t}dt$$

integrál az ún. Euler-féle gammafüggvény $p>0$ esetén konvergens. Ha p egész szám, akkor

$$\Gamma(p) = (p-1)!.$$

X valószínűségi változó (a, p) paraméterű *gamma eloszlású,* ha sűrűségfüggvénye

$$f(x) = \begin{cases} \dfrac{a^p}{\Gamma(p)} x^{p-1}e^{-ax}, & \text{ha} \quad x \geqq 0, \\ 0, & \text{ha} \quad x < 0, \end{cases}$$

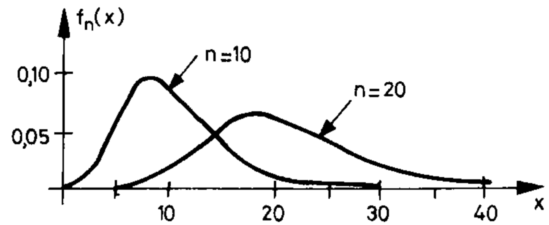
ahol $a>0$, $p>0$ állandók.

$$M(X) = \frac{p}{a}, \qquad D^2 = \frac{p}{a^2}.$$

Független, exponenciális eloszlású, azonos a paraméterű, p darab véletlen változó összege gammaeloszlású, (a, p) paraméterekkel. Független $(a, p_1), \ldots, (a, p_n)$ paraméterű, gammaeloszlású véletlen változók összege is gammaeloszlású $(a, p_1 + \ldots + p_n)$ paraméterekkel.

5. n számú független, standard normális eloszlású valószínűségi változó négyzetösszegének az eloszlását *n szabadságfokú, χ^2-eloszlásnak* nevezzük. Az előbbi négyzetösszeg négyzetgyöke *χ-eloszlású.*

Igazolható, hogy a χ^2-eloszlás $\left(\dfrac{1}{2}, \dfrac{n}{2}\right)$ paraméterű gammaeloszlás, így a χ^2-eloszlású változó sűrűségfüggvénye (6. ábra):



6. ábra

$$f_n(x) = \begin{cases} \dfrac{\left(\dfrac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\dfrac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x}, & \text{ha} \quad x \geqq 0, \\ 0, & \text{ha} \quad x < 0. \end{cases}$$

$$M(X) = n, \quad D^2(X) = 2n.$$

A 7. táblázatban megadjuk a χ^2-eloszlású változó eloszlásfüggvényének értékeit. Pl. $f = 11$ szabadságfok esetén $P(\chi^2 < 19{,}7) = 95\%$:

Szabadságfok	Valószínűség, %	
	90,0	95,0
11	17,3	19,7

Feltehető-e adott minta alapján, hogy a sokaság normális, exponenciális vagy egyéb eloszlású? Ennél az ún. *illeszkedésvizsgálatnál* fogunk találkozni a χ^2-próbával.

A χ-eloszlású változó sűrűségfüggvénye:

$$g_n(x) = \begin{cases} \dfrac{\left(\dfrac{1}{2}\right)^{\frac{n}{2}-1}}{\Gamma\left(\dfrac{n}{2}\right)} x^{n-1} e^{-\frac{1}{2}x^2}, & \text{ha } x \geqq 0, \\ 0, & \text{ha } x < 0. \end{cases}$$

6. Legyenek $Y, X_1, X_2, \ldots, X_f$ független, $N(0, 1)$-eloszlású változók. Ekkor *f szabadságfokú, Student- (vagy t-) eloszlásnak* nevezzük a

$$t = \frac{\sqrt{f}\, Y}{\sqrt{X_1^2 + X_2^2 + \ldots + X_f^2}}$$

valószínűségi változó eloszlását. Sűrűségfüggvénye:

$$f(x) = \frac{1}{\sqrt{f\pi}} \frac{\Gamma\left(\dfrac{f+1}{2}\right)}{\Gamma\left(\dfrac{f}{2}\right)} \frac{1}{\left(\dfrac{x^2}{f}+1\right)^{\frac{f+1}{2}}}.$$

Várható értéke csak $f \geqq 2$ esetén létezik: $M(X) = 0$. Szórásnégyzete csak $f \geqq 3$ esetén létezik: $D^2(X) = \dfrac{f}{f-2}$.

A sűrűségfüggvény a haranggörbéhez hasonló; ha $f \to \infty$, a sűrűségfüggvény tart az $N(0, 1)$-eloszlás sűrűségfüggvényéhez.

Annak a valószínűségét, hogy $|t|$ adott határ alatt marad, a 8. táblázatban adtuk meg. Pl. $f = 9$ szabadságfok esetén $n = f + 1 = 10$, $P(|t| < 2{,}262) = P(-2{,}262 < t < 2{,}262) = 95\%$:

$n = f+1$	Statisztikai biztonság		
	80%	90%	95%
10	1,383	1,833	2,262

A táblázat $f \to \infty$ sorában a standard normál eloszlás valószínűségei találhatók. A táblázatot a t próbánál gyakran fogjuk használni.

7. Legyenek $X_1, X_2, \ldots, X_m$ és $Y_1, Y_2, \ldots, Y_n$ valószínűségi változók függetlenek, és mind $N(0, 1)$-eloszlásúak. Ekkor az

$$F = \frac{\dfrac{1}{m} \sum_{i=1}^{m} X_i^2}{\dfrac{1}{n} \sum_{i=1}^{n} Y_i^2}$$

valószínűségi változó eloszlása (m, n) *szabadságfokú F-eloszlás*. Sűrűségfüggvénye:

$$f_{m,n}(x) = \begin{cases} \left(\dfrac{m}{n}\right)^{\frac{m}{2}} \dfrac{\Gamma\left(\dfrac{m+n}{2}\right)}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)} \dfrac{x^{\frac{m}{2}-1}}{\left(1+\dfrac{m}{n}x\right)^{\frac{m+n}{2}}}, & \text{ha} \quad x \geqq 0, \\ 0, & \text{ha} \quad x < 0. \end{cases}$$

Várható értéke $n \geqq 3$ esetén $M = \dfrac{n}{n-2}$, szórásnégyzete $n \geqq 5$ esetén $D^2 = \dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$. A 9. táblázatban megadjuk az eloszlásfüggvény értékeit. Ha pl. a számláló szabadságfoka $f_1 = 8$, a nevező szabadságfoka $f_2 = 20$, akkor $F(2{,}45) = P(F < 2{,}45) = 95\%$:

	f_1	
	7	8
$f_2 = 20$	2,52	2,45

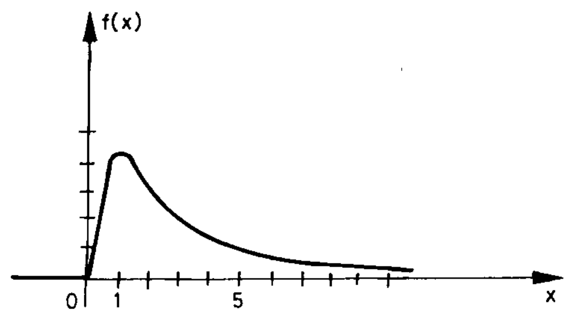
Az ún. F-próbánál gyakran fogjuk használni ezt a táblázatot.

8. Egy X valószínűségi változó *lognormális eloszlású*, ha logaritmusa $Y = \ln X$ normális eloszlású. Így eloszlásfüggvénye:

$$F(x) = P(X < x) = P(\ln X < \ln x) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\ln x} e^{-\frac{(u-m)^2}{2\sigma^2}} du,$$

ha $x > 0$, ha viszont $x \leqq 0$, akkor $F(x) = 0$.

Differenciálással adódik a sűrűségfüggvény (7. ábra):



7. ábra

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\,\sigma x} e^{-\frac{(\ln x - m)^2}{2\sigma^2}} & \text{ha} \quad x > 0, \\ 0, & \text{ha} \quad x \leqq 0. \end{cases}$$

$$M(X) = e^{m+\frac{\sigma^2}{2}}, \qquad D^2(X) = e^{2m+\sigma^2}(e^{\sigma^2} - 1).$$

9. Az X valószínűségi változó (p, q) paraméterekkel *bétaeloszlású*, ha sűrűségfüggvénye ($p > 0$ és $q > 0$ esetén):

$$f(x) = \begin{cases} \dfrac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}, & \text{ha} \quad 0 < x < 1, \\ 0, & \text{különben.} \end{cases}$$

$$M(X) = \frac{p}{p+q}, \qquad D(X) = \frac{1}{p+q}\sqrt{\frac{pq}{p+q+1}}.$$

10. Ha X folytonos valószínűségi változó, eloszlásfüggvénye $F(x)$, sűrűségfüggvénye $f(x)$, akkor az $Y = aX + b$ *lineáris transzformációval előállított véletlen változó* eloszlásfüggvénye:

$$G(y) = \begin{cases} F\left(\dfrac{y-b}{a}\right), & \text{ha } a > 0, \\ 1 - F\left(\dfrac{y-b}{a}\right), & \text{ha } a < 0; \end{cases}$$

sűrűségfüggvénye:

$$g(y)=\frac{1}{|a|}f\left(\frac{y-b}{a}\right).$$

Ha pedig az $y=r(x)$ szigorúan monoton és differenciálható függvény, amelyből $x=h(y)$, akkor az $Y=r(X)$ *transzformált véletlen változó* eloszlásfüggvénye:

$$G(y)=\begin{cases}F(h(y)), & \text{ha} \quad h'(y)>0,\\ 1-F(h(y)), & \text{ha} \quad h'(y)<0;\end{cases}$$

sűrűségfüggvénye:

$$g(y)=f(h(y))\,|h'(y)|.$$

8.3. Többdimenziós eloszlások

1. A többdimenziós eloszlásoknál is az az egyenletes eloszlás, amelynél k-szor akkora mérőszámú tartományba k-szor akkora valószínűséggel esik a változó, mint az eredeti tartományba. A sűrűségfüggvény állandó (mint az egydimenziós esetben).

Az (X, Y) valószínűségi változó a sík $|T|$ területű T tartományán egyenletes eloszlású, ha sűrűségfüggvénye:

$$f(x,y)=\begin{cases}\dfrac{1}{|T|}, & \text{ha} \quad (x,y)\in T,\\ 0, & \text{ha} \quad (x,y)\notin T.\end{cases}$$

Így a sík egy $|A|$ területű A tartományára való esés valószínűsége:

$$P((X,Y)\in A)=\frac{|A|}{|T|}.$$

Az elmondottak három vagy több dimenzióra is kiterjeszthetők.

2. (X, Y) *kétdimenziós normál eloszlású,* ha sűrűségfüggvénye:

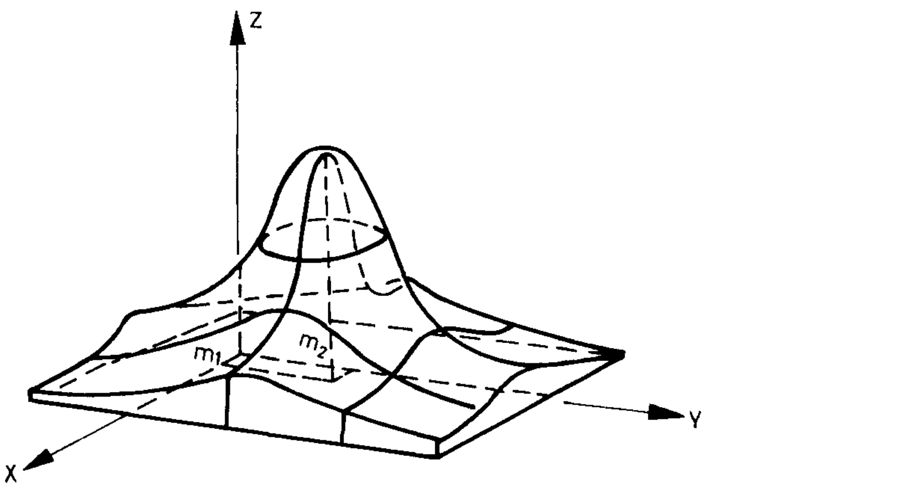
$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\,e^{-\frac{1}{2(1-r^2)}\left(\frac{(x-m_1)^2}{\sigma_1^2}-2r\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2}+\frac{(y-m_2)^2}{\sigma_2^2}\right)}$$

Itt m_1 és m_2 a peremeloszlások várható értékei, σ_1 és σ_2 a peremeloszlások szórásai, r pedig a két változó korrelációs együtthatója. A peremeloszlások sűrűségfüggvényei:

$$f_1(x)=\frac{1}{\sqrt{2\pi}\,\sigma_1}\,e^{-\frac{(x-m_1)^2}{2\sigma_1^2}},$$

$$f_2(y)=\frac{1}{\sqrt{2\pi}\,\sigma_2}\,e^{-\frac{(y-m_2)^2}{2\sigma_2^2}}.$$

X és Y tehát külön-külön normál eloszlásúak. Az együttes eloszlásuk sűrűségfüggvénye látható a 8. ábrán (haranghoz hasonló felület, $f(x, y)=$állandó síkmetszetei általában ellipszisek, amelyeknek tengelye általában nem párhuzamos az X, Y tengellyel).



8. ábra

3. Az $(X_1, X_2, X_3, \ldots, X_n)$ valószínűségi vektorváltozó *n-dimenziós normál eloszlású,* ha a változók együttes sűrűségfüggvénye:

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_1 \sigma_2 \ldots \sigma_n \sqrt{R}} e^{-\frac{1}{2R}\sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij}\frac{(x_i-m_i)(x_j-m_j)}{\sigma_i\sigma_j}}$$

Itt m_i az X_i változó várható értéke, σ_i pedig a szórása, r_{ij} az X_i és X_j változók korrelációs együtthatója. Az

$$R = \begin{vmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ r_{21} & r_{22} & \ldots & r_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} & \ldots & r_{nn} \end{vmatrix}$$

determináns a korrelációs együtthatókat tartalmazó korrelációs mátrix determinánsa, főátlójában minden $r_{ii}=1$ (mindegyik változó önmagával egyenlő, tehát a korrelációs együttható 1). A determináns szimmetrikus, $r_{ij}=r_{ji}$. Ezenkívül

$$0 < R \leqq 1.$$

Az R_{ij} az R determinánsban az r_{ij} korrelációs együtthatóhoz tartozó előjeles aldetermináns.

Minden X_i változó *peremeloszlása* normál eloszlás,

$$f_i(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-\frac{(x_i-m_i)^2}{2\sigma_i^2}}$$

peremsűrűség-függvénnyel.

Ha az $X_1, X_2, \ldots, X_n$ változók együttes eloszlása normális és a változók páronként korrelálatlanok ($r_{ij}=0$, ha $i \neq j$), akkor a változók függetlenek is.

4. A kétdimenziós normál eloszlású vektorváltozó elsőfajú regressziós függvénye

$$y = M(Y|X=x) = \frac{r\sigma_2}{\sigma_1}(x-m_1)+m_2,$$

ill.

$$x = M(X|Y=y) = \frac{r\sigma_1}{\sigma_2}(y-m_2)+m_1.$$

Jól látható, hogy *a kétdimenziós normál eloszlás regressziós függvényei lineárisak, a képük egyenes.* Pl. az első esetben $X=x$ feltétel esetén Y-ra egy eloszlást kapunk, $y = M(Y|X=x)$ ennek a várható értéke. Egy-egy x értéket megadva, meghatározva egy-egy $y = M(Y|X=x)$ ordinátát, az így kapott pontok a felírt egyenletű egyenesen sorakoznak.

9. A NAGY SZÁMOK TÖRVÉNYE. A CENTRÁLIS HATÁRELOSZLÁS-TÉTEL

1. Az $X_1, X_2, \ldots, X_n$ minta átlaga, $\bar{X}$ maga is véletlen változók (a mintaelemek) függvénye, ezért $\bar{X}$ is véletlen változó. *A mintaátlag várható értéke* az egész sokaság várható értéke, *szórása* az egész sokaság szórásának n-edrésze:

$$M(\bar{X}) = M(X), \qquad D(\bar{X}) = \frac{D(X)}{\sqrt{n}}.$$

2. A nagy számok törvényét először egy speciális esetre mondjuk ki: a p valószínűségű A esemény bekövetkezésével kapcsolatos $X_1, X_2, \ldots, X_n$ karakterisztikus változók számtani közepére, a $\frac{\sum_{i=1}^{n} X_i}{n} = \frac{X}{n}$ relatív gyakoriságra. *Bernoulli tétele szerint annak a valószínűsége, hogy* $\left|\frac{X}{n} - p\right|$ *ne legyen tetszőleges kicsi,* 0-*hoz tart, ha* $n \to \infty$:

$$\lim P\left(\left|\frac{X}{n} - p\right| \geqq \varepsilon\right) = 0.$$

Ezzel egyenértékű a következő: annak a valószínűsége, hogy $\left|\frac{X}{n}-p\right|$ tetszőleges kicsi, 1-hez tart:

$$\lim_{n\to\infty} P\left(\left|\frac{X}{n}-p\right| < \varepsilon\right) = 1.$$

Ennek az egyenlőtlenségnek a fennállását így is megfogalmazzuk: *az* $\frac{X}{n}$ *relatív gyakoriság sztochasztikusan konvergál a várható értékéhez, a p valószínűséghez* (gyenge konvergencia).

Általában Y_n sztochasztikusan konvergál Y-hoz, ha

$$\lim_{n\to\infty} P(|Y_n - Y| < \varepsilon) = 1$$

(gyenge konvergencia). Ha viszont

$$P(\lim_{n\to\infty} Y_n = Y) = 1,$$

akkor Y_n 1 valószínűséggel konvergál Y-hoz *(erős konvergencia).* Az erős konvergenciából a gyenge következik, de fordítva nem feltétlenül.

2. A *nagy számok törvénye* nemcsak karakterisztikus változók átlagára, hanem tetszőleges $X_1, X_2, \ldots, X_n$ független és azonos eloszlású valószínűségi változók átlagára kimondható, ha azoknak létezik $M(X_i) = m$ várható értéke és $D(X_i) = \sigma$ szórása. Annak a valószínűsége, hogy az átlag és várható érték eltérésének abszolút értéke tetszőleges kicsi lesz, határértékben 1-hez tart:

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - m\right| < \varepsilon\right) = 1.$$

Ez *a nagy számok törvényének Csebisev-féle alakja.*

4. *A központi (centrális) határeloszlás tétele* a következő. *Független, azonos eloszlású véletlen változók átlagának standardizáltja határértékben (aszimptotikusan) standard normál eloszlású:*

$$\lim_{n\to\infty} P\left(\frac{\bar{X}-m}{\sigma/\sqrt{n}} < x\right) = \lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - nm}{\sigma\sqrt{n}} < x\right) = \Phi(x).$$

A tétel az $X_1, X_2, \ldots, X_n$ karakterisztikus változókra így szól $\left(\sum_{i=1}^{n} X_i = X \text{ gyakoriság}\right)$:

$$\lim_{n\to\infty} P\left(\frac{X-np}{\sqrt{np(1-p)}} < x\right) = \Phi(x).$$

Ez az ún. *Moivre–Laplace-tétel,* amely szerint *a binomiális eloszlású X változó standardizáltja határértékben standard normál eloszlású.* Akkor viszont a binomiális eloszlású X változó közelítőleg normál eloszlású, np várható értékkel és $\sqrt{np(1-p)}$ szórással. Ebből kaphatók a binomiális eloszlás $P(X=k)$ valószínűségeire és az eloszlásfüggvényének $\sum_{k<x} P(X=k)$ értékeire a normális eloszlásnál megadott közelítő képletek.

5. Mind a nagy számok törvényét, mind a centrális határeloszlás tételét több, különböző változatban kimondták.

Ismertetünk mindegyikre egy-egy olyan tételt, amely nem követeli meg a változók azonos eloszlását.

Kolmogorov tétele: ha $X_1, X_2, \ldots$ független, véges szórású valószínűségi változók, $M(X_i) = 0$, $D(X_i) = \sigma_i$ $(i = 1, 2, \ldots)$, továbbá

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty,$$

akkor az $X_1, X_2, \ldots$ sorozatra teljesül a nagy számok erős törvénye:

$$P\left(\lim_{n\to\infty} \frac{X_1+X_2+\ldots+X_n}{n} = 0\right) = 1.$$

Laplace–Ljapunov-tétel: ha $X_1, X_2, \ldots$ olyan független valószínűségi változók, amelyeknek harmadik (és ennek következtében alacsonyabb rendű) momentumai léteznek, $M(X_k)=0$ $(k = 1, 2, \ldots)$, továbbá

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \beta_k}{S_n^3} = 0,$$

ahol

$$\beta_k = M(|X_k|^3), \qquad \sigma_k^2 = D^2(X_k), \qquad S_n = \sqrt{\sum_{k=1}^{n} \sigma_k^2},$$

akkor

$$\lim_{n\to\infty} P\left(\frac{X_1+X_2+\ldots+X_n}{S_n} < x\right) = \Phi(x).$$

II. MATEMATIKAI STATISZTIKA

1. A MATEMATIKAI STATISZTIKA TÁRGYA. STATISZTIKAI MINTA ÉS JELLEMZŐI. RENDEZETT MINTA

1. A matematikai statisztika a valószínűségszámításból fejlődött önálló tudományággá. Alapvető *feladata:* tapasztalati adatokból (mintából) következtetni az egész sokaság valószínűségeire, eloszlás- és sűrűségfüggvényére, azok paramétereire. Pl.:

– a mintában található selejtarány alapján következtetünk az egész sokaságbeli selejt-valószínűségre *(becsléselmélet);*

– statisztikai próbával megvizsgáljuk: a mintaátlagnak a várható értéktől való eltérése mellett az egész sokaságban feltehető-e, hogy teljesül egy adott szabványelőírás *(hipotézisvizsgálatok);*

– megvizsgáljuk, hogy a mintaátlag körül nagy valószínűséggel milyen intervallumban található az elméleti várható érték (az egész sokaság „átlaga"), ami az ún. *megbízhatósági- (konfidencia-) intervallum keresés;*

– adott mintabeli eloszlásfüggvény milyen elméleti eloszlásfüggvényhez illeszkedik jól (tehát milyen eloszlású az egész sokaság), ez az ún. *illeszkedésvizsgálat;*

– állítható-e két változóról, hogy egyforma eloszlású *(homogenitásvizsgálat)*;

– mérési eredmények alapján próbáljuk eldönteni, hogy milyen összefüggés áll fenn két változó között *(korreláció, regresszióanalízis)* stb.

Ezek a kérdések a műszaki életben igen fontosak, hiszen minőségellenőrzés során nincs lehetőség a teljes sokaság átvizsgálására, hanem csak egy n-elemű minta vételére van módunk. Három elnevezés is használatos arra, hogy az információ nem a teljes (elméleti) sokaság ismeretében jutott birtokunkba, hanem csak

a) n-elemű mintából nyert;
b) empirikus;
c) tapasztalati.

Mit jelent *az n-elemű minta?*
Egy $F(x)$ eloszlásfüggvénnyel rendelkező X valószínűségi változóra végezzünk n mérést, az eredmények:

$$x_{11},\ x_{12},\ x_{13},\ \ldots,\ x_{1n}.$$

Ha egy második, n-elemű méréssorozatot végzünk, az eredményként kapott

$$x_{21},\ x_{22},\ x_{23},\ \ldots,\ x_{2n}$$

sorozat általában nem ugyanaz, mint az első esetben volt. Célszerű tehát az *n-elemű statisztikai mintát mint*

$$X_1,\ X_2,\ X_3,\ \ldots,\ X_n$$

véletlen változók együttesét tekinteni, ahol az egyes változók (az egyes megfigyelések eredményei) X-szel azonos eloszlásúak és egymástól függetlenek.

2. A mintaelemekből alkotott függvényeket *statisztikai függvényeknek* (röviden: *statisztikáknak*) nevezzük. Ilyen statisztika (statisztikai függvény) pl. a *mintaátlag:*

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Mivel $\bar{X}$ a véletlentől függ, ezért maga is valószínűségi változó, van várható értéke, szórása stb. Mint már mondtuk, *a mintaátlag várható értéke az egész sokaság* (az eredeti X változó) *várható értéke:*

$$M(\bar{X}) = m,$$

a mintaátlag szórása pedig az eredeti X változó szórásának $\sqrt{n}$-edrésze:

$$D(\bar{X}) = \frac{D(X)}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}.$$

A mintaátlag eloszlásáról a centrális határeloszlásnál már beszéltünk, eredményként azt kaptuk, hogy

$$\lim_{n\to\infty} P\left(\frac{\bar{X}-m}{\sigma/\sqrt{n}} < x\right) = \Phi(x),$$

vagyis: a mintaátlag standardizáltja határértékben (aszimptotikusan) standard normál eloszlású. Ez azt jelenti, hogy maga *a mintaátlag aszimptotikusan $N(m;\ \sigma/\sqrt{n})$-eloszlású,* nagy n-ekre jó közelítésben normál eloszlású.

A mintaátlag sztochasztikusan az egész sokaság várható értékéhez konvergál (gyenge konvergencia). Bizonyos feltételek mellett (ha $D(X)$ létezik, vagy ha az $M(X^4)$ negyedik momentum létezik) fennáll az erős konvergencia is:

$$P(\lim_{n\to\infty} \bar{X} = M(X)) = 1.$$

Mivel a $\frac{Z_n}{n}$ relatív gyakoriság karakterisztikus változók átlaga, amelynek várható értéke az esemény p valószínűsége, ezért

$$\lim_{n\to\infty} P\left(\left|\frac{Z_n}{n} - p\right| > \varepsilon\right) = 0,$$

vagyis *egy esemény relatív gyakorisága sztochasztikusan konvergál az esemény valószínűségéhez.*

A mintaelemektől függő másik statisztika a *tapasztalati szórásnégyzet,* ez a mintaátlagtól való eltérések négyzetének az átlaga:

$$S_n^2 = \frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}.$$

Ennek pozitív négyzetgyöke a tapasztalati szórás: S_n.

A tapasztalati szórásnégyzet is véletlentől függő változó. Gyakorlatilag megköveteljük, hogy várható értéke az egész sokaság szórásnégyzete, σ^2 legyen. Ez azonban nem teljesül. Bizonyítható azonban, hogy ha n helyett $(n-1)$-gyel osztunk, akkor az ilyen módon *korrigált empirikus szórásnégyzet várható értéke* már az „elméleti" szórásnégyzet, σ^2 lesz. Tehát, ha

$$S_n^{*2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{n}{n-1} S_n^2,$$

akkor ennek várható értéke a véletlen változó szórásnégyzete:

$$M(S_n^{*2}) = \sigma^2.$$

A *korrigált tapasztalati szórásnégyzet* mint a valószínűségi változó *szórásnégyzete:*

$$D^2(S_n^{*2}) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right),$$

ahol $\mu_4 = M[(X-m)^4]$, az X változó negyedik centrális momentuma. Mivel a jobb oldal $n \to \infty$ esetén 0-hoz tart, ezért a Csebisev-egyenlőtlenséget felhasználva

$$P(|S_n^{*2} - \sigma^2| \geqq a) \leqq \frac{D^2(S_n^{*2})}{a^2} \to 0,$$

ha $n \to \infty$. Ez azt jelenti, hogy *a korrigált (és* vele együtt *a korrigálatlan) tapasztalati szórásnégyzet sztochasztikusan konvergál az elméleti szórásnégyzethez* (gyenge konvergencia).

Természetesen nagy mintaelemszám esetén S_n^2 és S_n^{*2} eltérése elhanyagolható. Mint véletlen változókat, nagybetűvel (vagy görög betűvel) szokás jelölni őket, de a statisztikai gyakorlatban kisbetű (s és s^*) is használatos. (Ugyanígy a mintaátlagnál $\bar{x}$ is szokásos jelölési mód.) *A zsebszámológépeken, számítógépeken gyakran van átlag- és szórásszámító szubrutin* (itt gyakran σ_n jelzi a korrigálatlan, σ_{n-1} a korrigált szórást). (Mi a melléje tett n idexszel azt jelöljük, hogy n-elemű mintából származnak.)

Ha a mintabeli adatok gyakorisággal vannak megadva (az x_i adat f_i-szer fordul elő), akkor a mintaátlag:

$$\bar{x} = \frac{\sum_{i=1}^{r} f_i x_i}{n},$$

a korrigált empirikus szórásnégyzet:

$$S_n^{*2} = \frac{\sum_{i=1}^{r} f_i (x_i - \bar{x})^2}{n-1}.$$

Itt r a ténylegesen különböző x_i értékek számát jelenti.

Az átlag és szórásnégyzet számításakor célszerű táblázattal dolgozni:

Adat, x_i	Gyakoriság, f_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
x_1 x_2 x_3 x_4 x_5	f_1 f_2 f_3 f_4 f_5				
Összesen:		$\Sigma f_i x_i$ $\bar{x} = \ldots$			$\Sigma f_i(x_i - \bar{x})^2$ $s^{*2} = \ldots$

Itt most kisbetűkkel jelöltünk (mintegy azt jelezve, hogy egy konkrét mintára is gondolhatunk a képletek alkalmazásakor).

A szórás és a mintaátlag aránya a *relatív szórás:* $\frac{S_n^*}{\bar{X}}$ (feltéve, hogy $\bar{X}>0$).

3. *A mintaátlag és a korrigálatlan szórásnégyzet eloszlása normál eloszlású sokaság esetén.* Ha a sokaság $N(m, \sigma)$-eloszlású, akkor $\bar{X}$ független, normál eloszlású változók összegének $\frac{1}{n}$-szerese, tehát maga is normális eloszlású. *Normális eloszlású sokaságnál az empirikus várható érték $N(m, \sigma/\sqrt{n})$-eloszlású.*

Normális eloszlású sokaságnál $\bar{X}$ és S^2 független valószínűségi változók. Ugyanígy $\bar{X}$ és S^{*2} is függetlenek.

$N(m, \sigma)$-eloszlású sokaság esetén $\frac{n}{\sigma^2}S^2$ χ^2-eloszlású valószínűségi változó, $n-1$ szabadságfokkal. Így $\frac{n-1}{\sigma^2}S^{*2}$ is χ^2-eloszlású $n-1$ szabadságfokkal.

4. *Számítástechnikai vagy számolástechnikai szempontból egyszerűbb képleteket* is adunk a tapasztalati szórásnégyzetre és mintaátlagra.

Az első az elméleti szórásnégyzetre megismert

$$D^2(X) = M(X^2) - M^2(X)$$

képlet megfelelője:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2,$$

$$S_n^{*2} = \frac{n}{n-1}S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\bar{X}^2.$$

Mintaátlagot és szórást könnyebb számolni ügyesen *megválasztott középérték segítségével.* A számolási időt csökkenti, ha kisebb számokkal dolgozunk. Ezért nagy számok vagy nagytömegű adat esetén

először durván megbecsüljük az adatok középértékét. Legyen ez a választott közép, c olyan, hogy az adatok tőle való eltérése minél kisebb és gyorsan számítható legyen. Ekkor az *átlag:*

$$\bar{X} = c + \frac{\sum_{i=1}^{n}(X_i - c)}{n},$$

a tapasztalati korrigált szórásnégyzet:

$$S^{*2} = \frac{1}{n-1}\left\{\sum_{i=1}^{n}(X_i-c)^2 - \frac{1}{n}\left[\sum_{i=1}^{n}(X_i-c)\right]^2\right\}.$$

Mintaátlag és szórásnégyzet gyors számítása osztályba sorolással. Ennél a statisztikában gyakran használatos módszernél minden adatot besorolunk egy-egy osztályba, azaz intervallumba (legyen az intervallumok hossza, azaz az osztályok terjedelme egyöntetűen k), utána *minden adatot helyettesítünk annak az osztálynak a középértékével, amelybe esett.* Így a mintaelemek átlaga jó közelítésben az osztályközepek átlaga, a szórás pedig az osztályközepek szórása.

Legyen valamelyik osztályközép (lehetőleg középen levő osztályközép) D. Az osztályközepek távolsága D-től:

$$\ldots, -3k, -2k, -k, 0, k, 2k, 3k, \ldots$$

Jelölje z_i azt, hogy az x_i osztályközép hányszor k-val tér el D-től, tehát a z_i értékek legyenek:

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$$

Ekkor a mintaátlag:

$$\bar{X} \approx D + \frac{k}{n}\sum_{i=1}^{r} f_i z_i,$$

a tapasztalati korrigált szórásnégyzet:

$$S_n^{*2} \approx \frac{k^2}{n-1}\left[\sum_{i=1}^{r} f_i z_i^2 - \frac{1}{n}\left(\sum_{i=1}^{r} f_i z_i\right)^2\right].$$

Itt r az osztályok számát jelenti. Az eljárásnál megállapodhatunk abban, hogy az osztályhatárra eső adatot az osztály alsó határával kezdődő osztályba soroljuk. A szükséges szummák kiszámítása *táblázattal* könnyen megoldható (egyszerűség kedvéért $r=5$ osztályt tartalmazó táblázatot mutatunk be):

Osztály	Osztályközép, x_i	Gyakoriság, f_i	z_i	$f_i z_i$	$f_i z_i^2$
1.	x_1	f_1	-2	$-2f_1$	$4f_1$
2.	x_2	f_2	-1	$-f_2$	f_2
3.	$x_3=D$	f_3	0	0	0
4.	x_4	f_4	1	f_4	f_4
5.	x_5	f_5	2	$2f_5$	$4f_5$
Összesen:		n		$\Sigma f_i z_i$	$\Sigma f_i z_i^2$

Figyeljük meg: a képletekhez az $f_i z_i$ és $f_i z_i^2$ oszlopokban álló számok összege szükséges, ez a táblázat utolsó sorában kiszámítható.

A mintaelemek osztályközéppel való helyettesítése során a szórásnégyzet torzítását csökkenti az ún. *Sheppard-korrekció*, $\frac{k^2}{12}$; a szórásnégyzet helyes közelítése:

$$\sigma^2 \approx S_n^2 - \frac{k^2}{12}.$$

Itt k az intervallumok (osztályok) hossza.

5. *Egyéb tapasztalati jellemzőket* is szokás kiszámítani a mintából. Rendezzük nagyság szerinti sorrendbe az $X_1, X_2, \ldots, X_n$ mintaelemeket, ez az ún. *rendezett minta:*

$$X_1^* \leqq X_2^* \leqq \ldots \leqq X_n^*.$$

Ha a mintaelemek száma páratlan, $n=2m+1$, akkor a minta *mediánja:* X_{m+1}^*. Ha a mintaelemek száma páros: $n=2m$, akkor a *mediánt* így definiáljuk:

$$\frac{X_m + X_{m+1}}{2}.$$

A minta terjedelme a minta legnagyobb és legkisebb elemének különbsége:

$$R = X_n^* - X_1^*.$$

Az r-edik *tapasztalati momentum* (a megfelelő elméleti momentum mintájára):

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r \quad (r = 1, 2, \ldots).$$

Az első momentum $m_1 = \bar{X}$, a második momentum $m_2 = S_n^2 + \bar{X}^2$.

A tapasztalati momentumok, mint a többi statisztikák, a megfelelő elméleti jellemzők becslésére használhatók.

Legyenek egy rendezett minta X_i elemei ($i = 1, 2, \ldots, n$):

$$\underline{1, 2, 3,}\ 4, 5, 6, 7, 8, 9, 10, 11, 12.$$

X_4^* az az elem, amelynél a mintaelemek $p=\frac{1}{4}$-része $(100\cdot\frac{1}{4}=25\%$-a) kisebb. Ezt az elemet nevezzük a minta negyedrendű kvantihsének. $n\cdot\frac{1}{4} = 12\cdot\frac{1}{4} = 3$ most egész szám, a negyedrendű kvantilis indexe ennél 1-gyel nagyobb.

Legyen most egy másik rendezett minta

$$\underline{1, 2,}\ 3, 4, 5, 6, 7, 8, 9, 10.$$

A negyedrendű kvantilis most $X_3^*=3$ (úgy tekintjük most is, hogy nála kisebb az elemek $\frac{1}{4}$-része, 25%-a). $n\cdot\frac{1}{4} = 10\cdot\frac{1}{4} = 2{,}5$ most nem egész szám, a negyedrendű kvantilis indexe *np* egészrészénél 1-gyel nagyobb.

Ugyanígy beszélhetünk a minta harmadrendű, ..., p-edrendű kvantiliséről.

A minta p-edrendű *kvantilise* az a mintaelem, amelynél kisebb a mintaelemek $100\,p$ %-a $(0<p<1)$. Ha az n darab mintaelem p-edrésze kisebb ennél a kvantilisnél, akkor $[np]$ darab kisebb van nála, vagyis ez az elem a rendezett mintából az $[np]+1$-edik: $X^*_{[np]+1}$. Itt $[np]$ az np szám egész részét, a nálánál nem nagyobb, hozzá legközelebb álló egész számot jelenti.

6. Egy-egy feladatban a folytonos (vagy diszkrét) elméleti sokaságot többnyire nem ismerjük, n-elemű mintát veszünk a becslésére. Ebből a diszkrét $X_1, X_2, \ldots, X_n$ rendezetlen mintából (vagy az $X_1^* \leqq X_2^* \leqq \ldots \leqq X_n^*$ rendezett mintából) becsüljük az elméleti sokaság jellemzőit, a várható értéket, szórást, mediánt stb., a diszkrét mintából nyert $F_n(x)$ *tapasztalati eloszlásfüggvénnyel közelítjük az egész sokaság* $F(x)$ *elméleti eloszlásfüggvényét, a tapasztalati* $f_n(x)$ *sűrűségfüggvénnyel az elméleti* $f(x)$ *sűrűségfüggvényt.* Tehát voltaképpen egy diszkrét eloszlással közelítjük a folytonos (vagy diszkrét) elméleti eloszlást.

Ha a teljes sokaság diszkrét eloszlású, akkor sűrűségfüggvénye nincs. Itt a $P(X=x_k)$ valószínűségeket hasonlítják össze a mintából és a feltételezett elméleti eloszlásból. Ha ezek között csak kis eltérések vannak, akkor feltehetjük az elméleti eloszlás fennállását. Természetesen ez is egy elsődleges tájékoztatásra szolgáló módszer, egzakt, számításos illeszkedésvizsgálatról később esik szó.

Az $X_1, X_2, \ldots, X_n$ mintához tartozó $F_n(x)$ *tapasztalati eloszlásfüggvény* ugyanúgy szerkesztendő, mint azt a diszkrét eloszlásnál mutattuk: a függvény az X tengellyel párhuzamos szakaszokból álló lépcsős függvény, amelynek minden egyes felvett x_i értéknél $\frac{1}{n}$ ugrása van, ha x_i-t egyszer kaptuk a mintában; $\frac{k}{n}$ ugrása van, ha x_i k-szor fordult elő a mintában. A függvény monoton nem csökkenő, balról folytonos, $F_n(-\infty)=0$ (sőt, $F_n(x)=0$, ha $x \leqq X_1^*$, ahol X_1^* a legkisebb mintaelem), $F_n(\infty)=1$ (sőt, $F_n(x)=1$, ha $x>X_n^*$, ahol X_n^* a legnagyobb mintaelem). Az $F_n(x)$ értéke annyiszor $\frac{1}{n}$, ahány x-nél kisebb mintaelem van:

$$F_n(x) = \sum_{X_i<x} \frac{1}{n},$$

mert $F_n(x)$ tulajdonképpen a $X<x$ esemény *relatív gyakorisága.* Ugyanezen esemény *valószínűségét* az elméleti eloszlásfüggvény adja meg:

$$P(X<x) = F(x).$$

A relatív gyakoriság: $F_n(x)$ a nagy számok törvénye értelmében várhatóan jól közelíti az $F(x)$ valószínűséget, ha az n elég nagy. *Ha tehát* n *nagy,* $F_n(x)$ *az elméleti* $F(x)$ *eloszlásfüggvény jó közelítésének tekinthető.*

Vizsgálták még, hogy mit tudunk mondani $F_n(x)$ és $F(x)$ eltérésének a maximumáról. Az eltérés abszolút értékének legkisebb felső korlátja az ún. supremum. Glivenko tétele szerint:

$$P(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x)-F(x)| = 0) = 1,$$

vagyis $F_n(x)-F(x)$ maximális eltérése abszolút értékben 1 valószínűséggel 0 felé tart. Másképpen: nagy a valószínűsége, hogy n növekedésével az $F_n(x)$ empirikus eloszlásfüggvény az egész számegyenesen egyenletesen tart az $F(x)$ elméleti eloszlásfüggvényhez. Ezt a fontos tételt szokták *a matematikai statisztika alaptételének* is nevezni.

Kimutatható, hogy $F_n(x) \to F(x)$ 1 valószínűséggel (erős konvergencia). Ekkor természetesen a gyenge konvergencia is teljesül.

$D_n = \sup_x |F_n(x)-F(x)|$ zérushoz való konvergenciájának gyorsaságáról szólnak Kolmogorov tételei: $\sqrt{n}\,D_n$ mint valószínűségi változó véges érték körül ingadozik, pontosabban:

$$\lim_{n\to\infty} P(\sqrt{n}\,D_n<z) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2z^2} = K(z) \qquad (z>0).$$

A Kolmogorov-féle $K(z)$ függvény értékeit táblázatba foglalva, segítségével a későbbiek folyamán azt tesztelhetjük: elfogadható-e, hogy X valószínűségi változó adott eloszlású.

Vezessük be a

$$D_n^+ = \sup_{-\infty<x<\infty} (F_n(x)-F(x)),\quad D_n^- = \sup_{-\infty<x<\infty} (F(x)-F_n(x))$$

jelöléseket! (Tehát pl. D_n^+ az a maximális érték, amennyivel az empirikus eloszlásfüggvény meghaladja az elméleti eloszlásfüggvényt.) Szmirnov tétele szerint

$$\lim_{n\to\infty} P(\sqrt{n}\,D^+ < z) = \lim_{n\to\infty} (\sqrt{n}\,D^- < z) = 1-e^{-2z^2} = S(z) \quad (z \geq 0).$$

A Szmirnov-féle $S(z)$ függvény értékeihez csak az e^x, e^{-x} táblázatra van szükség.

Ha n elég nagy, akkor a Kolmogorov–Szmirnov-tételek jobb oldalán álló kifejezések jól közelítik a $\sqrt{n}\,D$, $\sqrt{n}\,D^+$, $\sqrt{n}\,D^-$ változók eloszlását.

Említettük, hogy a minta tulajdonképpen egy diszkrét eloszlás. *Hogyan beszélhetünk mégis a minta $f_n(x)$ sűrűségfüggvényéről?* Olyan függvényt akarunk megszerkeszteni, amely alatti terület egy-egy részintervallumban megadja a változó intervallumba esésének a relatív gyakoriságát, azaz közelítőleg az intervallumba esés valószínűségét (hiszen ez a sűrűségfüggvény alapvető tulajdonsága).

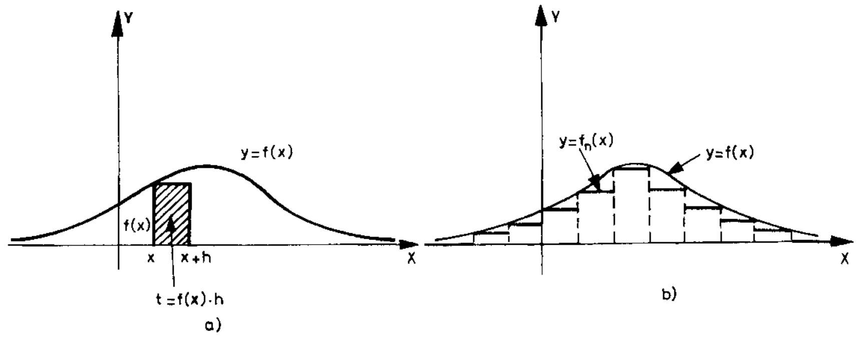
A 9a ábrán feltüntettük az ismeretlen elméleti $f(x)$ sűrűségfüggvényt. Az $x \leqq X < x+h$ intervallumban az X tengellyel párhuzamost húzva rajzoltunk egy h alapú, $f(x)$ magasságú téglalapot.

Az $y=f(x)$ függvény alatti terület az $(x, x+h)$ intervallumban annak a valószínűségét adja meg, hogy a változó az intervallumba essék. Ez a terület közelítőleg a téglalap területével egyezik meg:

$$P(x \leqq X < x+h) = \int_x^{x+h} f(x)\,dx \approx f(x)h.$$

Innen megkapjuk a már ismert képletet:

$$f(x) \approx \frac{P(x \leqq X < x+h)}{h}.$$



9ab ábra

Mivel az intervallumba esés valószínűsége közelítőleg a relatív gyakorisággal (az intervallumra eső adatok számának és az összes adatok számának arányával, $\frac{k}{n}$-nel) egyezik meg, ezért

$$f(x) \approx \frac{k}{nh} = \frac{\text{az intervallumba eső adatok száma}}{\text{(összes adat száma) (az intervallum hossza)}}.$$

Így a kívánt tulajdonsággal rendelkező $f_n(x)$ *tapasztalati sűrűségfüggvényt (hisztogramot)* a következőképpen szerkeszthetjük:

a) megvizsgáljuk, hogy a mintabeli adatok milyen (a, b) intervallumba esnek (a lehet a legkisebb mintaelem, vagy egy nálánál kisebb, hozzá közel eső, kerek szám, b a legnagyobb mintaelem, vagy egy közeli, nálánál valamivel nagyobb kerek szám);

b) az (a, b) intervallumot felosztjuk részintervallumokra (legtöbbször egyenlő hosszúságúakra). Minden részintervallum fölé az X tengellyel vízszintest húzva, szerkesztünk egy

$$f_n(x) = \frac{k}{nh} = \frac{\text{az intervallumba eső adatok száma}}{\text{(összes adat száma) (az intervallum hossza)}}$$

magasságú téglalapot. Az (a, b) intervallumon kívül $f_n(x)=0$. *Az így kapott lépcsős függvény $f_n(x)$, a tapasztalati sűrűségfüggvény (sűrűséghisztogram), amely közelíti az ismeretlen $f(x)$ elméleti sűrűségfüggvényt* (l. a 9b ábrát).

Ha ez a hisztogram egy haranggörbét közelít, akkor az eloszlást (legalábbis jó közelítésben) normálisnak tekinthetjük. Ha más, ismert sűrűségfüggvényt közelít, akkor a szóban forgó eloszlást tételezzük fel. *Ez egy első tájékoztatásra alkalmas illeszkedésvizsgálat.* Nehéz megvonni a határt, hogy milyen mértékű illeszkedés esetén fogadjuk el, vagy vetjük el a feltételezett eloszlást. Később majd szó lesz egzaktabb, számításos illeszkedésvizsgálatról is. Az elméleti és tapasztalati eloszlásfüggvény összehasonlítása, egymáshoz való illesztése is lehetséges, de nem olyan gyakori, mert az eloszlásfüggvény alakja nem olyan jellegzetes, mint a sűrűségfüggvényé.

Megemlítjük, hogy $f_n(x)$ tulajdonképpen az $F_n(x)$ tapasztalati eloszlásfüggvény differenciahányadosa,

$$f_n(x) = \frac{\Delta F_n(x)}{\Delta x},$$

amiből határértékben az

$$f(x) = \frac{dF(x)}{dx}$$

ismert összefüggés következik.

Kimutatható, hogy igen általános feltételek mellett $f_n(x)$ az n növelésével konvergál az $f(x)$ elméleti sűrűségfüggvényhez abban az értelemben, hogy

$$\max_x |f_n(x) - f(x)| \to 0, \quad \text{ha} \quad n \to \infty.$$

Ahhoz, hogy $f_n(x)$ jól mutassa $f(x)$ alakját (jól jellemezze az eloszlást), az is szükséges, hogy az osztópontok számát jól eltaláljuk. Ezt néha próbálgatni kell. Túl kevés vagy túl sok részintervallum felvételével hamis képet kaphatunk. A közök számát többnyire 6 és 14 között szokták felvenni, de nagy elemszámú mintánál több köz választása lehet alkalmas.

7. Legyen X folytonos véletlen változó $F(x)$ eloszlásfüggvénnyel és $f(x)$ sűrűségfüggvénnyel. Az $X_1, X_2, \ldots, X_n$ időrendi (mérési) sorrendben megadott mintát nagyság szerint rendezhetjük, az

$$X_1^* \leqq X_2^* \leqq \ldots \leqq X_n^*$$

rendezett mintát nyerjük. A rendezett mintaelemek szintén tekinthetők valószínűségi változóknak, azonban

a) már *nem függetlenek* (nagyságrendi reláció áll fenn közöttük, pl. X_2^* értéke sosem lehet kisebb a legkisebb mintaelemnél, X_1^* értékénél);

b) nem azonos eloszlásúak, X_1^* legkisebb mintaelem eloszlása más, mint a második legkisebbé, X_2^*-é stb. (egyikük eloszlásfüggvénye sem azonos az elméleti sokaság $F(x)$ eloszlásfüggvényével).

Jelölje $F_{n,k}(x)$ a „nagyságra nézve k-adik" mintaelem, X_k^* eloszlásfüggvényét! Háromféle megadás is ismeretes $F_{n,k}(x)$-re:

$$a) \quad P(X_k^* < x) = F_{n,k}(x) = \sum_{i=k}^{n} \binom{n}{i} [F(x)]^i [1 - F(x)]^{n-i},$$

a legkisebb és legnagyobb mintaelem eloszlásfüggvénye:

$$P(X_1^* < x) = F_{n,1}(x) = 1 - [1 - F(x)]^n,$$
$$P(X_n^* < x) = F_{n,n}(x) = [F(x)]^n;$$

$$b) \quad P(X_k^* < x) = F_{n,k}(x) =$$
$$= n\binom{n-1}{k-1} \int_{-\infty}^{x} [F(t)]^{k-1} [1 - F(t)]^{n-k} f(t)\, dt.$$

c) Végezzük el az $u = F(t)$, $du = f(t)\,dt$ helyettesítést! A határok:

$$t \to -\infty \text{ esetén } u = F(t) \to 0,$$
$$t = x \quad \text{ esetén } u = F(t) = F(x).$$

Így az eloszlásfüggvényre egyszerűbben áttekinthető alakot nyerünk:

$$P(X_k^* < x) = F_{n,k}(x) = n\binom{n-1}{k-1} \int_0^{F(x)} u^{k-1}(1-u)^{n-k}\, du,$$

ebből $k=1$ és $k=n$ esetén megkapjuk a legkisebb és legnagyobb mintaelem eloszlásfüggvényét:

$$P(X_1^* < x) = F_{n,1}(x) = n \int_0^{F(x)} (1-u)^{n-1}\, du,$$

$$P(X_n^* < x) = F_{n,n}(x) = n \int_0^{F(x)} u^{n-1}\, du.$$

Ha a *b)* vagy *c)* alatti eredményben $F_{n,k}(x)$-et deriváljuk x szerint, megkapjuk X_k^* *sűrűségfüggvényét:*

$$f_{n,k}(x) = n\binom{n-1}{k-1}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x).$$

Az

$$R_n = X_n^* - X_1^*$$

mintaterjedelem eloszlásfüggvénye:

$$P(R_n < r) = W_n(r) = n \int_{-\infty}^{\infty} [F(r+u)-F(u)]^{n-1} f(u)\, du.$$

Gyakorló feladatok

1. Automata töltőgép előírt súlyú mosóport tölt dobozokba. A dobozba töltött mosópor súlya X valószínűségi változó, az előírástól felfelé is, lefelé is eltér. Az eltéréseket osztályba sorolták. Az $n=12$ mérés közül a mosópor súlyeltérése -3 és -2 század newton közé esett 1 alkalommal, -2 és -1 közé 2 alkalommal stb. (l. a következő táblázatot).

Súlyeltérés-osztályhatárok, század N	Osztályközép, század N, x_i	Gyakoriság
$(-3;\ -2)$	$-2{,}5$	1
$(-2;\ -1)$	$-1{,}5$	2
$(-1;\ 0)$	$-0{,}5$	3
$(0;\ 1)$	$0{,}5$	3
$(1;\ 2)$	$1{,}5$	2
$(2;\ 3)$	$2{,}5$	1
Összesen:		$12=n$

a) Szerkesszük meg az adatokból a súlyeltérés *tapasztalati sűrűségfüggvényét,* $f_n(x)$-et; *b) tapasztalati eloszlásfüggvényét,* $F_n(x)$-et; *c)* számítsuk ki a *mintaátlagot* és *d)* a *tapasztalati korrigált szórást!*

A *b), c), d)* pontoknál tekintsük úgy, hogy a súlyeltérés értéke az osztályközép!

a) A tapasztalati sűrűségfüggvény egy intervallumon:

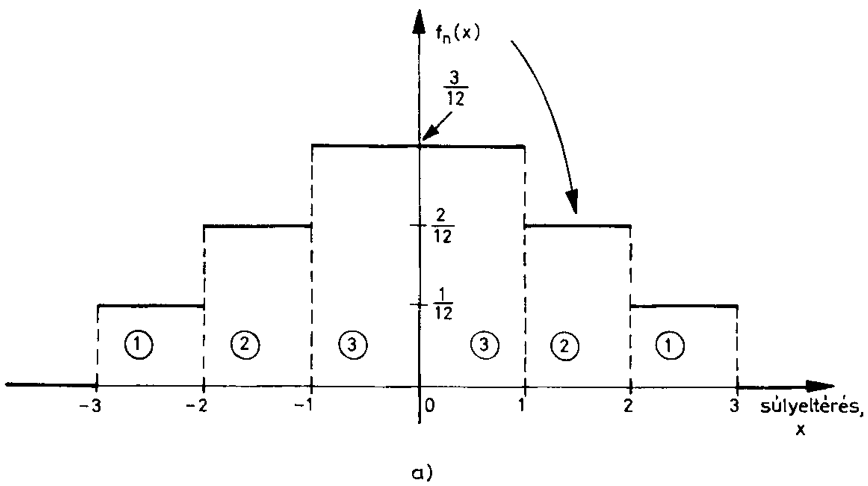
$$f_n(x) \approx \frac{\text{az intervallumra eső adatok száma}}{(\text{összes adatok száma})\ (\text{intervallum hossza})}.$$

Így a $(-3;\ -2)$ intervallumon a sűrűségfüggvény értéke az intervallum minden pontjában $\frac{1}{12\cdot 1} = \frac{1}{12}$, a $(-2;\ -1)$ intervallum minden pontjában $\frac{2}{12}$ stb. A 10*a* ábrán látható, hogy a súlyeltérés eloszlása

– az $x=0$-ra szimmetrikus, $(-n;\ -n+1)$ intervallumba ugyanannyi adat esik, mint $(n-1;\ n)$ intervallumba (n egész);

– a $(-1; 0)$ és $(0; 1)$ intervallumokba esik a legtöbb adat (3–3 adat), a $(-2;\ -1)$ és $(1; 2)$ intervallumokba kevesebb (2–2 adat), a $(-3;\ -2)$ és $(2; 3)$ intervallumokba még kevesebb (1–1 adat).

Mivel -3 alá, 3 fölé nem esett adat, $(-\infty)$-től (-3)-ig, 3-tól ∞-ig a sűrűségfüggvényt 0-nak tekintjük.



10*a* ábra

b) Az eloszlásfüggvény egy x helyen vett értéke annak a valószínűségét adja meg, hogy a változó x-nél kisebb legyen. Mivel $-2{,}5$ alá nem esik súlyeltérés, ezért a tapasztalati eloszlásfüggvény a $-2{,}5$ helyen 0:

$$F_n(-2{,}5) = P(X < -2{,}5) = 0,$$

ugyanígy

$$F_n(-3) = P(X < -3) = 0 \text{ stb.},$$

az $F_n(x)$ függvény $(-\infty)$-től $(-2{,}5)$-ig mindenütt 0.

$-1{,}5$ alá egy érték esik a 12 közül (a $-2{,}5$ osztályközép), ezért

$$F_n(-1{,}5) = P(X < -1{,}5) = \frac{1}{12},$$

$$F_n(-2) = P(X < -2) = \frac{1}{12} \text{ stb.},$$

az $F_n(x)$ értéke $-2{,}5 < x \leqq -1{,}5$ félig nyílt intervallumon mindenütt $\frac{1}{12}$.

Az $F_n(x)$ tapasztalati eloszlásfüggvény értéke a $(-2{,}5;\ -1{,}5]$, $(-1{,}5;\ -0{,}5]$, $(-0{,}5;\ 0{,}5]$, $(0{,}5;\ 1{,}5]$, $(1{,}5;\ 2{,}5]$ félig nyílt intervallumokban állandó. Értéke

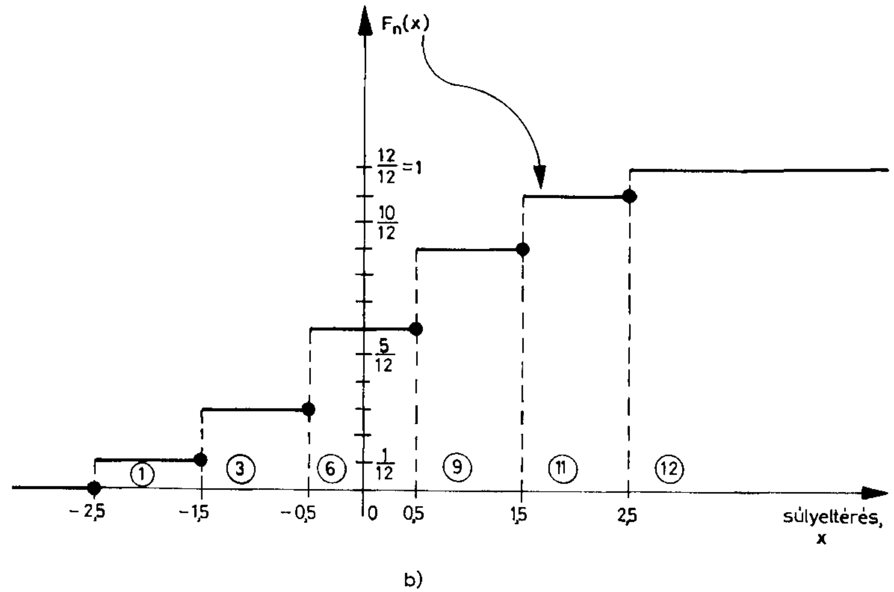
egy-egy intervallum végpontjában annyiszor $\frac{1}{12}$-del ugrik meg, ahányszor az intervallum végpontjához tartozó értéket a változó felveszi. $(-\infty)$-től $(-2{,}5)$-ig $F_n(x) = 0$, 2,5-től ∞-ig $F_n(x) = 1$ (l. a 10*b* ábrát).

c) A súlyeltérés eloszlása szimmetrikus az $x = 0$-ra, így világos, hogy a *mintaátlag:*

$$\bar{x} = 0.$$

Megkaphatjuk ezt az eredményt számítással is, ha figyelembe vesszük, hogy a megegyezés szerint a $-2{,}5$ osztályközepet 1-szer, a $-1{,}5$-et 2-szer, a $-0{,}5$-et 3-szor veszi fel a változó:

$$\bar{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \frac{-2{,}5 + 2(-1{,}5) + 3(-0{,}5) + 3 \cdot 0{,}5 + 2 \cdot 1{,}5 + 2{,}5}{12} = 0;$$



10*b* ábra

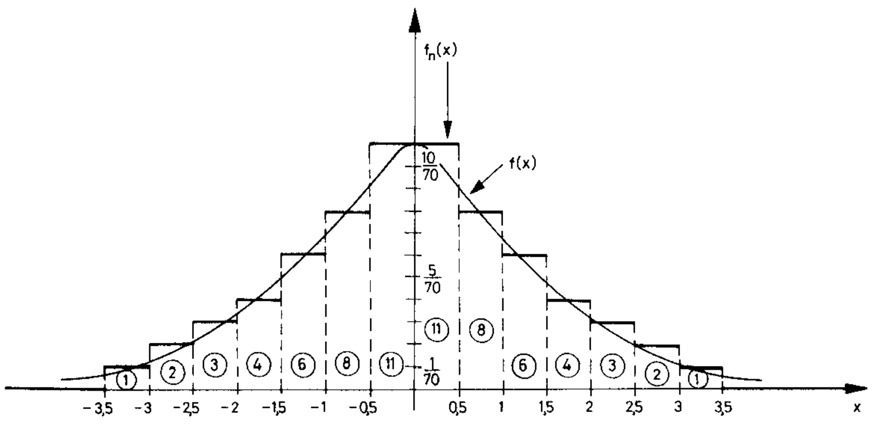
d) A tapasztalati korrigált szórás:

$$s^* = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} =$$

$$= \sqrt{\frac{(-2{,}5)^2+2(-1{,}5)^2+3(-0{,}5)^2+3\cdot 0{,}5^2+2\cdot 1{,}5^2+2{,}5^2}{11}} =$$

$$= 1{,}45 \text{ (század N).}$$

A vizsgálat során a súlyeltérés értékeit észrevétlenül $(-\infty)$-től $(+\infty)$-ig kiterjesztettük. Ez célszerű volt, mert ha nem 12 esetben ellenőrizzük a súlyeltérést, hanem pl. 70 esetben, akkor már nem -3 és $+3$ között változnak az eltérések, hanem esetleg nagyobb intervallumban. Egy $n=70$-elemű mintához tartozó tapasztalati sűrűségfüggvény látható a 10*c* ábrán, az osztályok terjedelme itt már nem 1, hanem 0,5.

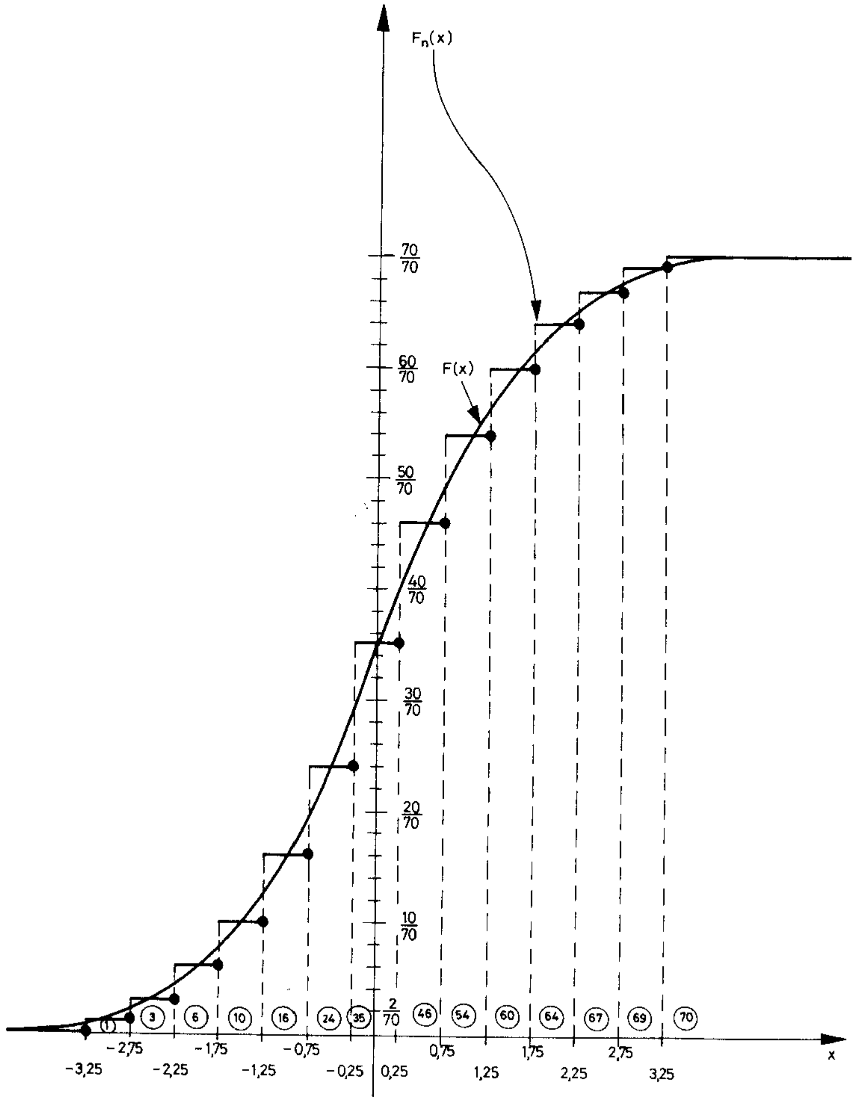


c)

10*c* ábra

Minél kisebb terjedelmű osztályokba sorolunk és minél nagyobb a mérésszám, a tapasztalati sűrűségfüggvény annál inkább egy y tengelyre szimmetrikus haranggörbét közelít meg (l. a 10*c* ábrán a folytonos görbével rajzolt függvényt). Úgy tekinthetjük, hogy az $f_n(x)$ véges sok mérésből származik, az $f(x)$ haranggörbe pedig a végtelen sok értéket felvevő X valószínűségi változó sűrűségfüggvénye.

Ugyanigy a 10*d* ábrán megszerkesztettük az $n=70$ mérési adatból származó $F_n(x)$ tapasztalati eloszlásfüggvényt. Ha az osztályok terjedelme egyre kisebb lesz és a mintaelemszám $n \to \infty$, a tapasztalati eloszlásfüggvény a 10*d* ábrán látható $F(x)$ eloszlásfüggvényhez tart nagy valószínűséggel.



d)

10*d* ábra

2. Az 1. feladatban mutatott mosóportöltési feladatnál az előírástól számított súlyeltérés jól közelíthető volt olyan normál eloszlással, amelynek várható értéke $m = \bar{x} = 0$, szórása $\sigma = s^* = 1{,}45$. Mennyi a valószínűsége annak, hogy a súlyeltérés *a)* (-1)-nél kisebb legyen; *b)* 1-nél nagyobb legyen?

---

a) $$P(X < -1) = F(-1) = \Phi\left(\frac{-1-0}{1{,}45}\right) = \Phi(-0{,}69) = 1 - \Phi(0{,}69) =$$
$$= 0{,}2451.$$

Kb. 24,5%-ban kapunk olyan mosóporos dobozt, amelyben az előíráshoz viszonyítva több mint 1 század N-nal kevesebb a betöltött mosópor.

b) Két megoldást is mutatunk: Ha (-1)-nél kisebb a változó 24,51%-ban, vagyis a haranggörbe alatti terület $X < -1$ esetén 0,2451, akkor a szimmetria miatt a haranggörbe alatti terület $X \geqq 1$ esetén is ugyanannyi lesz,

$$P(X \geqq 1) = 0{,}2451.$$

De számolhatunk az

$$F(x) = \Phi\left(\frac{x-m}{\sigma}\right)$$

összefüggéssel is.

$$P(X \geqq 1) = 1 - P(X < 1) = 1 - F(1) = 1 - \Phi\left(\frac{1-0}{1{,}45}\right) = 1 - \Phi(0{,}69) =$$
$$= 1 - 0{,}7549 = 0{,}2451.$$

3. A–50-es (MSZ 500-as) acél szakítási szilárdság ellenőrzésére az egész sokaságból $n = 31$ mérést végeztek. A mért értékek (N/mm^2-ben):
470 | 481, 483, 488, 489 | 490, 491, 492, 493, 493, 495, 496, 497, 498, 499 | 500, 501, 502, 503, 504, 505, 506, 507, 508, 509 | 512, 514, 516, 519 | 529, 530.

a) Mutassuk ki, hogy az adatok alapján a szakítószilárdság eloszlása *jól közelíthető normál eloszlással* (szerkesszük meg a tapasztalati sűrűségfüggvényt)!

b) Becsüljük meg az *egész sokaság m várható értékét* (az $\bar{x}$ mintaátlaggal), *az egész sokaság σ szórását* (az s^* korrigált tapasztalati szórással)!

c) Egy bizonyos munkánál selejtesnek tekintjük azt az acélt, amelynek szakítási szilárdsága 485 N/mm^2 alatt marad. Becsüljük meg *a selejt valószínűségét* (a Φ függvény segítségével)!

d) Végezzük el az előző becslést *az eloszlásfüggvény* megszerkesztésével, grafikus úton!

e) Az adatokat 470–480, 480–490, ..., 520–530 osztályokba sorolva, *helyettesítsünk minden adatot az osztályközéppel!* Becsüljük meg így a várható értéket és a korrigált tapasztalati szórást, szerkesszük meg így a tapasztalati eloszlásfüggvényt, becsüljük meg grafikus úton a selejt-valószínűséget!

---

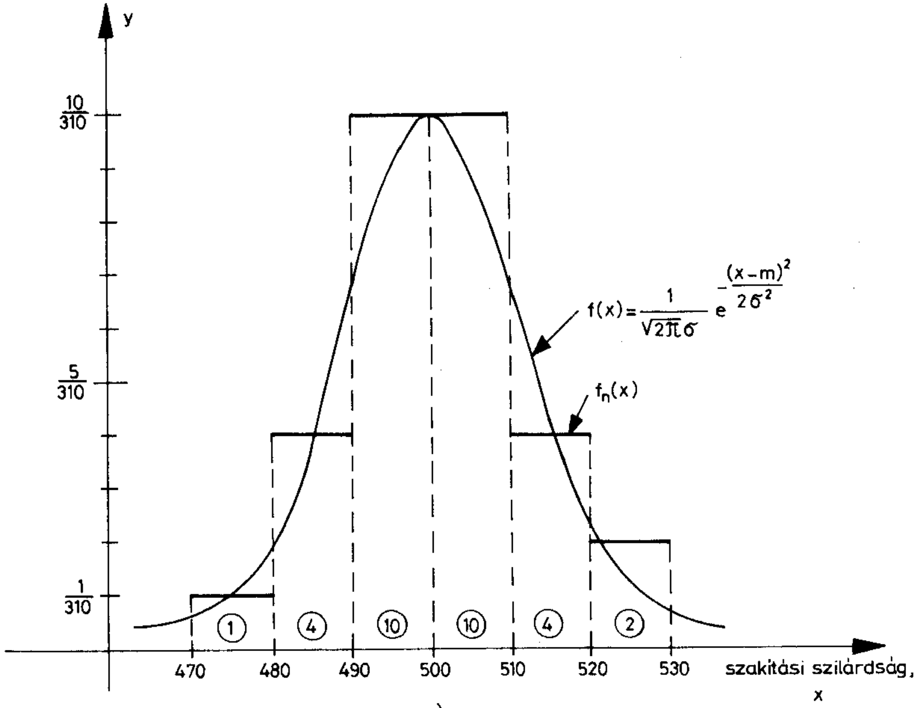
a) A növekvő sorrendbe rendezett adatokat tulajdonképpen már osztályokba soroltuk:

Osztályhatárok, N/mm^2	Gyakoriság
470–480	1
480–490	4
490–500	10
500–510	10
510–520	4
520–530	2

A gyakoriságértékeket $31 \cdot 10 = 310$-zel osztva, a tapasztalati sűrűségfüggvény (hisztogram) majdnem szimmetrikus, jól közelíthető haranggörbével. *A szakítási szilárdság eloszlása jól közelíthető normál eloszlással* (11*a* ábra).

b) A mérési értékek összege osztva a mérések számával adja a *mintaátlagot* (az m várható érték becslését):

$$m \approx \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{470 + 481 + \ldots + 530}{31} = 500{,}645.$$



a)

11*a* ábra

Az egész sokaság szórására jó becslés *a korrigált tapasztalati szórás:*

$$\sigma \approx s^* = \sqrt{\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} =$$

$$= \sqrt{\frac{(470-500{,}645)^2+\ldots+(530-500{,}645)^2}{30}} = 13{,}144.$$

Feltűnő, hogy a szórás milyen kicsi az átlaghoz képest:

$$\frac{s^*}{\bar{x}} = \frac{13{,}144}{500{,}645} = 0{,}026,$$

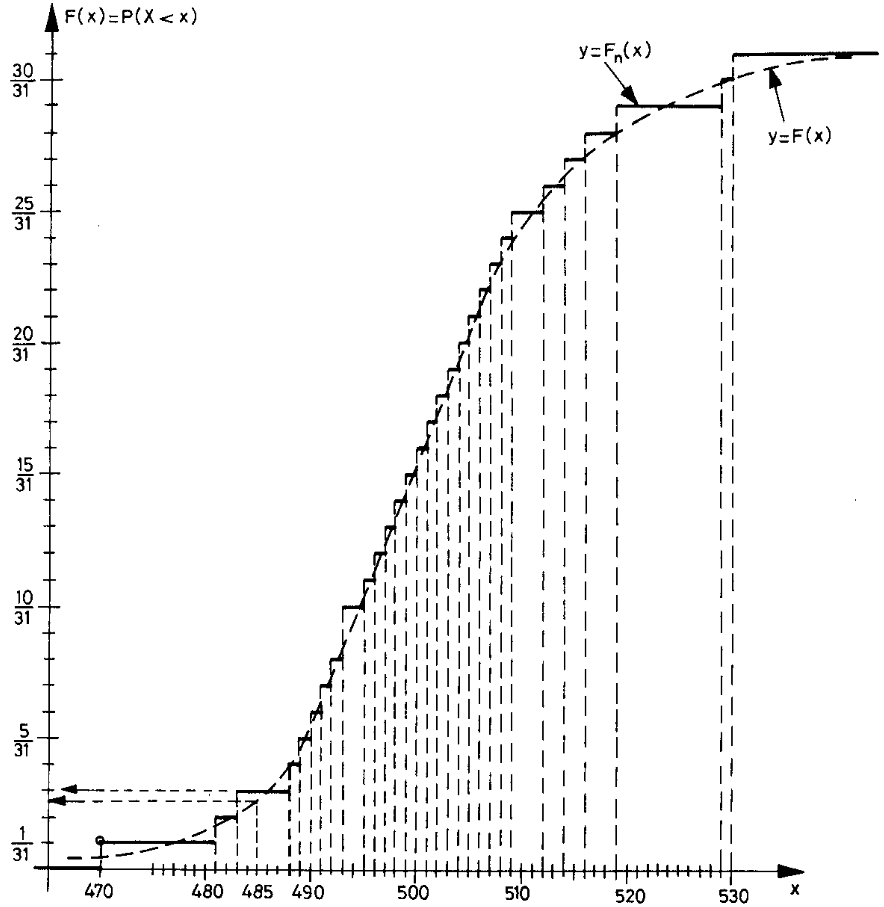
a 2,6%-os *relatív szórás azt mutatja, hogy a szakítási szilárdság értékei kevéssé szóródnak a várható érték körül.*

c) A szakítási szilárdság eloszlását normálissal közelítve, *a selejt-valószínűség becsülhető* a $\varnothing$ függvény segítségével:

$$P(X<485) = F(485) = \varnothing\left(\frac{485-500{,}645}{13{,}144}\right) =$$
$$= \varnothing(-1{,}19) = 1-\varnothing(1{,}19) = 1-0{,}8830 = 0{,}1170.$$

11,7%-*ra becsülhetjük a selejt valószínűségét,* ha *az eloszlást normálissal közelítjük.*

d) A tapasztalati eloszlásfüggvény a mért értékeknél $\frac{1}{31}$ valószínűséggel ugrik (a kétszer kapott 493-nál $\frac{2}{31}$ valószínűséggel). A kapott $F_n(x)$ függvény jól közelíti a normál eloszlás eloszlásfüggvényét, $F(x)$-et (l. a 11*b* ábrát).



11*b* ábra

b)

Grafikus úton megbecsülve annak a valószínűségét, hogy a szakítószilárdság 485 N/mm^2 alatt marad, *az eloszlásfüggvény $x=485$ abszcisszánál vett értékét kell leolvasnunk.* Ha a tapasztalati $F_n(x)$ függvényről olvassuk le, akkor

$$P(X<485) = F_n(485) = \frac{3}{31} \approx 0{,}0968,$$

ha pedig a normál eloszlás folytonos eloszlásfüggvényéről, $F(x)$-ről, akkor

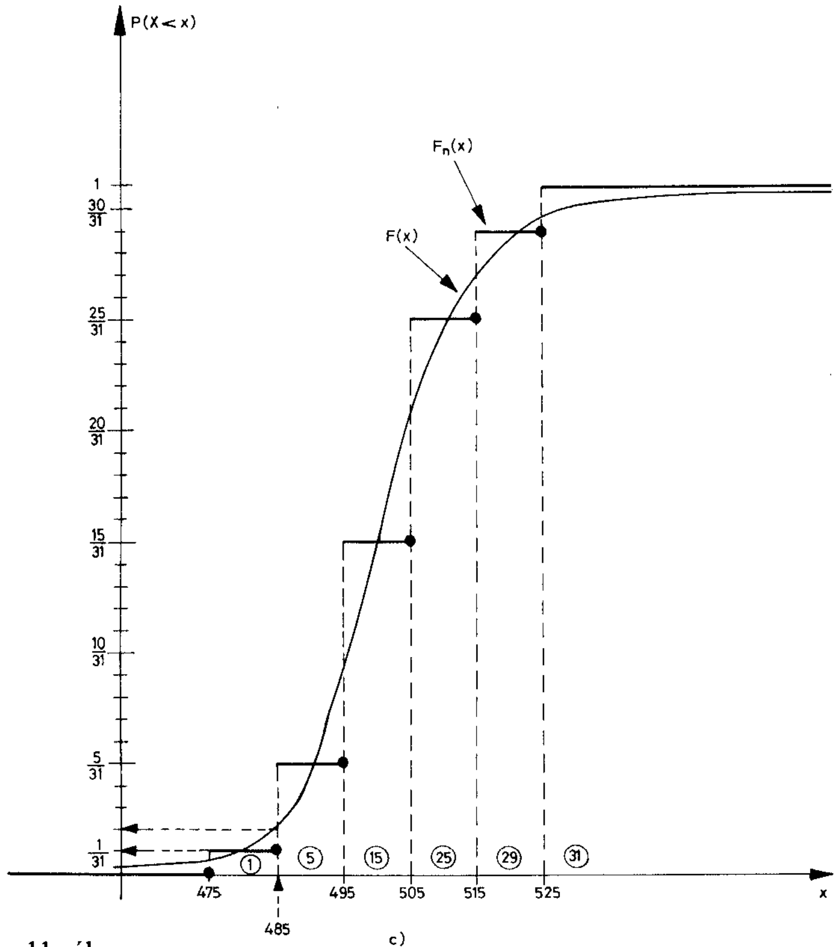
$$P(X<485) = F(485) = \frac{2{,}6}{31} \approx 0{,}0839.$$

e) Hosszadalmas az egyes mért értékekből megszerkeszteni az $F_n(x)$ tapasztalati eloszlásfüggvényt, ha az adatok száma nagy. Ilyenkor az adatokat osztályba sorolva, *a mért adatokat az osztályközepekkel helyettesítjük.*

Segítségképpen táblázatba foglaljuk az adatokat, az osztályközepeket és azok p_i felvételi valószínűségeit, valamint az $F_n(x_i) = P(X<x_i) = \sum_{X<x_i} p_i$ eloszlásfüggvény-értékeket:

Osztály-határok	Osztály-közép, x_i	Gyakori-ság, f_i	Valószínű-ség, p_i	Eloszlásfüggvény, $F_n(x_i) = \Sigma p_i$
470–480	475	1	1/31	$F_n(475) = 0$
480–490	485	4	4/31	$F_n(485) = 1/31$
490–500	495	10	10/31	$F_n(495) = 5/31$
500–510	505	10	10/31	$F_n(505) = 15/31$
510–520	515	4	4/31	$F_n(515) = 25/31$
520–530	525	2	2/31	$F_n(525) = 29/31$
				$F_n(526) = 31/31$

Úgy tekintjük, mintha a változó a 475 értéket 1-szer, $p_1 = \frac{1}{31}$ valószínűséggel, a 485 értéket 4-szer, $p_2 = \frac{4}{31}$ valószínűséggel stb. vette volna fel. Így könnyebben kaphatjuk meg az $F_n(x)$ *tapasztalati eloszlásfüggvényt* (l. a 11*c* ábrát, ami természetesen durvább becslés a normál eloszlás $F(x)$ *eloszlásfüggvényére,* mint az előbbi.



11*c* ábra

A selejt-valószínűség becslése a tapasztalati eloszlásfüggvénnyel:

$$P(X<485) \approx F_n(485) = \frac{1}{31} \approx 0{,}032,$$

ill. az $F(x)$ eloszlásfüggvénnyel:

$$P(X<485) \approx F(485) = \frac{2}{31} \approx 0{,}065.$$

A várható érték becslése a mintaátlaggal, *a szórás becslése* a korrigált tapasztalati szórással most így történhet:

Osztályközép, x_i	Gyakoriság, f_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$
475	1	−25,8	665,64
485	4	−15,8	249,64
495	10	−5,8	33,64
505	10	4,2	17,64
515	4	14,2	201,64
525	2	24,2	585,64

$$\bar{x}=\frac{1\cdot 475+4\cdot 485+10\cdot 495+10\cdot 505+4\cdot 515+2\cdot 525}{31}=500{,}806,$$

$$s^*=\sqrt{\frac{1\cdot 665{,}64+4\cdot 249{,}64+10\cdot 33{,}64+10\cdot 17{,}64+4\cdot 201{,}64+2\cdot 585{,}64}{30}}=$$

$$=11{,}768.$$

A most kapott értékek természetesen durvább becslések, mint az előzők, viszont rövidebben megkaphatók, mint azok.

Megjegyzések

a) A szakítási szilárdság eloszlását a normál eloszlás csak *jól közelíti,* hiszen a szilárdság negatív nem lehet. (Negatív x-ekre a közelítésül vett normál eloszlás sűrűségfüggvénye az x tengelyhez lapul, az alatta levő terület majdnem 0.)

b) A gyakorlatban *nagyon sok valószínűségi változó eloszlása közelíthető jól normál eloszlással,* pl. szakítószilárdság, méretek, mérési értékek előírástól való eltérése, élettartam, átlagérték stb. A normál eloszlás gyakori előfordulását többek között a központi határeloszlás-tétel indokolja: azonos eloszlású, független valószínűségi változók összege határértékben (vagyis ha $n\to\infty$) normál eloszlású.

A szakítószilárdság gyakran az elemi szálak egymástól független szakítószilárdságának az összege, ezért gyakran lesz közelítőleg normál eloszlású. Az átlagérték n db azonos eloszlású, független változó összegének $\frac{1}{n}$-szerese, $\bar{x}=\frac{x_1+x_2+\ldots+x_n}{n}$, ezért lesz nagy n-re közelítőleg normál eloszlású.

c) Óvakodni kell attól a gyakori tévedéstől, hogy ha egy változónál egyszer normál eloszlást tételezhetünk fel a mérési eredmények alapján, akkor az *mindig* normál eloszlású. Ezt a feltevést mindig meg kell vizsgálnunk a mérési eredmények alapján.

4. Egy bizonyos típusú elektroncső élettartamára vonatkozó követelmény, hogy a 950 óránál kisebb élettartamú csövek aránya legfeljebb 4% lehet. Az élettartam jó közelítésben normál eloszlású. Egy adott időszakban több ezer cső élettartamát 250 elemű véletlen minta alapján ellenőrzik. A vizsgálat eredménye (osztályba sorolással):

Osztályok, óra	Gyakoriság, f_i
850,5– 900,5	10
900,5– 950,5	16
950,5–1000,5	28
1000,5–1050,5	40
1050,5–1100,5	60
1100,5–1150,5	45
1150,5–1200,5	25
1200,5–1250,5	17
1250,5–1300,5	9

a) Állapítsuk meg, hogy eleget tesz-e a gyártott tétel a minőségi követelményeknek!

b) Ha a gyártott mennyiség tényleges átadásakor az átvevő csak akkor veszi át a tételt, ha 50-elemű véletlen mintában legfeljebb 3 db 950 óránál kisebb élettartamú csövet talál, akkor mekkora lesz egy tétel átvételének valószínűsége az előző adatok alapján?

c) Ha a következő gyártási időszakban 1050 óra átlagos élettartamot kívánnak biztosítani, és a minőségi előírásnak is eleget akarnak tenni (950 óra alattiak aránya legfeljebb 4%), akkor mekkora relatív szórás engedhető meg?

a) Nem, mert $P(X<950)=0{,}0951$; *b)* $np=5$, az átvétel valószínűsége közelítve 0,265; *c)* 5,44%.

5. Tömeggyártásban készített extrudált rudacskák fontos jellemzője a *külső átmérő,* amelyre vonatkozó minőségi előírás: 8,3 ± 0,1 mm. Az átmérő ellenőrzésére a többezres darabszámú tételből $n = 145$ elemű mintát vettek (0,01 mm pontossággal mérve).

Az adatokat osztályokba sorolva rendezték. Tekintsük úgy, hogy egy-egy osztályban az átmérő értéke az osztályhatárok számtani közepe (osztályközép)!

a) Számítsuk ki az átmérő *átlagát, szórását, relatív szórását!*

b) A *sűrűségfüggvény* megszerkesztésével mutassuk ki, hogy az átmérő eloszlása jól közelíthető normális eloszlással!

c) Szerkesszük meg az eloszlásfüggvényt! Becsüljük meg ennek segítségével, valamint a $\Phi(x)$ függvény segítségével, mennyi a *várható selejtszázalék* (a 8,2 alá, ill. a 8,4 fölé eső átmérő selejtet jelent)!

Osztályhatárok, mm	Osztályközép, x_i	Gyakoriság, f_i
8,125–8,155	8,14	2
8,155–8,185	8,17	3
8,185–8,215	8,20	4
8,215–8,245	8,23	5
8,245–8,275	8,26	14
8,275–8,305	8,29	21
8,305–8,335	8,32	55
8,335–8,365	8,35	23
8,365–8,395	8,38	7
8,395–8,425	8,41	6
8,425–8,455	8,44	2
8,455–8,485	8,47	1
8,485–8,515	8,50	2
Összesen		$145 = n$

---

a) $\bar{x} = 8{,}314$ mm, $s^* = 0{,}059$, $\frac{s^*}{\bar{x}} = 0{,}0071$ (igen kicsi!); *b)* a sűrűségfüggvény haranggörbe; *c)* $P(\text{átmérő} < 8{,}2) = 0{,}0268$, $P(\text{átmérő} > 8{,}4) = 0{,}0721$, selejt-valószínűség az előbbiek összege, kb. 10% (ha az elméleti eloszlásfüggvénnyel számolunk, ez a megbízhatóbb). Ha a tapasztalati eloszlásfüggvénnyel becslünk, a selejt-valószínűség kb. 7%.

6. C10 5-órás diffúziósan izzított acélnál a hajlító határfeszültségre a következő értékeket mérték (10 N/mm²-ben):

167,1	171,6	174,1	153,2	171,6	226,3
167,1	142,7	188,0	174,1	146,2	174,1
139,3	111,4	149,7	167,1	164,6	184,5

a) Mutassuk ki a *sűrűségfüggvény* megszerkesztésével, hogy a hajlító határfeszültség jó közelítésben normális eloszlású!

b) Számítsuk ki az *átlagot, a szórást és a relatív szórást!*

c) A $\Phi(x)$ függvény segítségével becsüljük meg, mekkora valószínűséggel nagyobb 140-nél a hajlító határfeszültség!

---

a) A sűrűségfüggvény haranggörbe; *b)* $\bar{x} = 165{,}15$, $s^* = 23{,}98$, $\frac{s^*}{\bar{x}} = 0{,}15$; *c)* 0,8531.

7. Kétféle fonal fonalfinomsági számának eloszlását vizsgálták:

Nm 85 fonalfinomsági számosztályok	Gyakoriság	Nm 120 fonalfinomsági számosztályok	Gyakoriság
77, –78,5	1	118–120	1
78,5–80,	6	120–122	4
80, –81,5	7	122–124	12
81,5–83,	10	124–126	16
83, –84,5	6	126–128	15
84,5–86,	13	128–130	8
86, –87,5	8	130–132	3
87,5–89,	9	132–134	1

Melyik esetben állíthatjuk, hogy a fonal finomsági száma normális eloszlással közelíthető? A második esetben milyen intervallumba esnek 99,7% valószínűséggel a finomsági szám értékei?

A második esetben látszik, hogy az eloszlás jól közelíthető normálissal, $\bar{x}=125{,}7$, $s^*=2{,}848$, a finomsági szám pl. $\bar{x}-3s^* = 117{,}156$ és $\bar{x}+3s^* = 134{,}244$ közé esnek 99,7%-os (nagy) valószínűséggel. (Logikus, hogy szimmetrikus eloszlásnál a várható érték becslésére, $\bar{x}$-ra szimmetrikus intervallumot kerestünk. De nem csak $\bar{x}$-ra szimmetrikus intervallum lehet a megoldás! Számtalan sok megoldás van.)

8. Egy ipari robot kezének egy előírt egyenestől való távolságára a minőségi előírás: $(12{,}8 \pm 1)$ mm. A mért távolságokat osztályokba sorolva gyakoriságukkal együtt a következő táblázatban láthatjuk:

Osztályhatárok, mm	Gyakoriság
10–11	1
11–12	5
12–13	13
13–14	4
14–15	2

a) Sűrűséghisztogram megszerkesztésével mutassuk ki, hogy a mért adatok eloszlása jó közelítéssel normál eloszlás!

b) Határozzuk meg az adatok átlagát, korrigált szórását és relatív szórását!

c) Várhatóan milyen százalékban teljesültek a minőségi előírások?

---

a) A sűrűségfüggvényre jól illeszthető haranggörbe, az eloszlás jól közelíthető normálissal; *b)* $\bar{x}=12{,}54$, $s^*=0{,}93$, $\frac{s^*}{\bar{x}} = 0{,}07$; *c)* $F(13{,}8)-F(11{,}8) = \Phi(1{,}35)-\Phi(-0{,}8) = 0{,}6996 \approx 70\%$.

9. Igazoljuk, hogy az n-elemű minta átlagának várható értéke az eredeti változó várható értéke, a mintaátlag szórása pedig az eredeti változó szórásának $\sqrt{n}$-edrésze:

$$M(\bar{X})=M(X), \qquad D(\bar{X}) = \frac{D(X)}{\sqrt{n}}.$$

A mintaelemek független, az eredeti változóval azonos várható értékű és szórású (azonos eloszlású) változók. Ezt felhasználva:

$$M(\bar{X}) = M\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} M(\Sigma X_i) = \frac{1}{n}\Sigma M(X_i) = M(X)$$

(itt a függetlenség nem is szükséges). A szórásnégyzet kiszámításánál azonban már szükséges:

$$D^2(\bar{X}) = D^2\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2} D^2(\Sigma X_i) = \frac{1}{n^2}\Sigma D^2(X_i) = \frac{1}{n^2} n D^2(X) = \frac{D^2(X)}{n},$$

ebből valóban

$$D(\bar{X}) = \frac{D(X)}{\sqrt{n}}.$$

10. Egy elektroncső élettartama a véletlentől függő valószínűségi változó. Várható értéke (egyúttal előírt érték) $m=580$ óra, szórása $\sigma=30$ óra, $n=100$ elemű mintából az átlag $\bar{x}=573$ óra, kevesebb, mint az egész sokaság átlagára előírt 580 óra. Feltehető-e, hogy az egész sokaság átlaga 580 óra (tehát a csökkenés csak véletlen következménye)?

Ha a feltevés igaz, akkor az élettartamátlag közelítőleg normális eloszlású-nak tekinthető, $m = 580$ várható értékkel,

$$\sigma_{100} = \frac{\sigma}{\sqrt{100}} = \frac{30}{\sqrt{100}} = 3 \text{ óra szórással.}$$

95% valószínűséggel a normál eloszlású változó olyan $m-x$, $m+x$ (m-re szimmetrikus) intervallumba esik, amelyre nézve:

$$P(m-x \leqq X < m+x) = F(m+x)-F(m-x) =$$
$$= \Phi\left(\frac{x}{\sigma_{100}}\right)-\Phi\left(-\frac{x}{\sigma_{100}}\right) = 2\Phi\left(\frac{x}{\sigma_{100}}\right)-1 = 0{,}95,$$

ebből

$$\Phi\left(\frac{x}{\sigma_{100}}\right) = 0{,}9750,$$

$$\frac{x}{\sigma_{100}} = 1{,}96, \quad x = 1{,}96\sigma_{100} = 1{,}96 \cdot 3 = 5{,}88.$$

Ha tehát az $m = 580$ hipotézis igaz, akkor 95% a valószínűsége, hogy az élettartamátlag

$$580-5{,}88 = 574{,}12 \quad \text{és} \quad 580+5{,}88 = 585{,}88$$

közé esik.

Mivel 573 a mondott intervallumon kívül esik (és erre csak 2,5% a valószí-nűség), ezért 97,5% biztonsággal kijelenthetjük, hogy m nem az előírt 580 óra, nem véletlen az elektroncső-élettartamátlag csökkenése a mintában, vagyis feltehetően romlott a minőség.

11. Egy elektroncső élettartamának a várható értéke a szabvány szerint $m = 580$ óra, $\sigma = 30$ óra szórással. $n = 225$ elemű mintában az élettartam átlaga $\bar{x} = 584$ óra volt. Állíthatjuk-e 99% biztonsági szinten, hogy az élettartam átlagában lényeges javulás történt?

Tegyük fel, hogy nem, vagyis $m = 580$ továbbra is. P (élettartam $< m+x$) =

$$= F(m+x) = \Phi\left(\frac{m+x-m}{\frac{\sigma}{\sqrt{n}}}\right) = \Phi\left(\frac{x}{\frac{30}{15}}\right) = 0{,}99, \ \frac{x}{2} = 2{,}33, \ x = 4{,}66. \text{ Tehát}$$

99% valószínűséggel 584,66 óra alá esik az élettartamátlag, lényeges (szignifi-káns) javulásra nem következtethetünk.

12. Igazoljuk a mintaátlag és a korrigált szórásnégyzet választott középpel való kiszámításának képletét:

$$\bar{x} = c + \frac{\sum_{i=1}^{n}(x_i-c)}{n},$$

$$s^{*2} = \frac{1}{n-1}\left\{\sum_{i=1}^{n}(x_i-c)^2 - \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-c)\right]^2\right\}.$$

A mintaátlagra adott képlet bizonyítása:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{\Sigma[(x_i-c)+c]}{n} = \frac{\Sigma(x_i-c)}{n} + \frac{nc}{n} = c + \frac{\Sigma(x_i-c)}{n}.$$

A korrigált szórásnégyzetre adott képlet igazolása:

$$s^{*2} = \frac{\Sigma(x_i-\bar{x})^2}{n-1} = \frac{\Sigma[(x_i-c)-(\bar{x}-c)]^2}{n-1} =$$
$$= \frac{1}{n-1}\{\Sigma(x_i-c)^2 - 2(\bar{x}-c)\,\Sigma(x_i-c) + n(\bar{x}-c)^2\} =$$
$$= \frac{1}{n-1}\left\{\Sigma(x_i-c)^2 - 2\frac{\Sigma(x_i-c)}{n}\Sigma(x_i-c) + \frac{[\Sigma(x_i-c)]^2}{n}\right\} =$$
$$= \frac{1}{n-1}\left\{\Sigma(x_i-c)^2 - \frac{1}{n}[\Sigma(x_i-c)]^2\right\}.$$

Megjegyzés

Érdemes megfigyelni, hogy a szórásnégyzetre kapott képlet korrekció nélkül (tehát ha nem $(n-1)$-gyel, hanem n-nel osztanánk az elején) a

$$D^2(X) = M(X^2) - M^2(X)$$

képlettel azonos szerkezetű. Ez érthető, mert a kapott szummák a várható értékeknek a mintából kapott becslései.

13. Számítsuk ki a következő adatok átlagát és korrigált szórását választott közép segítségével:

1243,1; 1244,2; 1242,7; 1244,3; 1242,6.

Az adatokat áttekintve látszik, hogy azok $c=1243$ egész szám körül szóródnak. Ez legyen a választott közép!

Adatok, x_i	Eltérések $c=1243$ választott középtől, x_i-c	Eltérések négyzete, $(x_i-c)^2$
1243,1	+0,1	0,01
1244,2	+1,2	1,44
1242,7	−0,3	0,09
1244,3	+1,3	1,69
1242,6	−0,4	0,16
Összesen:	$+1{,}9 = \Sigma(x_i-c)$	$3{,}39 = \Sigma(x_i-c)^2$

Az átlag:

$$\bar{x} = c + \frac{\Sigma(x_i-c)}{n} = 1243 + \frac{1{,}9}{5} = 1243{,}38,$$

a korrigált szórásnégyzet:

$$s^{*2} = \frac{1}{n-1}\left[\Sigma(x_i-c)^2 - \frac{1}{n}(\Sigma(x_i-c))^2\right] =$$

$$= \frac{1}{4}\left[3{,}39 - \frac{1}{5}1{,}9^2\right] = 0{,}667,$$

a korrigált szórás:

$$s^* = \sqrt{0{,}667} = 0{,}8167.$$

14. Igazoljuk a mintaátlagra és korrigált szórásnégyzetre adott

$$\bar{x} \approx D + \frac{k}{n}\sum_{i=1}^{r} f_i z_i,$$

$$s^{*2} \approx \frac{k^2}{n-1}\left[\sum_{i=1}^{r} f_i z_i^2 - \frac{1}{n}\left(\sum_{i=1}^{r} f_i z_i\right)^2\right]$$

képleteket! Egyszerűség kedvéért soroljuk most az adatokat $r=5$ osztályba!

Tüntessük fel, hogy az osztályközepek mennyivel térnek el a középső $D=x_3$ osztályközéptől (k az osztályok terjedelme):

$$x_1 = D-2k,\; x_2 = D-k,\; x_3 = D,\; x_4 = D+k,\; x_5 = D+2k.$$

A választott közép itt most D, az adatokat az osztályközepek helyettesítik. Jelentse f_i az i-edik osztályba eső adatok számát (gyakoriságát)! A *mintaátlag* választott középpel dolgozva:

$$\bar{x} \approx D + \frac{\sum_{i=1}^{5} f_i(x_i-D)}{n} =$$

$$= D + \frac{f_1(x_1-D)+f_2(x_2-D)+f_3(x_3-D)+f_4(x_4-D)+f_5(x_5-D)}{n} =$$

$$= D + \frac{f_1(-2k)+f_2(-k)+f_3(0\cdot k)+f_4\cdot k+f_5\cdot 2k}{n} =$$

$$= D + \frac{k}{n}\,\frac{f_1(-2)+f_2(-1)+f_3\cdot 0+f_4\cdot 1+f_5\cdot 2}{n} =$$

$$= D + \frac{k}{n}\sum_{i=1}^{5} f_i z_i,$$

hiszen a z_i értékeket így választottuk:

$$z_1 = -2,\ z_2 = -1,\ z_3 = 0,\ z_4 = 1,\ z_5 = 2.$$

A *korrigált szórás* négyzete választott közép felhasználásával:

$$s^{*2} \approx D + \frac{1}{n-1}\left\{\sum_{i=1}^{5} f_i(x_i - D)^2 - \frac{1}{n}\left[\sum_{i=1}^{5} f_i(x_i - D)\right]^2\right\}.$$

Láttuk már az előbb, hogy

$$x_1 - D = -2k = z_1k,\ x_2 - D = -k = z_2k, \ldots,$$

általában

$$x_i - D = z_ik.$$

Ezt a képletbe helyettesítve:

$$s^{*2} \approx D + \frac{1}{n-1}\left\{\sum_{i=1}^{5} f_iz_i^2k^2 - \frac{1}{n}\left[\sum_{i=1}^{5} f_iz_ik\right]^2\right\} =$$

$$= D + \frac{k^2}{n-1}\left\{\sum_{i=1}^{5} f_iz_i^2 - \frac{1}{n}\left(\sum_{i=1}^{5} f_iz_i\right)^2\right\}.$$

Megjegyzések

1. $\bar{x}$-ra és s^{*2}-re azért kaptunk csak közelítő értéket, mert az adatokat az osztályközepekkel helyettesítettük.

2. A bizonyítás $r = 5$ osztály helyett ugyanígy elvégezhető tetszőleges számú osztály esetén.

3. $\bar{x}$ és s^{*2} itt is véletlen változók. Itt most éltünk azzal a statisztikai gyakorlattal, hogy ezt nem tüntettük fel. Ennek az az oka, hogy olyankor is lehet átlagot és szórást számolni, ha nem egy véletlen változónak az értékeit mértük, hanem egy determinisztikus változóra egyetlen adatsorunk van.

15. A textiliparban az elemi szál szakítóvizsgálatánál 100 méréssel a következő szakadási értékeket kapták század N-ban, mindjárt nagyság szerinti sorrendbe rakva:

| 13; | 15; 16; | 17; 17; 17; 17; 17; 18; 18; | 19; 19; 19; 19; 19; 20; 20; 20; 20; 20; 20; 20; 20; | 21; 21; 21; 21; 21; 21; 21; 21; 22; 22; 22; 22; 22; 22; 22; 22; 22; | | 23; 23; 23; 23; 23; 23; 23; 23; 23; 23; 23; 24; 24; 24; 24; 24; 24; 24; | 25; 25; 25; 25; 25; 25; 25; 25; 25; 25; 25; 25; 25; 25; 25; 26; 26; 26; 26; 26; 26; | 27; 27; 27; 27; 27; 27; 27; 27; 28; 28; 28; 28; | 29; 29; 30; 30; 30; 30; 30; 30; | 32. |

Osztályba sorolással állapítsuk meg a szakítóerő várható értékét (középértékét), szórását, %-os szórását!

Az $n = 100$ adatot 13-tól 33-ig $k = 2$ század N-onként osztályba soroljuk (a függőleges elválasztójel jelzi az egyes osztályba jutottakat). Az osztályok száma $r = 10$.

Az i-edik osztályba jutó elemek számát (gyakoriságát) f_i jelöli (pl. $f_1 = 1$), az osztály számtani közepét x_i (pl. $x_1 = 14$).

Az $x_1 = 14$, $x_2 = 16, \ldots, x_{10} = 32$ osztályközepek közül válasszuk ki középsőnek pl. $D = 24$-et, az ezzel kiválasztott osztály sorszáma 0, a táblázatban fölötte állóké $-1, -2, -3, -4, -5$, a táblázatban alatta állóké $+1, +2, +3, +4$ (ezek a z_i értékek).

Az adatokat táblázatba foglalva:

Osztály	Osztály-közép, x_i	Gyakoriság, f_i	z_i	f_iz_i	$f_iz_i^2$
13–15	14	1	−5	−5	25
15–17	16	2	−4	−8	32
17–19	18	7	−3	−21	63
19–21	20	13	−2	−26	52
21–23	22	17	−1	−17	17
23–25	24 = D	18	0	0	0
25–27	26	21	+1	+21	21
27–29	28	12	+2	+24	48
29–31	30	8	+3	+24	72
31–33	32	1	+4	+4	16
Összeg:		100		$\sum_{i=1}^{10} f_iz_i = -4$	$\sum_{i=1}^{10} f_iz_i^2 = 346$

A mért adatok számtani közepe:

$$\bar{x} \approx D + \frac{k}{n}\sum_{i=1}^{r} f_iz_i = 24 + \frac{2}{100}(-4) = 23{,}92.$$

A mért adatok szórásnégyzete:

$$s^2 \approx \frac{k^2}{n-1}\left[\sum_{i=1}^{r} f_i z_i^2 - \frac{1}{n}\left(\sum_{i=1}^{r} f_i z_i\right)^2\right] =$$

$$= \frac{2^2}{99}\left[346 - \frac{1}{100}(-4)^2\right] = \frac{1383{,}36}{99}.$$

A szórás:

$$s \approx \sqrt{\frac{1383{,}4}{99}} = 3{,}72.$$

A %-os szórás a szórásnak a középértékhez mért aránya %-ra átszámolva, tehát a százalékos szórás:

$$\frac{s}{\bar{x}} = \frac{3{,}72}{23{,}92} = 0{,}155 = 15{,}5\%.$$

16. Azonos gyártási feltételek mellett előállított, adott hosszúságú szövetmintákon a szövethibák száma:

A hibák száma, X	0	1	2	3	4	5	6
Hányszor fordult elő a hiba, f_i	327	340	160	53	16	3	1

Az előző tapasztalati értékek alapján állíthatjuk-e, hogy a hibák száma Poisson-eloszlású?

A szövethibák átlaga:

$$\frac{0 \cdot 327 + 1 \cdot 340 + 2 \cdot 160 + 3 \cdot 53 + 4 \cdot 16 + 5 \cdot 3 + 6 \cdot 1}{327 + 340 + 160 + 53 + 16 + 3 + 1} = 1.$$

Ez torzítatlan becslés az egész sokaság várható értékére. Hasonlítsuk össze az egész sokaságban (amelynél Poisson-eloszlást tételezünk fel $\lambda = 1$ várható értékkel) és a mintában a 0, 1, 2, 3, 4, 5, 6 hiba valószínűségét:

A hibák száma	0	1	2	3	4	5	6
Valószínűség a mintában	0,3633	0,3778	0,1778	0,0589	0,0178	0,0033	0,0011
Valószínűség az egész sokaságban (l. a 3. táblázatot)	0,3679	0,3679	0,1839	0,0613	0,0153	0,0031	0,0005

A valószínűségek nagyon kevéssé térnek el, feltételezhető, hogy a szövethibák száma Poisson-eloszlású. (Feltehető a kérdés: milyen eltérés esetén fogadjuk el, mikor vetjük el a feltevést? Erre egzaktabb módszer lesz a χ^2-próba.)

17. Hogyan kell *átlagot és szórást számítani* a HT PTK–1050 *zsebszámológéppel?*

Az *átlag* ($\bar{x}$) és *korrigálatlan szórás* (s^2, a gépen látható jelölés σ^2) számítására *beépített áramkör van a zsebszámológépen* (mint nagyon sok más, korszerű zsebszámológépen). Az $x_1, x_2, \ldots, x_n$ adatok *bevitele* így történik:

x_1 2nd Σ^+
x_2 2nd Σ^+
- - - - - -
x_n 2nd Σ^+

Ha egy adat többször fordul elő, minden alkalommal be kell nyomnunk az adat értékét is a 2nd Σ^+ billentyűzés előtt (egyes gépeken az adatot nem kell még egyszer beütni, elég csak a Σ^+ billentyűt többször megnyomni).

Ennek a billentyűzésnek a hatására az egyes tárak így töltődnek fel:

Tár	*Tartalom*
0.	$M0=n$
1.	$M1=\Sigma x_i$
2.	$M2=\Sigma x_i^2$

A 2nd $\bar{x}$ billentyűzésre megjelenik az adatok átlaga, a 2nd σ^2 billentyűzésre pedig a korrigálatlan szórás négyzete. A „szubrutint" (az integrált áramkört) nyilván úgy csinálták meg, hogy az a

$$\frac{\Sigma x_i}{n}=\frac{M1}{M0},\quad \frac{\Sigma x_i^2}{n}-\left(\frac{\Sigma x_i}{n}\right)^2=\frac{M2}{M0}-\left(\frac{M1}{M0}\right)^2$$

műveletek elvégzésére legyen képes.

Példa. Adatok (az x_i értékek): 1, 2, 3.

Billentyűzés	Az utasítások hatása
1 2nd Σ^+ 2 2nd Σ^+ 3 2nd Σ^+	az adat sorszáma, ami végül n a kijelzőn és a 0. tárban; Σx_i az 1. tárban; Σx_i^2 a 2. tárban
2nd $\bar{x}$	$\bar{x}=2=$ átlag
2nd σ^2 $\sqrt{x}$	$\sigma_n=0{,}8165=$ korrigálatlan szórás
x(3 ÷ 2) $\sqrt{x}$=	$\sigma_{n-1}=1=$ korrigált szórás

Nagy tömegű adat esetén előfordulhat, hogy egy adatot rosszul adunk meg. *Hibás adat törlése* így történik. Ha 1, 2, 3 helyett véletlenül 1 2nd Σ^+ 2 2nd Σ^+ 4 2nd Σ^+ volt a billentyűzés, akkor 4 INV 2nd Σ^+ hatására a 4 törlődik, a kijelző 2-t jelez, és a tárak tartalma is korrigálódott.

3 2nd Σ^+ a megfelelő befejezés.

Megjegyzés

Több zsebszámológépen a statisztikai áramkört előbb be kell kapcsolni (a Statistical Decisions = SD gombbal, a Statistical Mode beállításával stb.).

18. Adjunk folyamatábrát, BASIC programot az átlag- és szórásszámításra!

a) Ha az

$$\bar{x}=\frac{\Sigma x_i}{n},\quad s^{*2}=\frac{\Sigma x_i^2}{n-1}-\frac{n}{n-1}\bar{x}^2$$

képletekkel dolgozunk, akkor

– az x_i értékeket egyetlen X tárba tehetjük el (takarékosság a tárban);

– az x_i és x_i^2 értékek összegezését egy ciklusban megszervezhetjük (futási idő csökkentése).

Történjék az x_i értékek összegezése az SU tárban, az x_i^2 értékek összegezése az SN tárban. Az összegező tárakat először 0-ra kell állíttatni. A ciklusban a SU tár új tartalma úgy keletkezik a régiből, hogy ahhoz az új x_i (röviden: X) értéket hozzáadjuk:

$$SU \Leftarrow SU+X,$$

ugyanígy

$$SN \Leftarrow SN+X^2.$$

A ciklusban 3 lépést kell $i=1$-től n-ig megismételni:

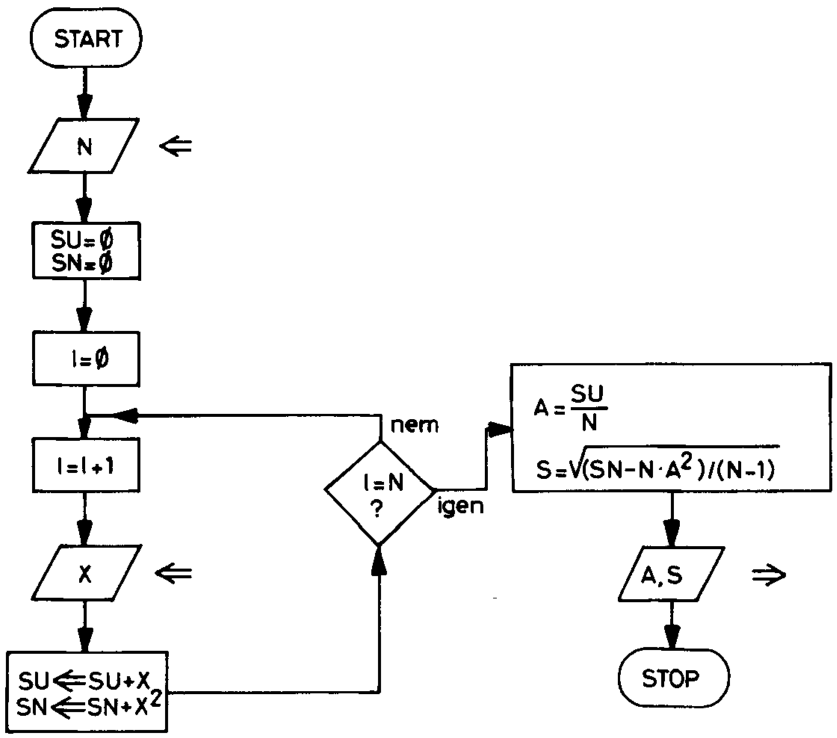
x_i (röviden: X) érték beolvasása;
X hozzáadása a SU tár tartalmához;
X^2 hozzáadása az SN tár tartalmához.

Így alakult ki az eljárásnak a 12*a* ábrán látható blokkdiagramja. A BASIC program:

```
1010 INPUT "N="; N:SU=0: SN=0: 1=0
1020 I=I+1: INPUT "X="; X: SU=SU+X: SN=SN+X↑2
1030 IF I < N THEN 1020
1040 A=SU/N: S=SQR ((SN-N*A↑2)/(N-1))
1050 PRINT "ATLAG="; A; "SZORAS="; S:END
```

b) Ha az

$$\bar{x}=\frac{\Sigma x_i}{n},\quad s^{*2}=\frac{\Sigma(x_i-\bar{x})^2}{n-1}$$



12*a* ábra a)

képletekkel dolgozunk, akkor a korrigált szórásnégyzet kiszámításához

– az x_i értékeket külön-külön tárolnunk kell (indexes $X(I)$ változóval n db tárat kell lefoglalni);

– előbb az $\bar{x}$ értéket kell kiszámíttatni (nem futhat le egy ciklusban $\bar{x}$ és s^{*2} kiszámítása).

A most közlendő program *hátránya* az előbbivel szemben, hogy

– nagyobb a tárigénye;

– hosszabb a futási idő.

Előnye viszont: az x_i értékek utólagos ellenőrzésre, javításra, felhasználásra lehívhatók a tárból (az előbbinél ez lehetetlen).

Az x_i értékek összegezését most egy FOR utasításos ciklusban, egy Z változóban összegezzük (a Z nullázása után a ciklusban Z = Z + X(I) utasítással). Ha a Z tartalmára (Σx_i-re) a későbbiek során nem lesz szükség, akkor egy második FOR utasításos ciklusban az $(x_i - \bar{x})^2$ értékek összegét számíttathatjuk ki ugyanazon Z változóban.

A folyamatábra a 12*b* ábrán látható. A program:

```
1005 REM ATLAG ES SZORAS
1010 INPUT "N="; N
1020 DIM X(N)
1030 Z=0
1040 FOR I=1 TO N
1050 PRINT "X(";I;")=": INPUT X(I)
1060 Z=Z+X(I)
1070 NEXT I
1080 XA=Z/N: Z=0
1090 FOR I=1 TO N
1100 Z=Z+(X(I)-XA)↑2
1110 NEXT I
1120 SC=SQR (Z/(N-1))
1130 PRINT "ATLAG="; XA
1140 PRINT "KORRIGALT SZORAS="; SC
1150 END
```

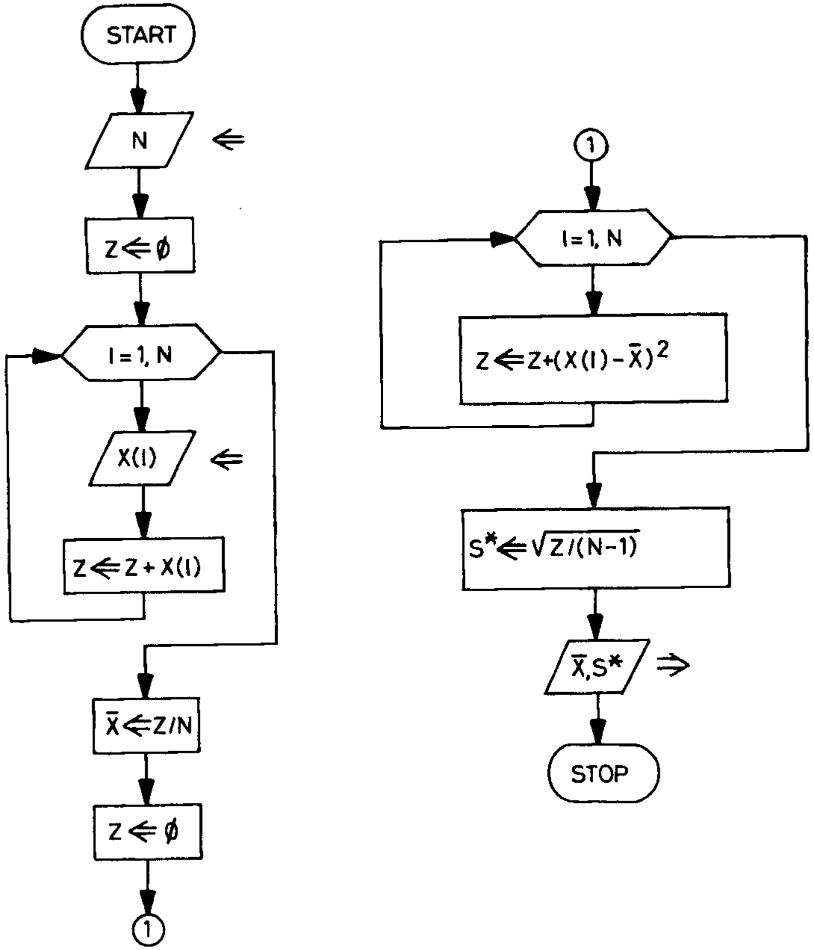
Megjegyzés

A program elején az 1020 DIM X(N) az X tömb (vektor) dimenzióját, méretét határozza meg. Ezzel az utasítással az adatoknak előre helyet foglalunk a gép tárában. Csak akkor kötelező kitenni, ha a dimenzió: DIM > 10.

Példa. Az adatok: 1, 2, 2, 3, 3, 3, 3, 4, 4, 5; itt $n = 10$, $\bar{x} = 3$, $s^* = 1{,}095\,445\,1$.

A következőkben különböző *adatrendező, mintát elemző statisztikai programokat* mutatunk be.

19. Adott n-elemű mintából kerestessük ki számítógéppel a legnagyobbat és a legkisebbet! Számítassuk ki a terjedelmet!



12*b* ábra

A beolvasott $a_1, a_2, \ldots, a_n$ elemek közül először az első kettőt hasonlítsuk össze! Ha $a_2 \geqq a_1$, akkor a_2 a legnagyobb, a_1 pedig a legkisebb (az eddig vizsgáltak közt). Tegyük el a_2-t a legnagyobbat őrző LN tárba, a_1-et pedig a legkisebbet tároló LK tárba! – Ha $a_2 \geqq a_1$ nem teljesül, akkor LN=A(1), LK=A(2).

Ezekután I=3-tól N-ig vizsgáltassuk meg, hogy vajon

$$a_i < \mathrm{LK}$$

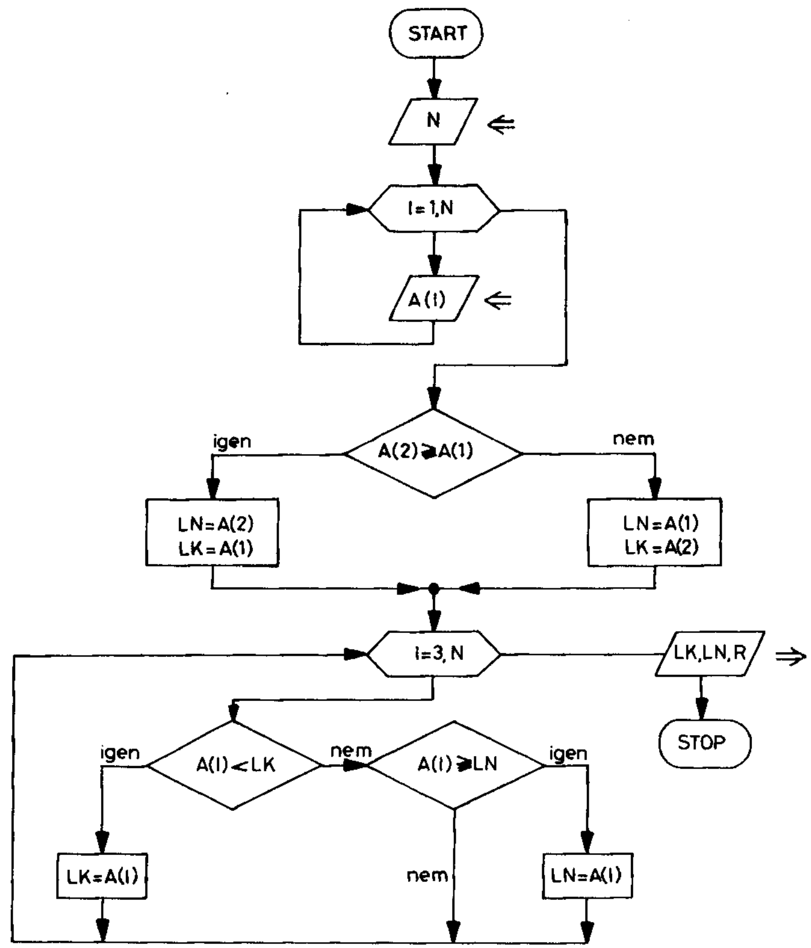
teljesül-e! Ha igen, akkor a_i a legkisebb az eddigiek közül (LK=A(I)), és ezután folytassuk a ciklust újabb elem bevonásával! Ha nem, akkor nézzük meg, hogy

$$a_i \geqq \mathrm{LN}$$

fennáll-e! Ha igen, akkor az eddigi elemek közül a_i a legnagyobb (LN=A(I)), és folytassuk a ciklust. Ha nem, folytassuk a ciklust!

A ciklus teljes lefutása után az LK tár a legkisebb mintaelemet, LN a legnagyobb mintaelemet tartalmazza. A terjedelem (range) R=LN−LK. A három tár tartalmának kiíratásával a program lefutott.

A folyamatábra a 13. ábrán, a program a következőkben látható:



13. ábra

```
  5 REM LEGNAGYOBB-LEGKISEBB KIVALASZTASA
 1Ø INPUT "N="; N
 12 DIM A(N)
 15 FOR I=1 TO N
 2Ø PRINT "A"; I: INPUT A(I)
 25 NEXT I
 3Ø IF A(2)>=A(1) THEN 5Ø
 4Ø LN=A(1): LK=A(2): GOTO 6Ø
 5Ø LN=A(2): LK=A(1)
 6Ø FOR I=3 TO N
 7Ø IF A(I)<LK THEN 11Ø
 8Ø IF A(I)>=LN THEN 1ØØ
 9Ø GOTO 12Ø
1ØØ LN=A(I): GOTO 12Ø
11Ø LK=A(I)
12Ø NEXT I
13Ø ? "LEGKISEBB ELEM="; LK
14Ø ? "LEGNAGYOBB ELEM=" LN
15Ø ? "TERJEDELEM=R="; LN-LK
16Ø END
```

Példa. N=6, A(I)=4, 5, 6, 1, 2, 3.

LEGKISEBB ELEM=1,
LEGNAGYOBB ELEM=6,
TERJEDELEM=R=5.

20. Válasszuk ki *n* megadott számból a különbözőket, adjuk meg a gyakoriságukat!

Ez a program felhasználható más programokban, ahol szükség van az adatok gyakoriságára.

---

a) Olvassuk be az adatokat: 5, 4, 3, 5, 3, 1, 5, 1, 1.

b) Legyen I=1, a legelső X(1)=5 elemet tegyük el egy E(1) tárba! Vizsgáljuk meg sorra, hogy *az E(1) tárban eltett érték megegyezik-e az utána következő X(2)-vel, X(3)-mal, ..., X(9)-cel!* Ha nem, menjünk tovább a megegyezés vizsgálatával; ha igen, a gyakoriságát számláló D(1) számláló értékét növeljük

meg 1-gyel és utána menjünk tovább! Így J = I+1 = 2-től N=9-ig azt vizsgálva, hogy

E(K)=X(J)

fennáll-e, a D(K), vagyis most a D(1) darabszámláló 3-ra fog felfutni, hiszen az E(1)=5 első elem D(1)=3-szor fordul elő. Itt K jelenti a jelenleg megtalált különböző elemek sorszámát, E(K) a K-adik különböző elemet, D(K) pedig a K-adik különböző elem darabszámát (gyakoriságát). Az első elem többivel való összehasonlításakor K=1.

c) I nem érte még el az N-et. Növeljük I értékét 1-gyel (I = I+1)! Az X(2)=4 elemet *hasonlítsuk össze az előzővel,* tehát vizsgáljuk meg, hogy X(I) (azaz most X(2)) megegyezik-e E(J)-vel (azaz X(1)-gyel)! A vizsgálatot nyilván ciklusszervezéssel, J=1-től K-ig (azaz most 1-ig) hajtjuk végre. X(I)=E(J), azaz X(2)=X(1) most nem teljesül. Jelezze ezt az, hogy *ilyenkor egy V segédváltozó Ø értéket vesz fel. Ez esetben folytassuk a programot úgy, hogy a különböző elemek sorszámát változtassuk kettőre* K = K+1 utasítással, hiszen megtaláltuk a második különböző elemet, E(K) = E(2) = X(I) = X(2) = 4-et! Egyúttal a D(2)=D(K) darabszámláló értékét 1-re kell állítanunk.

Az előzőleg a *b)* pontban adott vizsgálatot, azaz *X*(2) *összehasonlítását az utána következő* X(3), X(4), ..., X(9) tagokkal végezzük el ismét! Pontosabban J=1-től N-ig szervezett ciklusban vizsgáljuk meg, teljesül-e, hogy E(K)=X(J)! Nem fog teljesülni, tehát a ciklus végén

E(2)=4, D(2)=1

lesz.

Mivel I még mindig nem érte el az N-et, I-t 1-gyel növeljük, vizsgáljuk meg, hogy X(3) *egyenlő-e az előtte levőkkel,* X(1)-gyel és X(2)-vel. Ez a *c)* pontban adott vizsgálat, J=1-től K-ig vizsgáljuk, hogy X(I)=E(J) teljesül-e! Nem, tehát a V segédváltozó ismét Ø marad, K-t 1-gyel növelve, előbb a *b)* alatti lépéssorozatot végezzük el (eredmény: a *K*=3. különböző elem E(3)=3, D(3)=2), utána a *c)* alatti lépéssorozatot. Itt azt találjuk, hogy X(4)=5 már előfordult, X(I)=E(J) teljesül. *Ezt a V segédváltozó 1-re való állításával rögzítjük. Most a különbözők számát, K-t nem kell növelni,* hanem az I értékét 1-gyel növelve, az X(5)=3-ról vizsgáljuk, új elem-e vagy már előfordult (*c)* alatti lépéssorozat).

d) Így megy a *b)* és *c)* lépések váltogatása addig, míg I=N nem lett. Az X(9) előzőekkel való összehasonlítása után egy ciklusban J=1-től K-ig *kiíratjuk a különböző elemeket és gyakoriságukat,* vagyis az E(J) és D(J) változókat.

A futtatás eredménye nyilván:

E(1) = 5, D(1) = 3,
E(2) = 4, D(2) = 1,
E(3) = 3, D(3) = 2,
E(4) = 1, D(4) = 3.

Akinek ez nehéz és hosszú volt, annak számára még egyszer összefoglaljuk, hogy milyen lépéssorozatot végzünk egy-egy ciklusban.

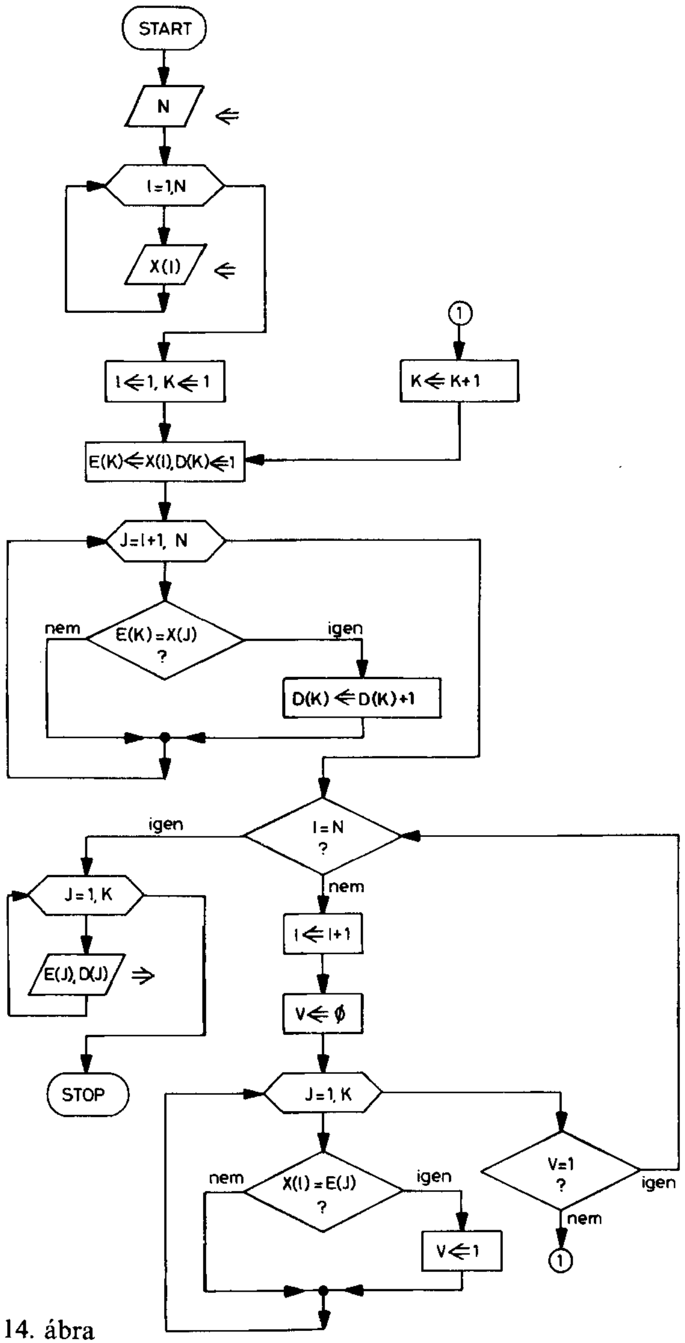
a) X(I) értékek *beolvasása* I = 1-től N-ig. Legyen I a megvizsgálandó tag indexe, K a különbözők száma az eddig megvizsgáltak között, E(K) a K-adik különböző elem, D(K) ennek az elemnek a darabszáma (gyakorisága). Állítsuk I-t, K-t 1-re!

b) X(I) értéket tegyük az E(K) tárba, D(K) legyen 1! *Hasonlítsuk össze X(I)-t az utána levőkkel!* Egy J = I + 1-től N-ig terjedő ciklusban megvizsgáljuk, hogy E(K) = X(J) teljesül-e. Ha igen, D(K)-t 1-gyel növeljük, ha nem teljesül, akkor nem növeljük. Folytatjuk a ciklust N-ig.

c) Megnézzük, hogy I elérte-e N-et. Ha még nem, I-t 1-gyel növeljük. Megvizsgáljuk, hogy *az új X(I) egyezik-e az előző, különbözőnek talált elemekkel,* X(I) = E(J) fennáll-e. Ha nem, akkor V = 0 jelzi, hogy X(I) új elem, vagyis a különbözők számát 1-gyel növelve (K = K + 1), a *b)* alatti lépéssorozatot ismételjük meg! Ha igen, akkor V = 1 jelzi, hogy K-t nem kell növelnünk (nincs új különböző tag), helyette a *c)* alatti eljárást kell megismételnünk.

d) Ha I már elérte N-et, akkor *kiíratjuk a különbözőnek talált elemeket és gyakoriságukat.*

Lássuk ezek után a 14. ábrán található blokkdiagramot és a programot:



14. ábra

```
700 REM ADATOK GYAKORISAGA
710 INPUT "ADATOK SZAMA"; N: DIM X(N), E(N), D(N)
720 FOR I=1 TO N: INPUT X(I): NEXT I
730 I=1: K=1
740 E(K)=X(I): D(K)=1
750 FOR J=I+1 TO N
760 IF E(K)=X(J) THEN D(K)=D(K)+1
770 NEXT J
780 IF I=N THEN GOTO 850
790 I=I+1: V=0
800 FOR J=1 TO K
810 IF X(I)=E(J) THEN V=1
820 NEXT J
830 IF V=1 THEN GOTO 780
840 K=K+1: GOTO 740
850 FOR J=1 TO K
860 ? "E(";J;")="; E(J); "D(";J;")="; D(J)
870 NEXT J: END
```

Megjegyzés

Ha a folyamatábrán V = 1 nem teljesül, a program az 1-gyel jelölt ponton folytatódik tovább (visszakanyarodik az elejére).

21. Az $x_1, x_2, x_3, \ldots, x_n$ mintaelemeket rendeztessük növekvő sorrendbe úgy, hogy mindig a legkisebbet rendezzük előre! Írassuk ki a kapott $x_1^* \leqq x_2^* \leqq \leqq x_3^* \leqq \ldots \leqq x_n^*$ rendezett mintát! *(Növekvő sorrendbe rendezés a legkisebb kiválasztásával.)*

A gépnek két lépéssorozatot kell elvégeznie:

a) az *i*-edik legkisebbet megkeresnie;

b) a megtalált legkisebbet két változó értékének a felcserélésével a megfelelő helyre hoznia.

A következő egyszerű *példa* mutatja, mit csinálunk.

Legyen a rendezetlen adatsor:

x_1	x_2	x_3	x_4	x_5
5	4	3	2	4

a) Az I = 1-nek megfelelő ciklusban vizsgáljuk az első elemet, x_1-et. Nyilván az eddig vizsgáltak közül ez a legkisebb, indexét jegyeztessük meg egy IN tárban (IN = I, most IN = 1).

Hasonlítsuk össze az eddigi legkisebb elemet a következőkkel, $x_2, x_3, \ldots, x_n$-nel, azaz J = I + 1-től N-ig vizsgáljuk, hogy

X(J) < X(IN)

fennáll-e! Ha nem áll fenn, az összehasonlító ciklust tovább kell folytatni. Ha fennáll, akkor az X(J) elem indexe lesz az eddig legkisebb elem indexe,

IN = J,

és csak ezután folytatódik az összehasonlító ciklus, most már az új IN indexű taggal.

Példánkban először x_2-t találja a gép kisebbnek (IN = 2), aztán x_2 és x_3 összehasonlításával x_3-at (IN = 3), x_3 és x_4 összehasonlításával x_4-et (IN = 4), $x_4 = 2$ és $x_5 = 4$ közül azonban x_4 a kisebb (IN marad 4).

A J = I + 1-től N-ig mozgó ciklusban *megtaláltuk az első (I-edik) legkisebbet, index IN = 4.*

b) Ezekután x_1-et és x_4-et (azaz X(I)-t és *X*(IN) tárak tartalmát) fel kell cserélni. Az *elemcsere* úgy történhet, amint ahogy egy A és B hordóban levő bor tartalmát felcseréljük egy S „segédhordó" segítségével:

S-be áttöltjük A tartalmát;
A-ba áttöltjük B tartalmát;
B-be áttöltjük S tartalmát.

Legyen a tárban a „segédhordó" Y, az előbbi lépéseknek a gépben a következő lépések felelnek meg:

Y = X(IN),
X(IN) = X(I),
X(I) = Y.

Ezzel a cserével *az első legkisebb elem az első helyre került.* Az elemek most:

x_1	x_2	x_3	x_4	x_5
2	4	3	5	4

a) Ezekután az I = 2-nek megfelelő lépéssorozatban a legkisebb indexet először 2-re állítva (IN = I), x_2-t összehasonlítjuk x_3, x_4, x_5-tel, vagyis J = = I + 1-től N-ig megnézzük, fennáll-e

X(J) < X(IN).

Ha valamelyik J-re fennáll, akkor az az X(J) elem lesz az eddig vizsgáltak közül a második legkisebb, a további összehasonlítást ezzel az elemmel kell végezni. Ha nem áll fenn, tovább folytatjuk a J-s ciklust (az összehasonlítást). Példánkban $x_3=3$ lesz *a második legkisebb.*

b) Az így megtalált második legkisebb cseréljen helyet x_2-vel, azaz *X(IN) cseréljen helyet X(I)-vel.* Ez a már felírt

Y = X(IN),
X(IN) = X(I),
X(I) = Y

lépések végrehajtásával történik.

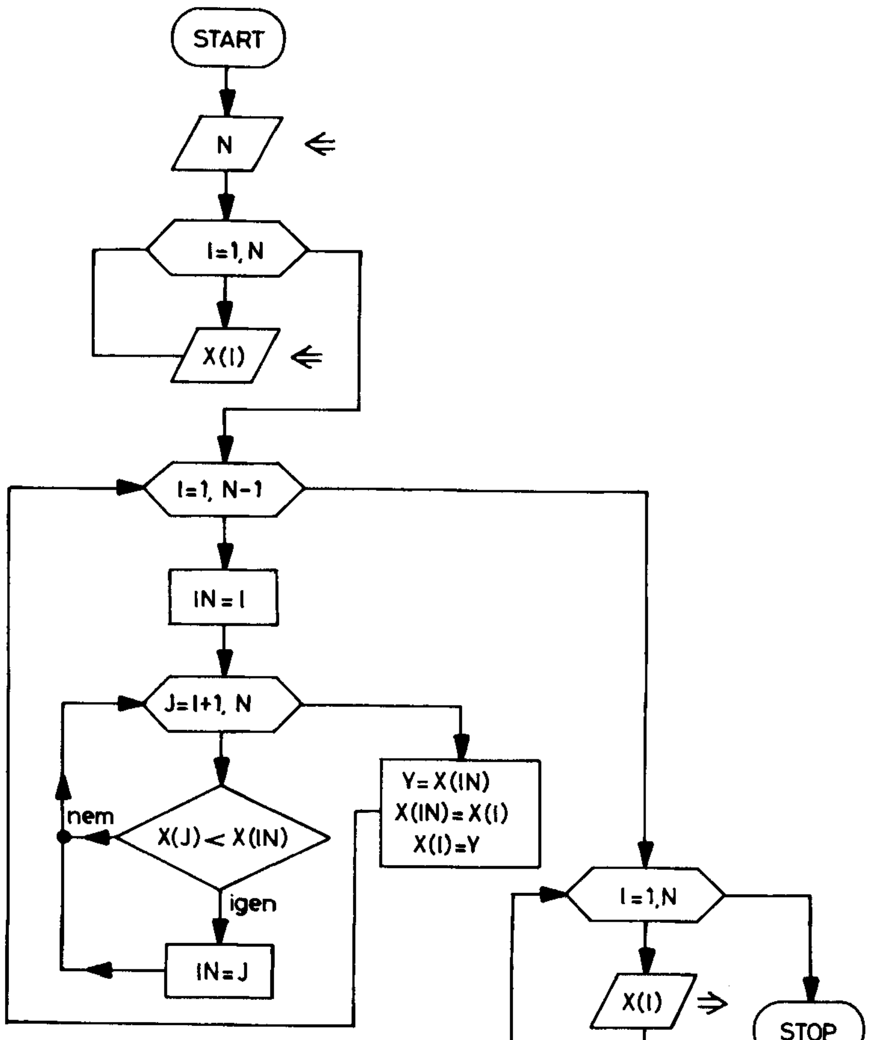
A mintaelemek sorrendje az I = 2-höz tartozó lépések után:

$$\frac{x_1\ x_2\ x_3\ x_4\ x_5}{2\ \ 3\ \ 4\ \ 5\ \ 4}.$$

I-t egyesével növelve, az *a)–b)* eljárást ismételgetve I = N − 1 = 4 után *kialakul a rendezett minta.* (Az I ciklus tehát csak N − 1-ig megy!)

Még egyszer összefoglalva az egészet:

Ciklusváltozó	x_1	x_2	x_3	x_4	x_5
	5	4	3	2	4
$I=1$	2	4	3	5	4
$I=2$	2	3	4	5	4
$I=3$	2	3	4	5	4
$I=4$	2	3	4.	4	5



15. ábra

A folyamatot a 15. ábra rögzíti. A *program:*

```
  5 REM RENDEZES LEGKISEBB KIVALASZTASAVAL
 10 INPUT "N="; N
 20 FOR I=1 TO N
 30 INPUT "X";I;"="; X(I)
 40 NEXT I
 50 FOR I=1 TO N-1
 60 IN=I
 70 FOR J=I+1 TO N
 80 IF X(J) < X(IN) THEN 100
 90 GOTO 110
100 IN=J
110 NEXT J
120 Y=X(IN): X(IN)=X(I): X(I)=Y
130 NEXT I
140 FOR I=1 TO N
150 PRINT X(I);",";
160 NEXT I
170 END
```

Megjegyzés

a) Az I=2-nek megfelelő sorrend után $x_3=4$ és $x_4=5$, valamint $x_3=4$ és $x_5=4$ összehasonlításával azt találja a gép, hogy x_3 a legkisebb (IN=3), ezt kell felcserélni x_3-mal (I=3). A csere tehát formális, de nem hibás.

b) A rendezett minta elkészítésére más eljárások is vannak (pl. egymás melletti elemek cseréjével, ha az elsőnek vizsgált elem nagyobb az utána következőnél stb.).

c) Ha a sorbarendezési program után a sűrűségfüggvényt, eloszlásfüggvényt is meg akarjuk állapíttatni (l. a következő feladatot), *akkor a sorbarendezési programot hagyjuk a gépbe betöltve,* az utasítások sorszámai a két feladatban össze vannak fésülve.

Ez esetben megadhatjuk a sorbarendező programot szubrutinként is, de akkor célszerű a sorbarendezésnél az utasítások sorszámait pl. 500-zal növelni. (Az 5-től 40-es utasításig nincs szükség az x_i értékek beolvasására, az az ottani programban is szerepel.)

22. A gépbe beolvasott n-elemű rendezetlen minta segítségével határozzuk meg a tapasztalati sűrűség- és eloszlásfüggvényt!

I. megoldás

Ennél a megoldásnál az eloszlásfüggvény az eredeti adatok segítségével van megadva. A mintaelemek nagyság szerinti sorrendbe történő rendezése után (l. az előző feladatot) a legnagyobb és legkisebb elemet kiíratva, megadjuk az A, B intervallumot, amelyen kívül a sűrűségfüggvény 0, egyúttal közöljük az osztályok számát is ($\mathcal{O}$-t). Így a gép a

$$H = \frac{B-A}{\mathcal{O}}$$

terjedelmű osztályokba sorolva a *sűrűségfüggvényt* a J-edik intervallumon a szokásos

$$f(x) \approx \frac{\text{az intervallumra eső adatok száma}}{\text{(összes adat száma) (intervallum hossza)}} = \frac{D(J)}{NH}$$

képlettel számolja. Előzőleg persze mindegyik gyakoriságszámlálót 0-ra állítjuk (D(J)=0 J=1-től $\mathcal{O}$-ig).

Az egyes intervallumokra eső adatok számának (vagyis a D(J) értékeknek a megállapításához) I=1-től N-ig azt kell vizsgálni, hogy az elem az intervallumra esik-e, pontosabban:

$$A+(J-1)H \leqq X(I) < A+JH$$

fennáll-e (J=1-től $\mathcal{O}$-ig). Ha nem esik az adott intervallumra, az azt jelenti, hogy ugyanaz az X(I) várhatóan a következő intervallumba esik (J-t kell 1-gyel növelni és vizsgálni az előző egyenlőtlenségek teljesülését). Ha viszont az adott intervallumra esik, akkor viszont D(J)-t kell 1-gyel növelni, és I-hez is 1-et hozzáadva, az új X(I)-nek az intervallumra esését vizsgálni.

Az I változójú külső ciklus lezajlása (I=1-től N-ig), a J változójú belső ciklus végrehajtása (J=1-től $\mathcal{O}$-ig) után a sűrűségfüggvény értékeit kiírathatjuk az adott $\mathcal{O}$ darab részintervallumban.

Az eloszlásfüggvény:

$$F(x)=0, \quad \text{ha} \quad x \leqq x_1,$$
$$F(x)=1, \quad \text{ha} \quad x > x_n.$$

Ezután ciklusszervezéssel (i=1-től n-ig) megállapítjuk $F(x_i)$ értékét. Ha

$$x_i \neq x_{i+1},$$

akkor

$$F(x_{i+1}) = \frac{i}{n}$$

(hiszen minden egyes felvett értéknél $F(x)$ $\frac{1}{n}$-et ugrik). Ezt az $F(x_i)$ értéket ki is íratjuk, és ezután folytatjuk a ciklust. Ha

$$x_i = x_{i+1},$$

akkor is ugrik x_i-nél az eloszlásfüggvény $\frac{1}{n}$-et, de $F(x_i)$-t nem kell kiíratni, mert x_{i+1}-nél minimum $\frac{2}{n}$-et ugrik az eloszlásfüggvény.

A ciklust $i=1$-től n-ig folytatva az eloszlásfüggvény értékeit a felvett x_i értékeknél kiíratjuk. Mivel az eloszlásfüggvény balról folytonos, az előző értékek alapján már megszerkeszthető a görbéje.

Megjegyzés

a) Pontosabb lett volna, ha $f_n(x)$-et és $F_n(x)$-et írunk (még pontosabb, ha feltüntetjük, hogy mindkét függvény értékei véletlen változók), de agyonbetűzni a beírást nem érdemes. A programba úgyis csak latin nagybetűket írhatunk.

b) Az

$$A+(J-1)H \leqq X(I) < A+JH$$

egyenlőtlenségek a programban az

$$A+(J-1)H \leqq X(I) \text{ AND } X(I) < A+JH$$

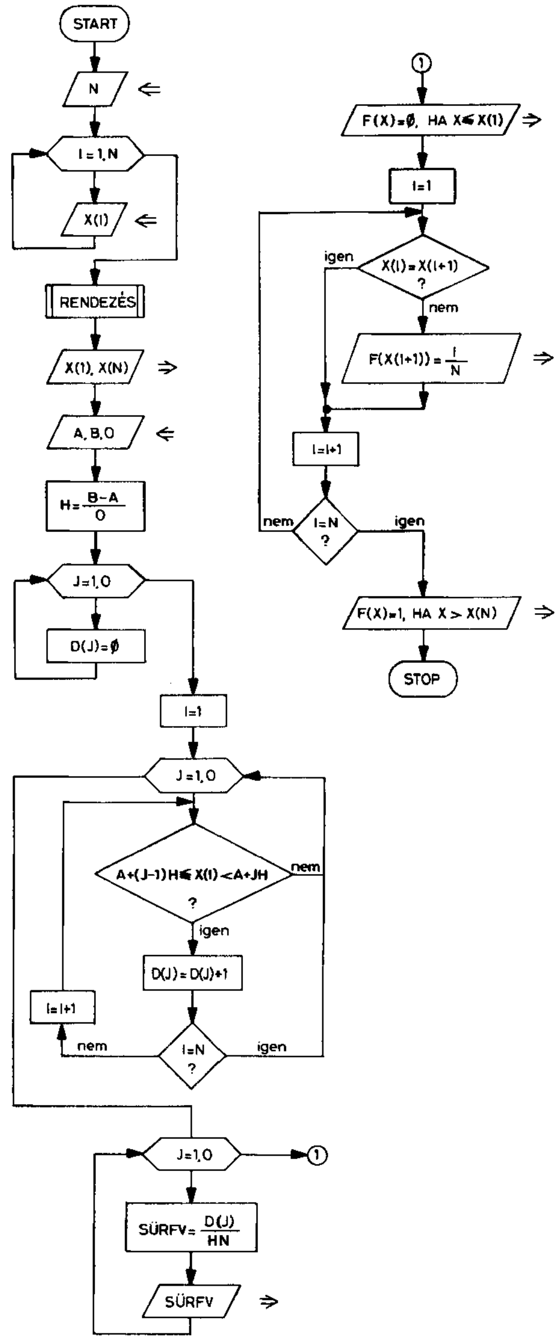
reláció teljesülésének a vizsgálatát jelenti (l. a 260. utasítást).

c) Ennek a feladatnak a lefutásához az előző sorbarendező program is szükséges. Ha a sorbarendező programot szabályos szubrutinként adjuk meg, akkor a program

```
 45 GOSUB 55Ø
570 RETURN
```

utasításokkal bővül.

A *folyamatábrát* l. a 16*a* ábrán.



16*a* ábra a)

A program:

```
5 REM SURUSEG- ES ELOSZLASFUGGVENY
10 INPUT "ADATOK SZAMA="; N
15 DIM X(N), D(N)
20 FOR I=1 TO N
30 ? "X"; I; "=": INPUT X(I)
40 NEXT I
50–160 l. előző feladatot
170
180 PRINT "LEGKISEBB="; X(I); "LEGNAGYOBB="; X(N)
190 INPUT "A"; A: INPUT "B="; B: INPUT "Ø"; Ø
200 H=(B-A)/Ø
210 FOR J=1 TO Ø
220 D(J)=0
230 NEXT J
240 I=1
250 FOR J=1 TO Ø
260 IF A+(J-1)*H<=X(I) AND X(I)<A+J*H THEN 280
270 GOTO 305
280 D(J)=D(J)+1
290 IF I=N GOTO 305
300 I=I+1: GOTO 260
305 NEXT J
310 FOR J=1 TO Ø
320 SURFV=D(J)/(H*N)
330 ? A+(J-1)*H; "<=X<"; A+J*H; "SURFV="; SURFV
340 NEXT J
350 ? "ELOFV=0, HA X<="; X(1)
370 I=1
380 IF X(I)=X(I+1) THEN 400
390 PRINT "F("X(I+1)")="; I/N
400 I=I+1
410 IF I=N THEN 430
420 GOTO 380
430 ? "ELOFV=1, HA X>"; X(N)
440 END
```


Egyszerű, kipróbáló példák:

a) N=5, X(I): 2, 4, 4.9, 3, 1,
A=1, B=5, Ø=4,

eredmény:

1, 2, 3, 4, 4.9, LEGKISEBB=1, LEGNAGYOBB=4.9
1≦X<2 SURFV= .2
2≦X<3 SURFV= .2
3≦X<4 SURFV= .2
4≦X<5 SURFV= .4
ELOFV=0, HA X≦1
F(2)= .2
F(3)= .4
F(4)= .6
F(4.9)= .8
ELOFV=1, HA X>4.9

b) N=5, X(I): 4.5, 2, 1, 3, 2,
A=1, B=5, Ø=4

Az eredmény hasonló az előzőhöz. Ezzel az adatsorral azt teszteltük, hogy hogyan működik az eloszlásfüggvényt számító program, ha egyenlő elemek is vannak:

F(2)= .2
F(3)= .6 (tehát $\frac{2}{n}$ = 0,4-del ugrik)
F(4.5)= .8

II. megoldás

Ennél a megoldásnál minden adatot az intervallum kezdőpontjának abszcisszájával helyettesítünk. Ezáltal

- nincs szükség az adatok sorbarendezésére;
- a sűrűségfüggvény változatlan;
- *az eloszlásfüggvény áttekinthetőbb, mint az előbb.*

(Ez különösen nagytömegű adat esetén előnyös.)

Az algoritmus azonos az előzővel, az eloszlásfüggvény értékeinek a kiszámítását kivéve. Az eloszlásfüggvény:

– ugrásai nem a mintaelemeknél, hanem az osztályok végpontjában vannak;

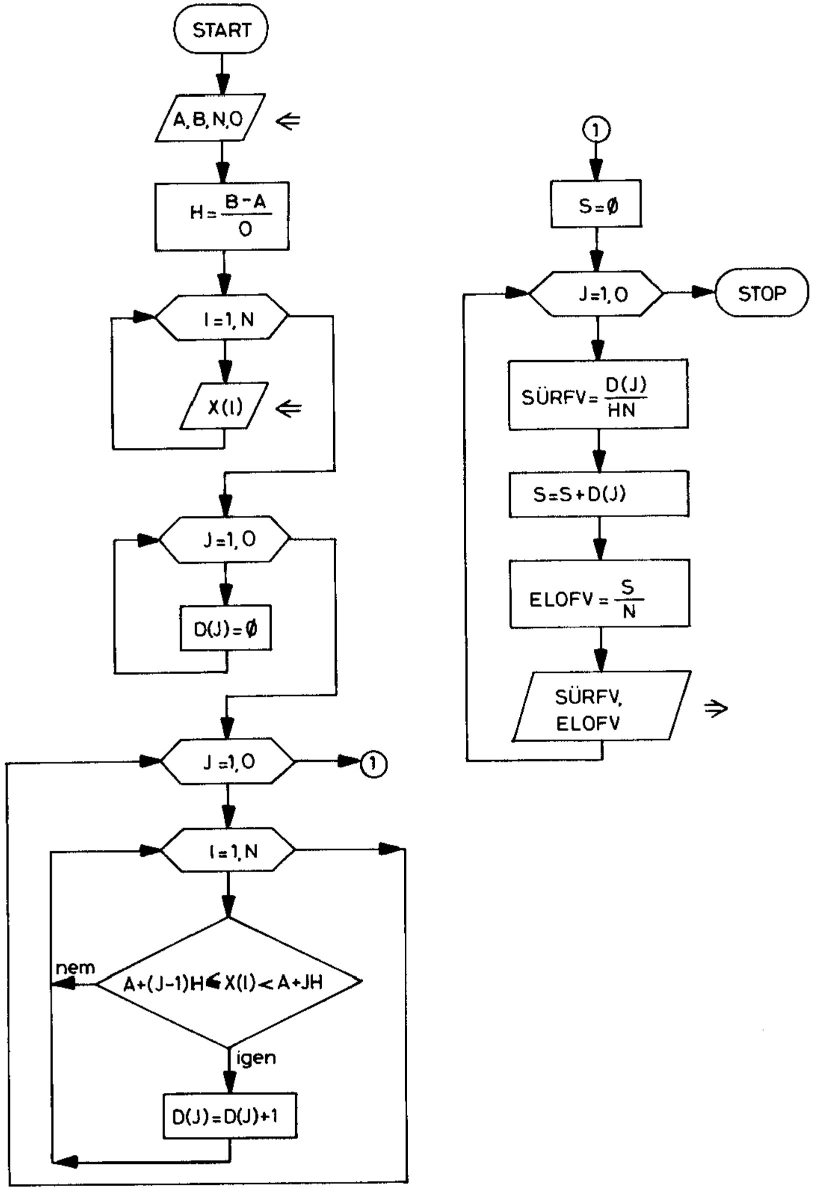
– az ugrások nagysága nem $\frac{1}{n}$ vagy $\frac{2}{n}$ stb., hanem $\frac{D(J)}{N}$ (D(J) a J-edik intervallumra eső adatok száma);

– egyszerűség kedvéért a sűrűség- és eloszlásfüggvényt azonosan, egy-egy intervallumra íratjuk ki, nem törődve, hogy melyik, hogyan folytonos.

Az algoritmust rögzítettük a 16*b folyamatábrán.* A hozzávaló *program:*

```
5 REM SURFV–ELOFV
1Ø INPUT A, B, N, O
2Ø H=(B–A)/O
25 DIM X(N), D(O)
3Ø FOR I=1 TO N
4Ø ? "X"I"=": INPUT X(I)
5Ø NEXT I
6Ø FOR J=1 TO O
7Ø D(J)=Ø
8Ø NEXT J
9Ø FOR J=1 TO O
1ØØ FOR I=1 TO N
11Ø IF A+(J–1)*H<=X(I) AND X(I)<A+J*H THEN D(J)=D(J)+1
12Ø NEXT I
13Ø NEXT J
14Ø S=Ø
15Ø FOR J=1 TO O
16Ø SURFV=D(J)/(H*N): S=S+D(J):ELOFV=S/N
17Ø ? A+(J–1)*H; "<X<"; A+J*H
18Ø ? "SURFV="; SURFV
19Ø ? "ELOFV="; ELOFV
195 ?
2ØØ NEXT J
21Ø END
```

Példa. N = 10, X(I) = $\underbrace{-1{,}2}_{D(1)=1}$, $\underbrace{-0{,}8;\ -0{,}7;\ -0{,}6;\ -0{,}3}_{D(2)=4}$, $\underbrace{0;\ 0{,}5;\ 0{,}7;\ 0{,}9}_{D(3)=4}$, $\underbrace{1{,}8}_{D(4)=1}$.



16*b* ábra

b)

$A=-2$ és $B=+2$ között $\varTheta=4$ osztály, eredmény:

$-2<X<-1$
SURFV=0.1
ELOFV=0.1

$-1<X<0$
SURFV=0.4
ELOFV=0.5

$0<X<1$
SURFV=0.4
ELOFV=0.9

$1<X<2$
SURFV=0.1
ELOFV=1.0

23. Számítsuk ki az n-elemű minta korrigálatlan szórásnégyzetének, S_n^2-nek a várható értékét! Legyen a véletlen változó várható értéke $M(X)=m$. Természetesen $M(X_i)=m$ is fennáll ($i=1, 2, \dots, n$).

Alakítsuk át S_n^2-et úgy, hogy az X_i-m értékek négyzete szerepeljen benne $(X_i-\bar{X})^2$ helyett; előbbinek ugyanis tudjuk a várható értékét:

$$M[(X_i-m)^2] = D^2(X_i) = D^2(X) = \sigma^2,$$

az $(X_i-\bar{X})^2$ értékek várható értékét viszont közvetlenül nem ismerjük.

A cél elérésére alakítsuk át S_n^2-et:

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n}\Sigma\,[(X_i-m)-(\bar{X}-m)]^2 = \\ &= \frac{1}{n}\Sigma\,(X_i-m)^2-\frac{2}{n}(\bar{X}-m)\,\Sigma\,(X_i-m)+\frac{1}{n}\cdot n(\bar{X}-m)^2 = \\ &= \frac{1}{n}\Sigma\,(X_i-m)^2-2\,(\bar{X}-m)\,(\bar{X}-m)+(\bar{X}-m)^2 = \\ &= \frac{1}{n}\Sigma\,(X_i-m)^2-(\bar{X}-m)^2. \end{aligned}$$

Vegyük a várható értékét:

$$\begin{aligned} M(S_n^2) &= \frac{1}{n}\Sigma\, M[(X_i-m)^2]-M[(\bar{X}-m)^2] = \frac{1}{n}\cdot n\sigma^2-D^2(\bar{X}) = \\ &= \sigma^2-\frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2. \end{aligned}$$

S_n^2 várható értéke tehát kisebb a sokaság szórásnégyzeténél $\left(\frac{n-1}{n}<1\right)$. Ha tehát sokszor veszünk n-elemű mintát, az S_n^2 értékek átlagai nem σ^2 körül ingadoznak.

Tekintsük a korrigált szórásnégyzetet:

$$S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2,$$

aminek várható értéke már σ^2 lesz:

$$M(S_n^{*2}) = \frac{n}{n-1}M(S_n^2) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2.$$

A „Statisztikai becslések" c. fejezetben ezt így fejezzük ki röviden: *a korrigált szórásnégyzet torzítatlan becslés az egész sokaság szórásnégyzetére* (a korrigált szórásnégyzet pedig torzított becslés).

Megjegyezzük még, hogy ha a mintaelemek száma elég nagy, akkor S_n^2 is elég jó becslés σ^2-re. Ugyanis, ha $n\to\infty$,

$$\lim_{n\to\infty} M(S_n^2) = \lim_{n\to\infty}\frac{n-1}{n}\sigma^2 = \lim_{n\to\infty}\left(1-\frac{1}{n}\right)\sigma^2 = \sigma^2.$$

Az ilyen tulajdonságú becslést *aszimptotikusan torzítatlannak* nevezzük (l. később).

24. Az $F_n(x)$ tapasztalati eloszlásfüggvény az $X_1, X_2, \dots, X_n$ mintaelemek (véletlen változók) függvénye, tehát maga is valószínűségi változó. Számítsuk ki várható értékét, szórását, a Csebisev-egyenlőtlenség segítségével adjunk becslést arra, hogy rögzített n esetén $F_n(x)$ és $F(x)$ (vagyis a tapasztalati és elméleti eloszlásfüggvény) mennyire térhet el egymástól!

Adott x helyen $F_n(x)$ annak a valószínűsége, hogy a *mintában* a változó x határ alá essék, vagyis: $F_n(x)$ az $X<x$ esemény *relatív gyakorisága.* Ugyanezen esemény elméleti *valószínűsége* $F(x)$.

Az $X<x$ esemény *gyakorisága:*

$$k = nF_n(x).$$

Egy esemény gyakorisága binomiális eloszlású. Így a binomiális eloszlásnál tanultakat alkalmazva:

$$a)\ P\left(F_n(x) = \frac{k}{n}\right) = P(nF_n(x) = k) = \binom{n}{k} p^k(1-p)^{n-k} =$$

$$= \binom{n}{k} [F(x)]^k [1-F(x)]^{n-k},$$

$$b)\ M(nF_n(x)) = np,$$
$$nM(F_n(x)) = nF(x),$$
$$M(F_n(x)) = F(x)$$

(a tapasztalati eloszlásfüggvény várható értéke az elméleti eloszlásfüggvény).

$$c)\ D(nF_n(x)) = \sqrt{np(1-p)},$$
$$nD(F_n(x)) = \sqrt{nF(x)\,(1-F(x))},$$
$$D(F_n(x)) = \sqrt{\frac{F(x)\,(1-F(x))}{n}}.$$

Ezzel megkaptuk $F_n(x)$ *szórását* is.

d) A Csebisev-egyenlőtlenség $M(Z)$ várható értékkel, $D(Z)$ szórással rendelkező Z véletlen változóra kimondja, hogy

$$P(|Z-M(Z)| < kD(Z)) > 1-\frac{1}{k^2}.$$

Alkalmazzuk ezt a $Z = F_n(x)$ véletlen változóra!

$$P\left(|F_n(x)-F(x)| < k\sqrt{\frac{F(x)\,(1-F(x))}{n}}\right) > 1-\frac{1}{k^2}.$$

Ha k nagy, akkor ez a valószínűség nagy. $F(x)\,(1-F(x))$ értékét felülről kellene becsülnünk. A

$$t(1-t) = t-t^2$$

parabolának maximuma van ott, ahol a deriváltja:

$$1-2t = 0,$$

vagyis $t = \frac{1}{2}$ esetén. A maximum értéke:

$$\frac{1}{2}-\left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

Ezért $F(x)\,(1-F(x)) \leqq \frac{1}{4}$, így: $|F_n(x)-F(x)|$ $\frac{1}{\sqrt{n}}$ nagyságrendű eltérés tetszőleges x-re. Ebből az is következik, hogy *n növekedésével $\frac{1}{\sqrt{n}}$ nagyságrendben a tapasztalati és elméleti eloszlásfüggvény eltérése abszolút értékben zérushoz konvergál.* (Amikor $n \to \infty$, egyrészt $F_n(x)$ ugrásai kisebbek, másrészt $F_n(x)$ alig tér el $F(x)$-től.)

Megjegyzések

1. A *d)* pontban mondottak szépen illusztrálják Glivenko, Kolmogorov, Szmirnov tételeit.

2. $D(F_n(x))$ maximumát így is kihozhatjuk. Ennek a szórásnak szélsőértéke ott lehet, ahol

$$\sqrt{F(x)\,(1-F(x))}$$

maximális. A mértani közép nem lehet nagyobb, mint a számtani közép:

$$\sqrt{F(x)\,(1-F(x))} \leqq \frac{F(x)+(1-F(x))}{2} = \frac{1}{2},$$

a maximális értéket (a számtani közepet) akkor éri el a mértani közép, ha a két mennyiség egyenlő:

$$F(x) = 1-F(x),$$
$$F(x) = \frac{1}{2}.$$

Ekkor $F(x)\,(1-F(x)) \leqq \frac{1}{4}$ (mint előbb).

Az $F(x) = \frac{1}{2}$ egyenlet megoldása x-re (ha létezik) a medián. *A mediánnál lehet $F_n(x)$ szórása a legnagyobb.*

Érdekes, hogy ez utóbbi eredményt az $F(x)(1-F(x))$ függvény deriválásával is megkapjuk. Maximum esetén a derivált 0:

$$f(x)(1-F(x)) - F(x)f(x) = 0,$$
$$f(x) = 2f(x)F(x), \qquad /: f(x) \neq 0$$
$$\frac{1}{2} = F(x).$$

3. Ha a Csebisev-egyenlőtlenség

$$P(|Z - M(Z)| \geqq a) \leqq \frac{D^2(Z)}{a^2}$$

alakját használjuk fel, akkor a következőt kapjuk:

$$P(|F_n(x) - F(x)| \geqq a) \leqq \frac{F(x)(1-F(x))}{a^2 n} \to 0,$$

ha $n \to \infty$. Ez pedig azt jelenti, hogy $F_n(x)$ *sztochasztikusan konvergál* $F(x)$-hez minden x-re, tehát egyenletesen gyenge a konvergencia.

25. Igazoljuk a „nagyságra nézve k-adik" mintaelem, X_k^* eloszlásfüggvényére, $F_{n,k}(x)$-re a fejezet elméleti tudnivalóinak 7. pontjában megadott képleteket.

---

a) Vizsgáljuk először a legnagyobb mintaelem eloszlásfüggvényét, a

$$P(X_n^* < x) = F_{n,n}(x)$$

függvényt.

Ha az X_n^* legnagyobb mintaelem kisebb, mint x, akkor mindegyik megfigyelési érték kisebb x-nél. Így

$$P(X_n^* < x) = P(X_1 < x, X_2 < x, \ldots, X_n < x) =$$
$$= P(X_1 < x)P(X_2 < x)\ldots P(X_n < x) = [F(x)]^n.$$

Felhasználtuk, hogy az X_i mintaelemek függetlenek és azonos eloszlásúak.

Vizsgáljuk másodszor az X_1^* legkisebb mintaelem eloszlásfüggvényét, $F_{n,1}(x)$-et! $X_1^* < x$ esemény ellentettje $X_1^* \geqq x$. Ennek a valószínűsége a mintaelemek függetlensége és azonos eloszlása alapján:

$$P(X_1^* \geqq x) = P(X_1 \geqq x, X_2 \geqq x, \ldots, X_n \geqq x) =$$
$$= P(X_1 \geqq x)P(X_2 \geqq x)\ldots P(X_n \geqq x) = [1 - F(x)]^n.$$

Így

$$F_{n,1}(x) = P(X_1^* < x) = 1 - P(X_1^* \geqq x) = 1 - [1 - F(x)]^n.$$

Vizsgáljuk harmadszor az X_k^* rendezett mintaelem $F_{n,k}(x)$ eloszlásfüggvényét!

Az $X_k^* < x$ esemény akkor következik be, ha legalább k db megfigyelési érték kisebb, mint x, vagyis $X_k^* < x$, ha

- vagy k darab megfigyelési érték $< x$, többi $(n-k)$ darab $\geqq x$,
- vagy $k+1$ darab megfigyelési érték $< x$, többi $(n-k-1)$ darab $\geqq x$,
- ...
- vagy n darab megfigyelési érték $< x$.

Ezek egymást kizáró események, így ezek valószínűségeinek összege adja meg az $X_k^* < x$ esemény valószínűségét.

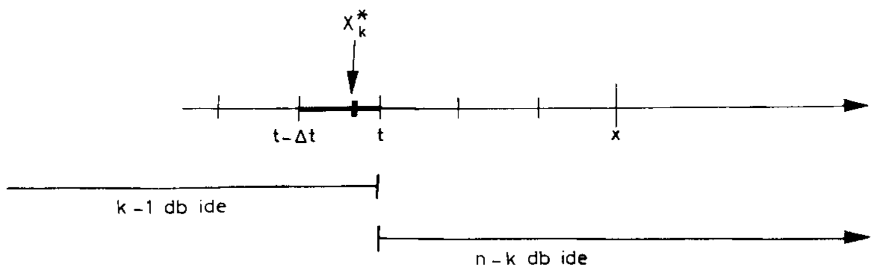
Annak a valószínűsége, hogy n darab független megfigyelés közül pontosan i számú megfigyelt érték kisebb x-nél, a többi $n-i$ számú megfigyelt érték pedig nagyobb, mint x, a binomiális eloszlással adható meg:

$$\binom{n}{i} F(x)^i [1 - F(x)]^{n-i}.$$

Ezáltal

$$P(X_k^* < x) = F_{n,k}(x) = \sum_{i=k}^{n} \binom{n}{i} F(x)^i [1 - F(x)]^{n-i}.$$

b) Osszuk fel a $(-\infty, x)$ intervallumot egymáshoz csatlakozó kis Δt hosszúságú intervallumokra (17. ábra)! Tudjuk a Δt értékét olyan kicsire választani, hogy:



17. ábra

I. X_k^* egy $(t-\Delta t, t)$ intervallumba essék;
II. $k-1$ darab megfigyelési érték $<t$, a többi $n-k$ megfigyelt érték $\geq t$.

Annak a valószínűsége, hogy $X_1, X_2, \dots, X_n$ közül *egy kiszemelt* esik a $(t-\Delta t, t)$ intervallumra, közelítőleg $f(t)\Delta t$. Annak a valószínűsége, hogy az n érték *valamelyike* (vagy az első vagy a második ... vagy az n-edik) esik ide,

$$nf(t)\Delta t$$

(I. alatti esemény valószínűsége).

A II. alatti esemény valószínűsége a binomiális eloszlás figyelembevételével

$$\binom{n-1}{k-1}[F(t)]^{k-1}[1-F(t)]^{n-k}.$$

A független események együttes bekövetkezésének a valószínűsége a valószínűségek szorzata. Ezért az I. és II. alatti események együttes bekövetkezésének a valószínűsége jó közelítésben:

$$n\binom{n-1}{k-1}[F(t)]^{k-1}[1-F(t)]^{n-k}f(t)\Delta t.$$

Ez most annak a közelítő valószínűsége, hogy X_k^* *egy bizonyos* $(t-\Delta t, t)$ intervallumra essék. Annak a valószínűsége, hogy *valamelyik* intervallumba essék, az egymást kizáró, fenti jellegű valószínűségek összege:

$$F_{n,k}(x) \approx \sum_{t=-\infty}^{x} n\binom{n-1}{k-1}[F(t)]^{k-1}[1-F(t)]^{n-k}f(t)\Delta t.$$

A közelítés annál pontosabb, minél kisebb Δt. Pontos értéket akkor kapunk, ha a jobb oldalon álló kifejezés határértékét vesszük, midőn $\Delta t \to 0$ (ez pedig egy határozott integrál lesz):

$$F_{n,k}(x) = \lim_{\Delta t \to 0} \sum_{t=-\infty}^{x} n\binom{n-1}{k-1}[F(t)]^{k-1}[1-F(t)]^{n-k}f(t)\Delta t,$$

$$F_{n,k}(x) = n\binom{n-1}{k-1}\int_{-\infty}^{x}[F(t)]^{k-1}[1-F(t)]^{n-k}f(t)\,dt.$$

A *c)*-ben adott képleteket az elméleti tudnivalók között igazoltuk.

26. Legyen X egyenletes eloszlású a $(0; 1)$ intervallumon. Adjuk meg X_1^*, X_n^*, X_k^* rendezett mintaelemek eloszlásfüggvényét.

Egyenletes eloszlás esetén X eloszlásfüggvénye:

$$F(x) = \begin{cases} 0, & \text{ha} \quad x \leq 0, \\ x, & \text{ha} \quad 0 < x \leq 1, \\ 1, & \text{ha} \quad 1 < x. \end{cases}$$

A *c)*-ben kapott képletekkel számolva:

$$P(X_1^* < x) = F_{n,1}(x) = n\int_0^x (1-u)^{n-1}du = 1-(1-x)^n,$$

$$P(X_n^* < x) = F_{n,n}(x) = n\int_0^x u^{n-1}du = x^n,$$

$$P(X_k^* < x) = F_{n,k}(x) = n\binom{n-1}{k-1}\int_0^x u^{k-1}(1-u)^{n-k}du.$$

Ez azt mutatja, hogy X_k^* ún. bétaeloszlású, k és $n-k+1$ paraméterrel. Ennek segítségével bizonyítható, hogy a k-adik rendezett mintaelem várható értéke és szórásnégyzete:

$$M(X_k^*) = \frac{k}{n+1} \qquad (k = 1, 2, \dots, n),$$

$$D^2(X_k^*) = \frac{k(n-k+1)}{(n+1)^2(n+2)}.$$

Megjegyzés

Az X valószínűségi változót (p, q) paraméterű, *bétaeloszlásúnak* nevezzük, ha sűrűségfüggvénye ($p>0$, $q>0$ esetén):

$$f(x) = \begin{cases} \dfrac{\Gamma(p+q)}{\Gamma(q)\Gamma(q)} x^{p-1}(1-x)^{q-1}, & \text{ha } 0<x<1, \\ 0 & \text{különben.} \end{cases}$$

Várható értéke és szórása:

$$M(X) = \frac{p}{p+q},$$

$$D(X) = \frac{1}{p+q}\sqrt{\frac{pq}{p+q+1}}.$$

Az eloszlás használatát táblázatok könnyítik meg.

27. Igazoljuk, hogy a mintaátlag sztochasztikusan konvergál a várható értékhez!

Alkalmazzuk a Csebisev-egyenlőtlenséget

$$\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}, \quad M(\bar{X}) = m, \quad D^2(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

esetén:

$$P\left(\left|\frac{\Sigma x_i}{n} - m\right| \geqq \varepsilon\right) \leqq \frac{\sigma^2}{\varepsilon^2 n}.$$

Ha $n \to \infty$, akkor $\dfrac{\sigma^2}{\varepsilon^2 n} \to 0$, ennélfogva

$$P\left(\left|\frac{\Sigma x_i}{n} - m\right| \geqq \varepsilon\right) \to 0.$$

Ez épp az igazolandó sztochasztikus konvergenciát bizonyítja. (Ebből következik, hogy egy esemény relatív gyakorisága is sztochasztikusan konvergál az esemény valószínűségéhez.)

2. STATISZTIKAI BECSLÉSEK

2.1. Problémafelvetés

Gyakori, hogy egy X valószínűségi változóról elméleti meggondolások vagy mérések alapján tudjuk, hogy normál eloszlású, vagy eloszlása jól közelíthető normál eloszlással, vagyis sűrűségfüggvénye:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

azonban m és σ olyan ismeretlen paraméterek, amelyeket statisztikai mintából kell becsülnünk. Ezért m és σ is véletlen változó, így tüntetik fel, hogy $f(x)$ tőlük is függ:

$$f(x) = f(x; m, \sigma).$$

Általánosítva: a statisztikai gyakorlatban sokszor előfordul, hogy *a vizsgált X véletlen változó eloszlása ismert* (ismerjük az $F(x)$ eloszlásfüggvényt vagy az $f(x)$ sűrűségfüggvényt), *azonban a sűrűségfüggvény vagy az eloszlásfüggvény olyan ismeretlen,* $T_1, T_2, \ldots, T_m$ *paramétertől függ, amelyeket a statisztikai mintából kell becsülnünk,* tehát

$$f(x) = f(x; T_1, T_2, \ldots, T_m),$$

$$F(x) = F(x; T_1, T_2, \ldots, T_m).$$

A T_i paraméter becslését így *jelöljük:* $\hat{T}_i$ $(i = 1, 2, \ldots, m)$, $\hat{T}_i$ a mintaelemek függvénye (véletlen változó).

Az eloszlás egyik ismeretlen a paraméterét becsülhetjük:

– egyetlen számértékkel, mint azt az előbb jelöltük (ez ún. *á pontbecslés*), az a pontbecslése a mintaelemek valamely $\alpha(X_1, X_2, \ldots, X_n)$ függvénye (maga is valószínűségi változó),

– egy (α_1, α_2) intervallummal, amely nagy valószínűséggel tartalmazza a-t *(intervallumbecslés)*. Ilyenkor tehát

$$P(\alpha_1 \leqq a < \alpha_2) = 1 - \varepsilon,$$

ahol ε a 0-hoz közeleső, kis valószínűség. $1-\varepsilon$ a fenti (α_1, α_2) *megbízhatósági (konfidencia-) intervallumhoz* tartozó valószínűség; %-ban kifejezett értéke: $100(1-\varepsilon)\%$ a *biztonsági szint*.

Megjegyezzük, hogy α_1 is, α_2 is a mintaelemek függvénye.

A következőkben az ilyen *pontbecslések* módszereiről és tulajdonságairól fogunk beszélni.

2.2. A pontbecslés módszerei

Több olyan módszer létezik, amellyel az ismeretlen $T_i = T_i(X_1, X_2, \ldots, X_m)$ paraméter „jól" becsülhető.

a) A momentumok módszere. Az elméleti momentumokat a kérdéses paraméterek függvényeként kifejezve, a megfelelő empirikus momentumokkal tesszük egyenlővé. Így az ismeretlen paraméterekre sok esetben könnyen megoldható egyenletrendszert nyerünk.

R. A. Fisher kimutatta, hogy a momentumokkal való paraméterbecslés erősen aszimmetrikus eloszlások esetén kevésbé hatékony.

Normál eloszlásnál a momentummódszer egyenértékű a most következő maximum-likelihood eljárással.

b) A maximum-likelihood módszer. A könnyebb megértés kedvéért *példával* kezdjük. Egy textilgyárban a fonalak szakadása egymástól független, kimutatható, hogy ilyenkor egy adott időtartam alatti fonalszakadások száma, X jó közelítésben Poisson-eloszlású.

I. Milyen becslést adhatunk az ismeretlen λ paraméterre, ha a fent jelzett időtartamban a fonalszakadások száma $X=k$ volt?

Célszerű λ-nak olyan értéket adni, amely esetén a fenti esemény $(X=k)$ valószínűsége maximális.

A maximális valószínűség angolul: *maximum-likelihood.* Az $L = L(k; \lambda)$ likelihood-függvény maximumát keressük:

$$L(k; \lambda) = P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

$$\frac{dL}{d\lambda} = \frac{1}{k!}(k\lambda^{k-1}e^{-\lambda} - \lambda^k e^{-\lambda}) = \frac{\lambda^k e^{-\lambda}}{k!}\left(\frac{k}{\lambda} - 1\right) = L\left(\frac{k}{\lambda} - 1\right) = 0,$$

mivel $L = \dfrac{\lambda^k e^{-\lambda}}{k!} > 0$, ezért

$$\frac{k}{\lambda} = 1,$$

$$\lambda \doteq k.$$

Minthogy $\lambda = k$ esetén a második derivált negatív, ezért biztosan maximális a minta valószínűsége, ha $\lambda = k$. Valóban:

$$\frac{d^2L}{d\lambda^2} = L'\left(\frac{k}{\lambda} - 1\right) + L\left(-\frac{k}{\lambda^2}\right) < 0.$$

Megmutatjuk, hogy még könnyebb a *maximum-likelihood függvény logaritmusának,* ln L-nek a maximumát megkapni:

$$\ln L = k \ln \lambda - \lambda - \ln k!,$$

$$\frac{d \ln L}{d\lambda} = \frac{k}{\lambda} - 1 = 0,$$

$$\lambda \doteq k,$$

$$\frac{d^2 \ln L}{d\lambda^2} = -\frac{k}{\lambda^2} < 0.$$

Természetesen L-nek és logaritmusának ugyanazon helyen van ugyanolyan jellegű szélsőértéke, hiszen a logaritmusfüggvény szigorúan monoton növekedő függvény. Mivel $L>0$, ezért $\ln L$ mindig értelmezve van.

– A probléma a valóságban nem egyetlen időtartam alatti vizsgálatként merül fel. Inkább úgy, hogy n mérési intervallumban a fonalszakadások száma:

$$X_1=k_1,\ X_2=k_2,\ \ldots,\ X_n=k_n,$$

milyen λ paraméterérték esetén maximális a fent kapott mintának a valószínűsége? A fonalszakadások független volta miatt:

$$L = L(k_1, k_2, \ldots, k_n; \lambda) = P(X_1=k_1, X_2=k_2, \ldots, X_n=k_n) =$$

$$= P(X_1=k_1)\,P(X_2=k_2)\ldots P(X_n=k_n) = \prod_{i=1}^{n} \frac{\lambda^{k_i}e^{-\lambda}}{k_i!} =$$

$$= e^{-n\lambda}\prod_{i=1}^{n}\frac{\lambda^{k_i}}{k_i!}.$$

Itt is egyszerűbb, ha L helyett $\ln L$ maximumát számoljuk:

$$\ln L = -n\lambda + \sum_{i=1}^{n}(k_i \ln\lambda - \ln k_i!),$$

$$\frac{d\ln L}{d\lambda} = -n + \sum_{i=1}^{n}\frac{k_i}{\lambda} = 0,$$

$$\lambda = \frac{\sum_{i=1}^{n} k_i}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{x},$$

$$\frac{d^2\ln L}{d\lambda^2} = -\frac{1}{\lambda^2}\Sigma k_i < 0.$$

A λ paraméter maximum-likelihood becslése tehát a mintaátlag. Ez azt jelenti, hogy *$\lambda=\bar{x}$ esetén maximális annak a valószínűsége, hogy az $X_1=k_1, X_2=k_2, \ldots, X_n=k_n$ mintát kapjuk, célszerű tehát λ-t $\bar{x}$-gal becsülni.*

Megjegyzés

A második derivált negatív voltának igazolása hozzátartozik a maximum létezésének igazolásához. Mégis, a gyakorlatban ezt néha az Olvasóra bízzák. Ha nem igazoljuk, akkor csak annyit mondhatunk: a szóban forgó helyen *lehet* a valószínűség maximuma (nem biztos, hogy van).

– Általánosítva a fentieket, *a maximum-likelihood módszer a következő.*

Ismerjük a sokaság eloszlását, de nem ismerjük az eloszlást jellemző paramétert (vagy paramétereket). A paraméter(ek) értékét olyan érték(ek)kel becsüljük, amely(ek) esetén az adott minta bekövetkezése volna a legnagyobb valószínűségű. A maximális valószínűséget az adott minta valószínűségét megadó likelihood-függvény maximumával vagy a logaritmusának a maximumával keressük meg.

Egy T paraméter esetén a likelihood-függvény a következő.

I. *Diszkrét esetben:*

$$L(X_1, X_2, \ldots, X_n, T) = P(X_1=x_1, X_2=x_2, \ldots, X_n=x_n) =$$

$$= P(X_1=x_1)\,P(X_2=x_2)\ldots P(X_n=x_n).$$

Szorzat helyett könnyebb összeget kezelni, tehát gyakran az

$$\ln L = \sum_{i=1}^{n}\ln P(X_i=x_i)$$

függvényt tekintik likelihood-függvénynek.

II. *Folytonos esetben* egy pont felvételének valószínűsége az n-dimenziós térben 0. Itt annak a valószínűségét kell maximálissá tenni, hogy a pont az $(x_1, x_2, \ldots, x_n)$ pont közvetlen környezetébe, pontosabban az

$$x_1 \leqq X_1 \leqq x_1+\Delta x_1,\ \ldots,\ x_n \leqq X_n \leqq x_n+\Delta x_n$$

n-dimenziós téglatestbe essék. Ennek valószínűsége:

$$f(x_1, T)\,f(x_2, T)\ldots f(x_n, T)\,\Delta x_1\,\Delta x_2\ldots\Delta x_n.$$

Ez ott maximális, ahol az

$$L(x_1, x_2, \ldots, x_n; T) = f(x_1, T)f(x_2, T) \ldots f(x_n, T)$$

függvény maximális. Ezt vagy ennek logaritmusát: az

$$\ln L = \sum_{i=1}^{n} \ln f(x_i, T)$$

függvényt tekintik likelihood-függvénynek, ennek megfelelően a

$$\frac{dL}{dT} = 0 \quad \text{vagy a} \quad \frac{d \ln L}{dT} = 0$$

egyenlet a likelihood-egyenlet.

– Több T_i paraméter esetében ugyanígy definiálható a likelihood-függvény. Ilyenkor a parciális deriváltak zérushelyeit keressük (ahol *lehet* a maximum). Hogy biztosan maximum *van-e,* azt a Hesse-determináns vagy egyéb vizsgálat dönti el, amit néha az Olvasóra bízunk.

A legnagyobb valószínűség elve elég könnyen teljesülő feltételek mellett aszimptotikusan konzisztens becsléshez vezet, amely elég nagy n-re közelítőleg normál eloszlású (l. később).

c) Fontos becslési módszer a *legkisebb négyzetek elve* (l. a „Korreláció- és regresszióelemzés" c. fejezetben).

2.3. A pontbecslés tulajdonságai

a) A $\hat{T} = \hat{T}(X_1, X_2, \ldots, X_n)$ statisztika *torzítatlan becslése* a T paraméternek, ha a $\hat{T}$ értékek átlaga T körül ingadozik, pontosabban, ha $\hat{T}$ várható értéke maga a becsült paraméter értéke:

$$M(\hat{T}) = T.$$

Aszimptotikusan torzítatlan a becslés, ha a fenti egyenlet csak határértékben (nagy n-re közelítőleg) áll fenn:

$$\lim_{n \to \infty} M(\hat{T}) = T.$$

b) Ésszerű kívánalom, hogy $\hat{T}$ ingadozásai ne legyenek túlságosan nagyok, és $\hat{T}$ szórása kicsi legyen. Két becslő statisztikai függvény közül *hatásosabb (efficiensebb) becslés* az, amelynek szórása (T minden lehetséges értéke mellett) kisebb a másikénál. Vagyis, ha

$$D_T^2(\hat{T}_1) \leqq D_T^2(\hat{T}_2)$$

minden T-re. Ekkor T_1 a hatásosabb becslés.

Ha T torzítatlan becslései között létezik egy legkisebb szórásnégyzetű $\hat{T}_0$ becslés, akkor ez a *leghatékonyabb vagy röviden: hatásos torzítatlan becslés* (minden lehetséges T esetén). Kimutatható, hogy ha van ilyen, akkor csak ez az egy ilyen becslés létezik. Ehhez viszonyítva adjuk meg egy tetszőleges $\hat{T}_1$ becslés hatékonyságindexét:

$$0 < \frac{\inf\limits_{T_0} D_T^2(\hat{T}_0)}{D_T^2(\hat{T}_1)} \leqq 1.$$

A becslés hatásfokát nem mindig vonatkoztatjuk az összes létező torzítatlan becslésre, hanem csak a becslések egy olyan szűkebb osztályára, amelyen belül rendszerint megtalálható a leghatásosabb becslés. Ilyenkor a szóban forgó *osztályra vonatkoztatott hatásfokról* beszélünk.

c) Harmadik elvárás, hogy növelve a mintaelemek számát, a becslés az előző szempontokból javuljon. *Konzisztens* a $\hat{T}$ becslés, ha $n \to \infty$ esetén $\hat{T}$ sztochasztikusan T-hez konvergál:

$$\lim_{n \to \infty} P(|\hat{T} - T| > \varepsilon) = 0$$

minden ε-ra. A konzisztens becslés nem feltétlen torzítatlan, de mindig aszimptotikusan torzítatlan.

A gyenge és erős konvergencia mintájára itt is beszélünk a konzisztens becslésen kívül erősen konzisztens becslésről. A T paraméter torzítatlan becslései közül *erősen konzisztens* az a $\hat{T}$ becslés, amelynek szórásnégyzete zérushoz tart, ha $n \to \infty$, azaz

$$D^2(\hat{T}) = M[(\hat{T} - T)^2] \to 0, \quad \text{ha} \quad n \to \infty.$$

Az erős konzisztenciából következik a konzisztencia (de fordítva nem feltétlenül). Ugyanis a Csebisev-egyenlőtlenség alapján:

$$P(|\hat{T}-T| > \varepsilon) < \frac{D^2(\hat{T})}{\varepsilon^2},$$

ha n minden határon túl való növelésével $D^2(\hat{T}) \to 0$, akkor

$$P(|\hat{T}-T| > \varepsilon) \to 0,$$

ami azt jelenti, hogy erősen konzisztens becslés egyúttal konzisztens is.

d) Erősen konzisztens becslés szórásnégyzete zérushoz tart, ha $n \to \infty$. Vajon *rögzített n* esetén van-e olyan becslés, amelynek tetszőleges kicsi a szórásnégyzete, vagy van egy olyan alsó korlát, amelynél a szórásnégyzet nem lehet kisebb? A *Cramer–Rao-egyenlőtlenség* szerint van ilyen alsó korlát. Ha a $\hat{T} = \hat{T}(X_1, X_2, \dots, X_n)$ statisztika torzítatlan becslés (T helyett mindjárt általánosabban) a $g(t)$ függvényre, akkor

$$D^2(\hat{T}) \geqq \frac{[g'(t)]^2}{I},$$

ahol I a Fisher-féle információmennyiség:

$$I = M\left[\left(\frac{d\ln f}{dT}\right)^2\right] = M\left[\left(\frac{1}{f}\,\frac{df}{dT}\right)^2\right].$$

Itt f a mintaelemek együttes sűrűségfüggvénye:

$$f = f(x_1, x_2, \dots, x_n;\, t) = \prod_{i=1}^{n} f(x_i;\, t).$$

A várható értéknél $x_1, x_2, \dots, x_n$ konkrét értékek helyett az $X_1, X_2, \dots, X_n$ véletlen változókat kell venni. A

$$\frac{[g'(t)]^2}{I}$$

mennyiség az információs határ. Van, amikor ezt elérhetjük (legjobb, torzitatlan becslés), van, amikor nem. Ha $\hat{T}$ magának T-nek a torzítatlan becslése, $g(t) = t$, akkor $g'(t) = 1$, az információs határ $\frac{1}{I}$.

Megmutatható, hogy azok a sűrűségfüggvények, amelyek esetén a minimális szórásnégyzet eléri az információs határt, a következő alakúak:

$$f(x_1, x_2, \dots, x_n;\, t) = f_1(x_1, x_2, \dots, x_n)e^{A(x_1, x_2, \dots, x_n)A_1(t) + A_2(t)}$$

(Tehát f_1 és A csak a változók, A_1 és A_2 csak a t paramétertől függenek.) Ez a feltétel szükséges, de nem elégséges. Mindenesetre ebből várható, hogy a normális, exponenciális, Poisson-eloszlások esetén esetleg a legjobb, torzítatlan becslést kapjuk.

e) Az *elégséges becslés* fogalmát célszerű példával kezdeni.

Legyen X $N(m, \sigma)$-eloszlású változó, a független $X_1, X_2, \dots, X_n$ mintaelemekből számított $\bar{X}$ becslés az ismeretlen m várható értékre. Kérdés, hogy minden információt tartalmaz-e, vagyis az egyes mintaelemek nem adnak-e több információt, mintha csak $\bar{X}$-gal becsüljük m-et.

A normál eloszlású X változó $\bar{X}$ átlaga $N(m, \sigma/\sqrt{n})$-eloszlású, $\bar{X}$ sűrűségfüggvénye:

$$f_1(\bar{x}) = \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}}\, e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}}.$$

Ez $\bar{X}$ eloszlását meghatározza.

Az $X_1, X_2, \dots, X_n$ független, $N(m, \sigma)$-eloszlású mintaelemek együttes sűrűségfüggvénye azok együttes eloszlását határozza meg:

$$f(x_1, x_2, \dots, x_n;\, m, \sigma) = \prod_{i=1}^{n} f(x_i;\, m, \sigma) =$$

$$= \frac{1}{(\sqrt{2\pi}\sigma)^n}\, e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k - m)^2}$$

Könnyen belátható, hogy

$$\sum_{k=1}^{n}(x_k-m)^2 = \sum_{k=1}^{n}(x_k-\bar{x})^2+n(\bar{x}-m)^2.$$

Így az együttes eloszlást meghatározó, együttes sűrűségfüggvény a következőképpen írható:

$$f(x_1, x_2, \ldots, x_n;\ m, \sigma) =$$

$$= \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}}e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}}\frac{1}{\sqrt{n}(\sqrt{2\pi}\sigma)^{n-1}}e^{-\frac{\Sigma(x_k-\bar{x})^2}{2\sigma^2}}.$$

Ha az $\bar{X}=\bar{x}$ mintabeli érték, akkor az $X_1, X_2, \ldots, X_n$ együttes feltételes sűrűségfüggvény határozza meg a mintaelemek eloszlását; ez az előző két függvény hányadosa:

$$f(x_1, x_2, \ldots, x_n;\ \sigma/\bar{X}=\bar{x}) = \frac{1}{\sqrt{n}(\sqrt{2\pi}\sigma)^{n-1}}e^{-\frac{\Sigma(x_k-\bar{x})^2}{2\sigma^2}}.$$

Nem véletlenül hagytuk ki az argumentumokból m-et: *a mintaelemek együttes, feltételes sűrűségfüggvénye, vagyis a mintaelemek eloszlása az m paramétertől független.* Az m-re nem kapunk több információt a mintaelemekből, mint amit az $\bar{x}$ becslésből kaptunk. *Az $\bar{x}$ mintaközép tehát elégséges becslés* az m várható értékre (normál eloszlás esetén).

A feladatot általánosítjuk. Ha a mintaelemekből olyan, a T paramétert becslő $\hat{T}$ statisztikai függvényt képeztünk, amely a becsült paraméterre minden információt tartalmaz, akkor azt *elégséges becslésnek* nevezzük. Ilyenkor a mintaelemek együttes sűrűségfüggvénye felbontható egy T-t és $\hat{T}$-t tartalmazó $g(\hat{T}, T)$ függvény és egy T-től független függvény szorzatává:

$$f(x_1, x_2, \ldots, x_n;\ T) = f(x_1, x_2, \ldots, x_n/T=t)g(\hat{T};\ T),$$

így a feltételes, együttes sűrűségfüggvény

$$f(x_1, x_2, \ldots, x_n/T=t) = \frac{f(x_1, x_2, \ldots, x_n;\ T)}{g(\hat{T}, T)}$$

már nem függ T-től (nem tartalmaz több információt T-re).

Bizonyítható, hogy ha $\hat{T}$ a $g(T)$ torzítatlan becslése és $D^2(\hat{T}) = g'^2(t)/I$, akkor $f(x_1, x_2, \ldots, x_n;\ T)$ a fenti szorzat alakra hozható és így $\hat{T}$ elégséges statisztika.

2.4. Intervallumbecslések

Az eddigiek során egy eloszlás egy ismeretlen T paraméterét egyetlen $\hat{T}$ számadattal becsültük. Az így kapott pontbecslés helyett a most tárgyalandó intervallumbecslést többnek, alaposabb információnak érezzük: a pontbecslés egyetlen értéket ad, és nem tudható, hogy ez hány százalék valószínűséggel (biztonsággal) fog bekövetkezni (egy véletlen változó egyetlen értéket 0% valószínűséggel vesz fel folytonos esetben). Az intervallumbecslésnél viszont

– figyelembe vesszük, hogy $\hat{T} = \hat{T}(X_1, X_2, \ldots, X_n)$ maga is véletlen változó, amelynek értékei szóródnak a T paraméter valódi értéke körül;

– ha ismerjük $\hat{T}$ eloszlásfüggvényét, azt is meg tudjuk mondani, *hány százalék valószínűséggel (biztonsági szinten) tartalmazza az* (α_1, α_2) *intervallum T valódi értékét.*

Az előre megadott (lehetőleg kicsiny) ε-hoz tartozó, (α_1, α_2), $1-\varepsilon$-*szintű konfidencia- (megbízhatósági) intervallumot* tehát az alábbi egyenlet definiálja:

$$P(\alpha_1 \leqq T \leqq \alpha_2) = 1-\varepsilon.$$

Az $1-\varepsilon$ az ún. *valószínűségi szint*, %-ban kifejezett értéke a *biztonsági szint*. Persze, α_1 és α_2 is véletlen változók (a mintaelemek függvényei):

$$\alpha_1 = \alpha_1(X_1, X_2, \ldots, X_n), \qquad \alpha_2 = \alpha_2(X_1, X_2, \ldots, X_n).$$

a) Megbízhatósági intervallum a normál eloszlás m várható értékére, ha ismert az elméleti σ szórás. Az m várható értéket az $\bar{X}$ mintaközéppel becsülve, $\bar{X}$ ugyancsak normál eloszlású, m várható értékkel és $\frac{\sigma}{\sqrt{n}}$

szórással. Az

$$\frac{\bar{X}-m}{\sigma/\sqrt{n}}$$

változó már y tengelyre szimmetrikus, standard normál eloszlású. A szimmetria miatt célszerű a konfidencia-intervallumot a 0-ra szimmetrikusan, $-u_\varepsilon$ és $+u_\varepsilon$ határokkal keresni úgy, hogy

$$P\left(-u_\varepsilon \leqq \frac{\bar{x}-m}{\sigma/\sqrt{n}} \leqq u_\varepsilon\right) = 2 \cdot \Phi(u_\varepsilon)-1 = 1-\varepsilon$$

legyen (ε előre adott, kis valószínűség). Az egyenlőtlenség-rendszert rendezve:

$$P\left(\bar{x}-u_\varepsilon\frac{\sigma}{\sqrt{n}} \leqq m \leqq \bar{x}+u_\varepsilon\frac{\sigma}{\sqrt{n}}\right) = 1-\varepsilon.$$

Az így kapott $\bar{x}$-ra szimmetrikus

$$\left(\bar{x}-u_\varepsilon\frac{\sigma}{\sqrt{n}};\ \bar{x}+u_\varepsilon\frac{\sigma}{\sqrt{n}}\right)$$

megbízhatósági intervallum az eseteknek átlagosan $(1-\varepsilon) \cdot 100\%$-ában tartalmazza az ismeretlen m várható értéket, ha az intervallumbecslést elég sokszor elvégezzük. Ha $\varepsilon = 0{,}05$, akkor átlagosan $\varepsilon \cdot 100\% = 5\%$-ban tévedünk, mikor az előbbi konfidencia-intervallumon keressük m-et. Ez ún. $1-\varepsilon$-szintű konfidencia-intervallum m-re.

Megjegyzés

Szokásos (és jogos) kérdések az előbbiek elhangzásakor:

- honnan ismerjük σ-t,
- hogyan számoljuk ki u_ε-t?

Az elméleti szórást vagy előzetes, nagyon sok mérés alapján tételezzük fel, vagy szabványkövetelményként írjuk elő.

Az u_ε meghatározása a

$$P\left(-u_\varepsilon \leqq \frac{\bar{X}-m}{\sigma/\sqrt{n}} \leqq u_\varepsilon\right) = \Phi(u_\varepsilon)-\Phi(-u_\varepsilon) =$$
$$= 2 \cdot \Phi(u_\varepsilon)-1 = 1-\varepsilon$$

egyenlet „végéről", a Φ függvény táblázatából visszakereséssel történik. Pl. $1-\varepsilon = 0{,}95$ esetén:

$$\begin{aligned} 2 \cdot \Phi(u_\varepsilon)-1 &= 0{,}95,\\ \Phi(u_\varepsilon) &= 0{,}975,\\ u_\varepsilon &= 1{,}96 \end{aligned}$$

(l. az 5. táblázatot):

x	$\varphi(x)$	$\Phi(x)$
1,96		0,9750

Mivel azonban a Student-eloszlás „határhelyzete" a normál eloszlás, ha $n \to \infty$, ezért u_ε a 8. táblázatból is kereshető, az $n \to \infty$ sorból:

n	Statisztikai biztonság		
	90%	95%	98%
∞		1,960	

b) Megbízhatósági intervallum a normál eloszlás ismeretlen m várható értékére, ha csak a mintabeli s szórás ismeretes.* Kimutatható, hogy

$$\frac{\bar{X}-m}{s^*/\sqrt{n}}$$

Student-eloszlású, $n-1$ szabadságfokkal. Ezért most a

$$P\left(-t_\varepsilon \leqq \frac{\bar{X}-m}{s^*/\sqrt{n}} \leqq t_\varepsilon\right) = 1-\varepsilon$$

egyenlet m-re az

$$\left(\bar{x}-t_\varepsilon\frac{s^*}{\sqrt{n}};\ \bar{x}+t_\varepsilon\frac{s^*}{\sqrt{n}}\right)$$

megbízhatósági intervallumot adja, amely szintén $\bar{x}$-ra szimmetrikus.

Például $n=30$, 95% biztonsági szint esetén a 8. táblázatból:

$n=f+1$	Statisztikai biztonság		
	90%	95%	98%
30	1,699	2,045	2,462

$t_\varepsilon=2{,}045$, ennek alapján a 95% szintű konfidencia-intervallum m-re:

$$\left(\bar{x}-2{,}045\cdot\frac{s^*}{\sqrt{30}};\ \bar{x}+2{,}045\cdot\frac{s^*}{\sqrt{30}}\right).$$

Megjegyzés

1. Ha a biztonsági szintet alacsonyabbra vesszük, a konfidencia-intervallum keskenyedik, ha magasabbra vesszük, az intervallum szélesedik (l. az előbbi táblázatban a különböző t értékeket). Ez érthető, mert ha pl. az esetek nagyobb %-ában tartalmazó intervallumot akarunk kapni, az nyilván szélesebb intervallum lesz (a változó nagyobb intervallumba nagyobb valószínűséggel esik).

2. Észrevehető, hogy ugyanazon biztonsági szinthez, ugyanazon $s^*=\sigma$ szóráshoz, azonos n-hez tartozó $(-u_\varepsilon;\ u_\varepsilon)$ konfidencia-intervallum keskenyebb, mint a $(-t_\varepsilon;\ t_\varepsilon)$ intervallum ($u_\varepsilon=1{,}96 < t_\varepsilon=2{,}045$).

Ez is érthető, hiszen a σ szórás ismerete több információt jelent, mint s^*-é (mint ha előzetesen lett volna „végtelen" sok mérésünk, amelyből σ-t ismerjük).

3. Ha m-ről semmit nem tudunk, célszerű mindkét esetben $\bar{x}$-ra szimmetrikus intervallumot keresni. A statisztika könyvek azonban nagyon gyakran megfeledkeznek arról, hogy néha a gyakorlatban tudjuk: nagy valószínűségű, hogy

$$m>\bar{x}.$$

Ilyenkor célszerűbb $u_2>u_\varepsilon$-nal $\bar{x}$-ra nem szimmetrikus

$$\left(\bar{x}-u_1\frac{s^*}{\sqrt{n}};\ \bar{x}+u_2\frac{s^*}{\sqrt{n}}\right)$$

megbízhatósági intervallumot keresni.

Nem kötelező tehát, hogy $\bar{x}$-ra szimmetrikus megbízhatósági intervallumot keressünk (még szimmetrikus eloszlás esetén sem). *A megbízhatósági intervallum megadása nem egyértelmű* (azonos n, szórás, megbízhatósági szint esetén sem).

Erre az észrevételre a matematikai statisztikai próbáknál még visszatérünk.

c) Konfidencia-intervallum $N(m;\sigma)$-eloszlás ismeretlen szórásnégyzetére. A szóban forgó eloszlás esetén $n\dfrac{S_n^2}{\sigma^2}$ véletlen változó χ^2-eloszlású, $(n-1)$ paraméterrel. Ha a 7. táblázatból úgy határozzuk meg a c_1 és c_2 értékeket, hogy (adott ε esetén) c_1 alá a változó $\dfrac{\varepsilon}{2}$, c_2 fölé $1-\dfrac{\varepsilon}{2}$ valószínűséggel essék, akkor nyilván c_1 és c_2 közé $1-\varepsilon$ valószínűséggel esik a változó:

$$P\left(c_1 \leqq n\frac{S_n^2}{\sigma^2} \leqq c_2\right) = 1-\varepsilon.$$

Ebből pedig kiszámítható σ-ra az $1-\varepsilon$-szintű konfidencia-intervallum:

$$\sqrt{n}\frac{S_n}{\sqrt{c_2}} \leqq \sigma \leqq \sqrt{n}\frac{S_n}{\sqrt{c_1}}.$$

d) Megbízhatósági intervallum az exponenciális eloszlás λ paraméterére. Az $F(x) = 1-e^{-\lambda x}$ eloszlásfüggvény esetében a λ paraméter maximum-likelihood (pont)becslése:

$$\lambda \doteq \sum_{i=1}^{n} \frac{X_i}{n} = \bar{X}.$$

A mintaelemek független, exponenciális eloszlású változók, szintén λ paraméterrel, tehát

$$P(\lambda X_i < x) = F\left(X_i < \frac{x}{\lambda}\right) = 1-e^{-x},$$

λX_i sűrűségfüggvénye:

$$f(x) = e^{-x}, \quad (i = 1, 2, \ldots, n),$$

ez exponenciális eloszlást jelent $\lambda = 1$ paraméterrel.

A gammaeloszlás bevezetésekor említettük, hogy független, exponenciális eloszlású, azonos a paraméterű, p darab véletlen változó öszege gammaeloszlású (a, p) paraméterekkel. Ezért az

$$n\lambda\bar{X} = \sum_{i=1}^{n} \lambda X_i$$

véletlen változó gammaeloszlású,

$$g(x) = \frac{1}{\Gamma(n)} x^{n-1} e^{-x}$$

sűrűségfüggvénnyel,

$$P(n\lambda\bar{X} < x) = G(x) = \int_0^x \frac{x^{n-1}}{(n-1)!} e^{-x}\,dx$$

eloszlásfüggvénnyel. Ennek segítségével kiszámítható olyan c_1 határ, amely alá $\frac{\varepsilon}{2}$ valószínűséggel, olyan c_2 határ, amely fölé $1-\frac{\varepsilon}{2}$ valószínűséggel esik az $n\lambda\bar{X}$ változó. Ebből a λ számára a

$$\frac{1}{n\bar{X}} c_1 \leqq \lambda \leqq \frac{1}{n\bar{X}} c_2$$

megbízhatósági intervallum adódik $(1-\varepsilon)\cdot 100\%$ biztonsági szinten.

Ha nem rendelkezünk gammaeloszlás-táblázattal, akkor a következő két módszert tudjuk ajánlani.

Mivel $2n\lambda\bar{X}$ változó sűrűségfüggvénye

$$\frac{1}{2} g\left(\frac{x}{2}\right) = \frac{\lambda}{2^n \Gamma(n)} x^{n-1} e^{-x},$$

ezért $2n\lambda\bar{X}$ változó $2n$ paraméterrel χ^2-eloszlású. Az előbbi pontban vázoltak szerint meghatározhatjuk rá a megbízhatósági intervallumot.

Ha viszont n elég nagy, akkor a centrális határeloszlás-tétel alapján $\bar{X}$ közelítőleg normál eloszlású, amelynek várható értéke és szórása:

$$M(\bar{X}) = M(X) = \frac{1}{\lambda}, \quad D(\bar{X}) = \frac{D(X)}{\sqrt{n}} = \frac{1}{\lambda\sqrt{n}}.$$

Így elég nagy n-re a standard normál eloszlás táblázata segítségével adhatunk λ számára nagy valószínűséggel tartalmazó intervallumot.

e) Konfidencia-intervallum a binomiális eloszlás ismeretlen p paraméterére. n független kísérletet végezve, a vizsgált A esemény ismeretlen p valószínűségére pontbecslés a

$$\hat{p} = \frac{k}{n}$$

relatív gyakoriság. A $k=n\hat{p}$ valószínűségi változó binomiális eloszlású,

$$M(n\hat{p}) = np, \quad D^2(n\hat{p}) = np(1-p)$$

várható értékkel, ill. szórásnégyzettel. A centrális határeloszlás-tétel szerint *elég nagy n esetén* a

$$\frac{n\hat{p}-np}{\sqrt{np(1-p)}}$$

standardizált véletlen változó jó közelítésben $N(0, 1)$-eloszlású, vagyis:

$$P\left(-x \leqq \sqrt{n}\frac{\hat{p}-p}{\sqrt{p(1-p)}} \leqq +x\right) \approx \Phi(x)-\Phi(-x) = 2\cdot\Phi(x)-1.$$

A zárójelben levő esemény rendezés után a

$$(\hat{p}-p)^2 \leqq x^2\frac{p(p-1)}{n}$$

egyenlőtlenségre vezet.

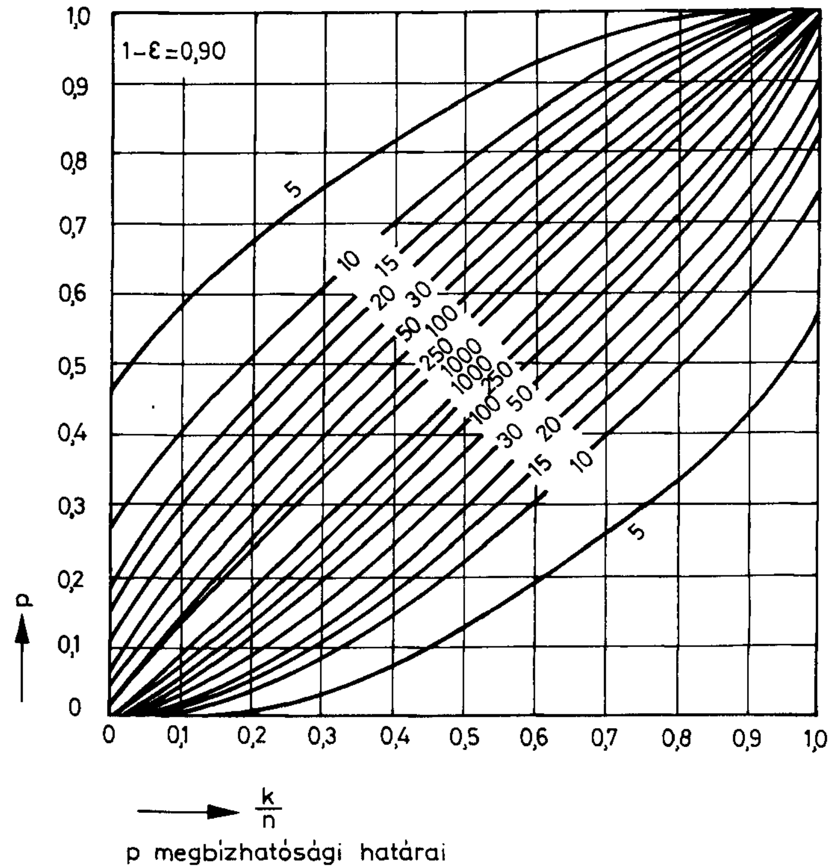
A másodfokú egyenlőtlenség megoldása p-re olyan $\hat{p}_1$ és $\hat{p}_2$ konfidencia-intervallum határokat ad meg, amelyekre:

$$P(\hat{p}_1 \leqq p \leqq \hat{p}_2) \approx 2\cdot\Phi(x)-1.$$

Itt $2\cdot\Phi(x)-1$ a biztonsági szint,

$$\hat{p}_{1,2} = \frac{\hat{p}+\dfrac{x^2}{2n}\pm\dfrac{x}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p})+\dfrac{x^2}{4n}}}{1+\dfrac{x^2}{n}}.$$

Ezzel tulajdonképpen egy A esemény p valószínűségére adhatunk intervallumbecslést. A feladat fontosságára való tekintettel *egy nomogramot* is adunk a gyors becsléshez.



18*a* ábra a)

A kísérletet n-szer függetlenül megismételve, p-re pontbecslés a $\frac{k}{n}$ relatív gyakoriság. Az adott n értékhez és különböző $\frac{k}{n}$ értékekhez az előbb vázolt módszerrel kiszámítható alsó, ill. felső intervallumhatárokat egy-egy görbével összekötötték. Így kapták a 18*a* ábrán látható nomogramot. A megbízhatósági szint 90%.

Ha pl. $n=100$, $k=20$, $\frac{k}{n}=0{,}20$, az ábrán az $n=100$-nak megfelelő két görbe 0,20 abszcisszájánál levő pontjának ordinátái: 0,14 és 0,28.

Tehát 90% valószínűséggel állíthatjuk, hogy az ismeretlen p valószínűséget a (0,14; 0,28) intervallum lefedi (vagyis tartalmazza).

Megjegyzés

A kapott másodfokú egyenlet gyökei $\hat{p} = \frac{k}{n}$-re szimmetrikusak, ha n elég nagy. A normál eloszlás azonban csak jó közelítés. Az ábra megszerkesztésénél nem végeztek közelítést. Ezek miatt a (0,14; 0,28) intervallum nem szimmetrikus $\frac{k}{n} = 0{,}20$-ra.

Kis és nagy n-ekre egyaránt használható intervallumbecslést p-re így is szerkeszthetünk.

A k gyakoriság binomiális eloszlású. Olyan c_1-et és c_2-t kell keresnünk, amelyre

$$P(c_1 \leqq k \leqq c_2) = \sum_{i=c_1}^{c_2} \binom{n}{i} p^i (1-p)^{n-i} = 1-\varepsilon$$

jó közelítésben teljesül. Ez a feltétel teljesül, ha az

$$P(k \leqq c_1) = \sum_{i=0}^{c_1} \binom{n}{i} p^i (1-p)^{n-i} = \frac{\varepsilon}{2},$$

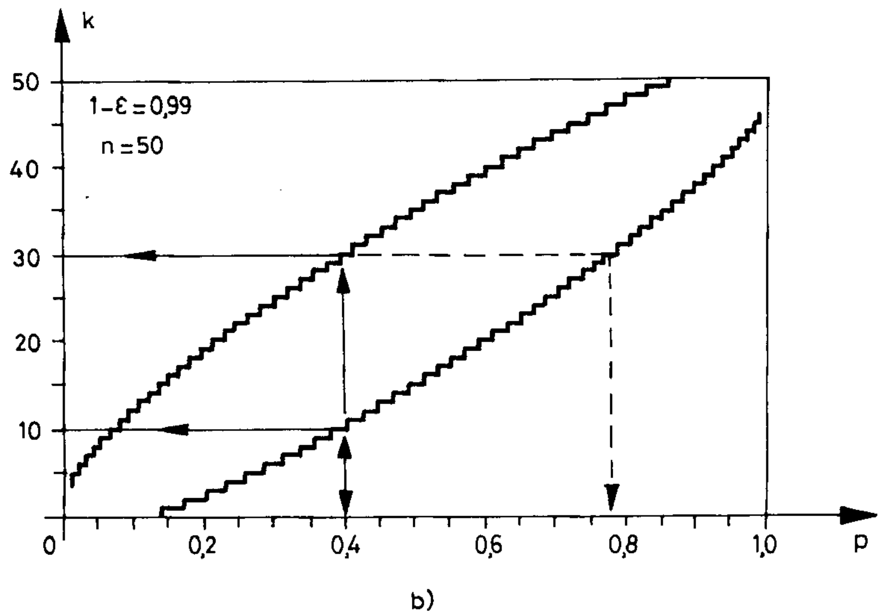
$$P(k \leqq c_2) = \sum_{i=0}^{c_2} \binom{n}{i} p^i (1-p)^{n-i} = 1-\frac{\varepsilon}{2}$$

egyenletek teljesülnek. Mivel p ismeretlen, p különböző, a (0; 1) intervallumon sűrűn választott értékeire kiszámítjuk a hozzájuk tartozó c_1 és c_2 határokat (18*b* ábra; $n = 50$ esetben, $1-\varepsilon = 0{,}99$ megbízhatósági szinten).

Ha most a kész ábrát fordítva használjuk, akkor $k = 30$ gyakoriság $\left(\frac{k}{n} = \frac{30}{50} = 0{,}60 \text{ relatív gyakoriság}\right)$ esetén p-re jó közelítésben

$$0{,}40 \leqq p \leqq 0{,}78$$

a megbízhatósági intervallum (l. az ábrát).



18*b* ábra

Így készült a 18*a* ábra is, csak több n-re, 90% biztonsági szinten.

f) Konfidencia-intervallum két normál eloszlású sokaság várható értékeinek eltérésére, ha az ismeretlen szórás a két sokaságban egyenlő.

Az $N(m_1, \sigma)$-eloszlású sokaságból vett n_1-elemű minta átlaga $\bar{x}_1$, korrigált szórása s_1^*, a másik, $N(m_2, \sigma)$-eloszlású sokaságból vett n_2-elemű minta átlaga $\bar{x}_2$, korrigált szórása s_2^*. Igazolható, hogy a következő statisztika,

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (m_1 - m_2)}{\sqrt{\frac{(n_1-1)s_1^{*2} + (n_2-1)s_2^{*2}}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Student-eloszlású valószínűségi változó, $f = n_1 + n_2 - 2$ szabadságfokkal. Így a 8. táblázatból meghatározhatók olyan $(-t_\varepsilon, t_\varepsilon)$ határok, amelyek közé t nagy, $1-\varepsilon$ valószínűséggel esik. Ebből kiszámítható $(m_1 - m_2)$-re az ugyancsak $1-\varepsilon$-szintű konfidencia-intervallum.

g) Konfidenciasáv az ismeretlen F(x)-eloszlásfüggvényre.

Az $\{X<x\}$ esemény relatív gyakorisága az $F_n(x)$ empirikus eloszlásfüggvény, valószínűsége az $F(x)$ elméleti eloszlásfüggvény.

A binomiális eloszlás p paraméterének intervallumbecslésénél mondottak adott x-nél $F(x)$-re konfidencia-intervallumot adnak.

$F(x)$-re az $F_n(x)$ körül az alábbi konfidenciasáv adható meg:

$$F_n(x)-\frac{1}{\sqrt{n}} \leqq F(x) \leqq F_n(x)+\frac{1}{\sqrt{n}}.$$

A 11. gyakorló feladatban fogjuk igazolni, hogy ez $F_n(x)$-re szimmetrikus, legalább 75% biztonsági szintű sáv.

h) Ha a normál eloszlás várható értékére adott megbízhatósági intervallum fél szélességét előre megszabjuk (pl. szabványban), akkor megadhatjuk, hogy *adott biztonsági szint esetén a követelmény betartásához hány mérést kell végeznünk.*

– Ha pl. a fél intervallumhossz legfeljebb q, a szórás σ, akkor 95% biztonsági szint esetén

$$q \geqq u_{95}\frac{\sigma}{\sqrt{n}},$$

$$n \geqq \left(1{,}96\frac{\sigma}{q}\right)^2.$$

– Ha csak a minta szórása ismert, akkor a hasonlóan kapható

$$\frac{q}{s^*} \geqq \frac{t}{\sqrt{n}}$$

képletből megállapítható (adott biztonsági szint esetén) a szükséges minimális n. Ezeket az értékeket a 10. táblázat tartalmazza. Pl. $q=0{,}50$, $s^*=2$, 95% biztonsági szintnél $\frac{q}{s^*}=\left(\frac{t}{\sqrt{n}}\right)=0{,}25$, a 10. táblázatból $f=n-1\approx 69$, $n\approx 70$.

$f=n-1$	95% $t/\sqrt{n}$
60	0,238
70	0,259

Gyakorló feladatok

1. *a)* A momentumok módszerével, *b)* a maximum-likelihood módszerrel adjunk becslést az exponenciális eloszlás λ paraméterére!

---

a) Ismerjük tehát az eloszlásfüggvényt és sűrűségfüggvényt:

$$F(x)=1-e^{-\lambda x}, \quad f(x)=\lambda e^{-\lambda x}.$$

Becsüljük az ismeretlen paramétert úgy, hogy az első, elméleti momentumot a mintabeli $x_1, x_2, \ldots, x_n$ értékekből kiszámítható első, tapasztalati momentummal becsüljük:

$$\frac{1}{\lambda}=M_1(X)=\int_0^\infty xf(x)\,dx \approx \sum_{i=1}^n x_i\cdot\frac{1}{n}=\frac{\Sigma x_i}{n}=\bar{x},$$

tehát

$$\hat{\lambda}=\frac{1}{\bar{x}}.$$

b) Exponenciális eloszlás esetén:

$$f(x_i;\lambda)=\lambda e^{-\lambda x_i}$$

így az $L(x_1, x_2, \ldots, x_n; \lambda)=\prod_{i=1}^n f(x_i;\lambda)$ likelihood-függvény:

$$L(x_1, x_2, \ldots, x_n; \lambda)=\prod_{i=1}^n \lambda e^{-\lambda x_i}=\lambda^n\prod_{i=1}^n e^{-\lambda x_i}.$$

Szorzatfüggvény helyett egyszerűbb összegfüggvényt deriválni, keressük tehát a fenti függvény logaritmusának a maximumát:

$$\ln L = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i,$$

$$\frac{d \ln L}{d\lambda} = \frac{n}{\lambda} - \Sigma x_i = 0,$$

$$\hat{\lambda} = \frac{n}{\Sigma x_i} = \frac{1}{\bar{x}}.$$

$\dfrac{d^2 \ln L}{d\lambda^2} = -\dfrac{n}{\lambda^2} < 0$, tehát $\hat{\lambda} = \dfrac{1}{\bar{x}}$ a „legjobb valószínűséget adó” becslés.

A két módszer ugyanarra az eredményre vezetett.

2. A momentumok, ill. a maximális valószínűség módszerével becsüljük meg a normál eloszlású változó m várható értékét és σ szórását az n-elemű, független minta alapján.

a) $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$,

$$m = M_1(x) = \int_{-\infty}^{\infty} x f(x)\, dx \approx \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x}, \qquad \text{tehát} \qquad \hat{m} = \bar{x},$$

$$m^2 + \sigma^2 = M_2(x) = \int_{-\infty}^{\infty} x^2 f(x)\, dx \approx \sum_{i=1}^{n} \frac{x_i^2}{n},$$

és ebből

$$\hat{\sigma}^2 = \frac{1}{n} \Sigma x_i^2 - \bar{x}^2 = \frac{1}{n}(\Sigma x_i^2 - 2n\bar{x}^2 + n\bar{x}^2) = \frac{1}{n}(\Sigma x_i^2 - 2\bar{x}\Sigma x_i + \Sigma \bar{x}^2) =$$

$$= \frac{1}{n} \Sigma (x_i - \bar{x})^2 = s_n^2.$$

I. megoldás

b) A maximum-likelihood módszernél tulajdonképpen a mintaelemek együttes sűrűségfüggvényéből indulunk ki, amely a függetlenség miatt az egyes sűrűségfüggvények szorzata. A sűrűségfüggvények szorzásakor az e kitevői öszeadódnak, így:

$$L(x_1, x_2, \ldots, x_n;\, m, \sigma) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - m)^2}{\sigma^2}}$$

legyen ennek logaritmusa az l likelihood-függvény:

$$l = \ln L = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \Sigma (x_i - m)^2,$$

$$\frac{\partial l}{\partial m} = \frac{\partial \ln L}{\partial m} = -\frac{1}{2\sigma^2}(-2)\Sigma(x_i - m) = 0,$$

$$\Sigma x_i - nm = 0,$$

$$\hat{m} = \frac{\Sigma x_i}{n} = \bar{x},$$

$$\frac{\partial l}{\partial \sigma} = \frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \Sigma (x_i - \bar{x})^2 = 0,$$

$$\Sigma (x_i - \bar{x})^2 = n\sigma^2,$$

$$\hat{\sigma}^2 = \Sigma \frac{(x_i - \bar{x})^2}{n} = s_n^2.$$

A likelihood-függvény kétváltozós. A Hesse-determinánshoz szükséges másodrendű parciális deriváltak (az $m = \bar{x}$ és $\sigma = s_n$ helyen):

$$l''_{mm} = \frac{1}{\sigma^2}(-n) = -\frac{n}{\sigma^2}, \qquad l''_{mm}(\bar{x}, s_n) = -\frac{n}{s_n^2},$$

$$l''_{m\sigma} = l''_{\sigma m} = -\frac{2}{\sigma^3} \Sigma (x_i - m), \qquad l''_{m\sigma}(\bar{x}, s_n) = -\frac{2}{s_n^3} \Sigma (x_i - \bar{x}) = 0,$$

$$l''_{\sigma\sigma} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \Sigma (x_i - m)^2, \qquad l''_{\sigma\sigma}(\bar{x}, s_n) = \frac{n}{s_n^2} - \frac{3}{s_n^4} \cdot n \frac{\Sigma (x_i - \bar{x})^2}{n} =$$

$$= \frac{n}{s_n^2} - \frac{3n}{s_n^4} s_n^2 = -\frac{2n}{s_n^2},$$

$$H = \begin{vmatrix} -\dfrac{n}{s_n^2} & 0 \\ 0 & -\dfrac{2n}{s_n^2} \end{vmatrix} = \frac{2n^2}{s_n^4} > 0,$$

mivel $-\dfrac{n}{s_n^2} < 0$, így biztosan maximum van.

II. megoldás

Elemi úton is megmutatjuk, hogy az $m=\bar{x}$, $\sigma=s_n$ helyen az eredeti L függvénynek (és akkor vele együtt l = ln L-nek) *abszolút maximuma van.*

Az L értéke az $m=\bar{x}$, $\sigma=s_n$ helyen:

$$L(\bar{x}, s_n) = \frac{1}{s_n^n(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2s_n^2}\cdot n\Sigma(x_i-\bar{x})^2/n} = \frac{1}{s_n^n(2\pi)^{\frac{n}{2}}} e^{-\frac{n}{2}}.$$

a) Először azt igazoljuk, hogy ez nem kisebb, mint

$$\frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\bar{x})^2}.$$

Ehhez meg kell mutatnunk, hogy

$$\frac{1}{s_n^n} e^{-\frac{n}{2}} \geqq \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}ns_n^2},$$

$$\left(\frac{\sigma}{s_n}\right)^n \geqq e^{-\frac{n}{2}\left(\frac{s_n}{\sigma}\right)^2+\frac{n}{2}},$$

$$n\ln\frac{\sigma}{s_n} \geqq -\frac{n}{2}\left(\left(\frac{s_n}{\sigma}\right)^2-1\right),$$

azaz $\frac{\sigma}{s_n} = z$ jelöléssel ($z>0$):

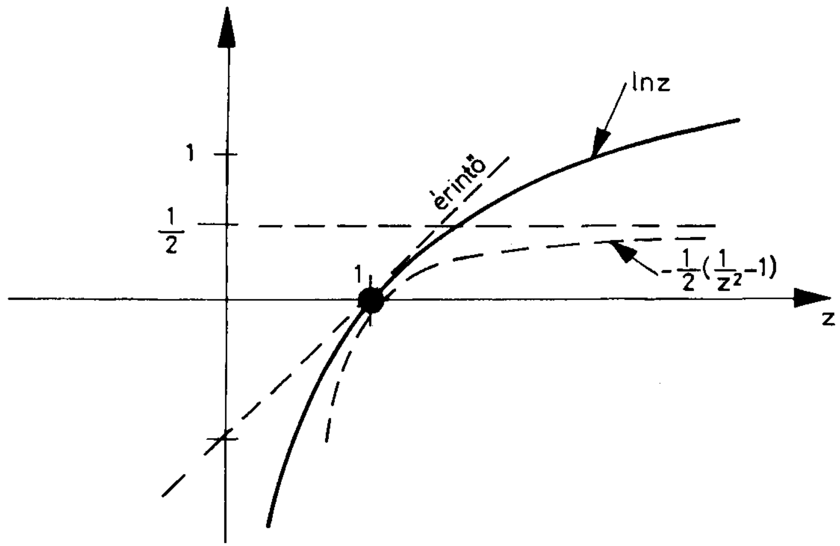
$$\ln z \geqq -\frac{1}{2}\left(\frac{1}{z^2}-1\right).$$

Ez azonban igaz, mert

– az $\ln z$ és $-\frac{1}{2}\left(\frac{1}{z^2}-1\right)$ függvénynek $z=1$-nél közös pontja és közös érintője van (a függvényérték 0, a derivált értéke 1);

– mindkettő alulról konkáv függvény.

Ezért $\ln z$ az (1; 0) közös pont kivételével a $-\frac{1}{2}\left(\frac{1}{z^2}-1\right)$ függvény *alatt*



19. ábra

helyezkedik el (19. ábra). Mivel a lépések visszafelé megfordíthatóak, ezzel valóban igazoltuk, hogy

$$\frac{1}{s_n^n(2\pi)^{\frac{n}{2}}} e^{-\frac{n}{2}} \geqq \frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\bar{x})^2}.$$

b) Másodszor igazoljuk, hogy a most bizonyított egyenlőtlenség jobb oldalán álló kifejezés nem kisebb az L függvény tetszőleges (m, σ) helyen felvett értékénél! Ehhez azt kell megmutatni, hogy

$$e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\bar{x})^2} \geqq e^{-\frac{1}{2\sigma^2}\Sigma(x_i-m)^2},$$

$$-\Sigma(x_i-\bar{x})^2 \geqq -\Sigma(x_i-m)^2,$$

$$\Sigma x_i^2 - 2\bar{x}\Sigma x_i + n\bar{x}^2 \leqq \Sigma x_i^2 - 2m\Sigma x_i + nm^2,$$

$$-2\bar{x}n\bar{x} + n\bar{x}^2 \leqq -2mn\bar{x} + nm^2,$$

$$0 \leqq n(\bar{x}-m)^2.$$

Ez utóbbi a változók minden értékére teljesülő egyenlőtlenség, a lépések megfordíthatósága miatt a *b)* pont elején felírt egyenlőtlenség is helyes.

Akkor pedig az *a)* és *b)* eredményekből következik, hogy az L függvény értéke az $(\bar{x}, s_n)$ pontban nem kisebb, mint tetszőleges (m, σ) pontban. *Az* $(\bar{x}, s_n)$ *pontban az L függvénynek valóban abszolút maximuma van.*

3. Egy szállítmányból N-szer veszünk n-elemű visszatevéses mintát, a selejtesek aránya az N darab mintában rendre:

$$X_1 = \frac{k_1}{n},\ X_2 = \frac{k_2}{n},\ \ldots,\ X_n = \frac{k_n}{n}.$$

Adjuk meg a p selejtvalószínűség maximum-likelihood becslését!

Annak a valószínűsége, hogy az i-edik mintában a selejtszám $\frac{k_i}{n}$ lesz,

$$\binom{n}{k_i} p^{k_i}(1-p)^{n-k_i}.$$

Ennek alapján a likelihood-függvény:

$$L = \prod_{i=1}^{N} \binom{n}{k_i} p^{k_i}(1-p)^{n-k_i},$$

$$\ln L = \sum_{i=1}^{N} \left(\ln \binom{n}{k_i} + k_i \ln p + (n-k_i) \ln (1-p) \right).$$

A likelihood-egyenlet:

$$\frac{\partial \ln L}{\partial p} = \sum_{i=1}^{N} \left(\frac{k_i}{p} - \frac{n-k_i}{1-p} \right) = \Sigma \frac{k_i - np}{p(1-p)} = 0,$$

$$\Sigma k_i - Nnp = 0,$$

$$\hat{p} = \frac{\sum_{i=1}^{N} k_i}{Nn}.$$

Ez a p selejtvalószínűség maximum-likelihood becslése. Mivel a binomiális eloszlásnál np a várható érték, így azt kaptuk, hogy az M várható érték maximum-likelihood becslése

$$\hat{M} = n\hat{p} = \frac{\sum_{i=1}^{n} k_i}{N},$$

vagyis az elméleti várható értéknek az adott mintát maximális valószínűséggel biztosító becslése a tapasztalati „várható érték" (mintaátlag).

A többi számolást az Olvasóra bízzuk.

Megjegyzés

a) $N=1$ db, n-elemű, *visszatevéssel vett* minta esetén az előbbiből az következik, hogy ilyenkor *a p paraméter (selejtvalószínűség) maximum-likelihood becslése a relatív gyakoriság (selejtarány):*

$$\hat{p} = \frac{k_1}{n}.$$

Ez az eredmény a

$$L(k_1, p) = P(X_1 = k_1) = \binom{n}{k_1} p^{k_1}(1-p)^{n-k_1}$$

deriváltjának zérushelyéből is megkapható.

b) Ha az egyszerű n-elemű mintavétel visszatevés nélkül történik, akkor a maximum-likelihood becslés:

$$\hat{p} = \frac{k_1}{n} \cdot \frac{Z+1}{Z},$$

ahol Z az összes munkadarabok számát jelenti, amelyből az n-elemű mintát kivesszük. Ha Z elég nagy, ez keveset tér el a binomiális eloszlásnál kapott $\frac{k_1}{n}$ becsléstől.

4. Egyenletes eloszlású sokaságból kaptuk az $X_i = x_i$ $(i = 1, 2, \ldots, n)$ mintát. Adjuk meg az (a, b) intervallumhoz a kezdő- és végpont abszcissza maximum-likelihood becslését! Itt (a, b) az az intervallum, amelyben egyenletes eloszlású az X változó.

Mivel az $f(x;\ a, b)$ sűrűségfüggvény csak az (a, b) intervallumon különbözik 0-tól, ezért elég csak erre az intervallumra szorítkozni.

$$L(x;\ a, b) = \prod_{i=1}^{n} f(x;\ a, b) = \left(\frac{1}{b-a}\right)^n,$$

$$\ln L = -n \ln (b-a),$$

$$\frac{\partial \ln L}{\partial a} = \frac{n}{b-a} = 0,$$

$$\frac{\partial \ln L}{\partial b} = -\frac{n}{b-a} = 0,$$

az egyenletrendszer ellentmondó, nincs megoldása.

Az egyenletes eloszlásnál a kezdő- és végpont abszcisszájára tehát nincs maximum-likelihood becslés. Kissé pongyolán: semmilyen minta nem határozza meg elég jól a-t és b-t.

5. Határozzuk meg kétparaméteres gammaeloszlás esetén az a és p paraméterek becslését *a)* a momentum-, *b)* a maximum-likelihood módszerrel! A sűrűségfüggvény:

$$f(x) = \begin{cases} \dfrac{a^p}{\Gamma(p)} x^{p-1} e^{-ax}, & \text{ha} \quad x \geqq 0, \\ 0, & \text{ha} \quad x < 0. \end{cases}$$

---

a) Az első és második elméleti momentumot közelítsük a megfelelő tapasztalati momentummal:

$$M_1(X) = \frac{a^p}{\Gamma(p)} \int_0^\infty x^p e^{-ax}\, dx = \frac{1}{a\Gamma(p)} \int_0^\infty (ax)^p e^{-ax}\, d(ax) =$$

$$= \frac{1}{a} \frac{\Gamma(p+1)}{\Gamma(p)} = \frac{p}{a} \approx \bar{X},$$

$$M_2(X) = \frac{a^p}{\Gamma(p)} \int_0^\infty x^{p+1} e^{-ax}\, dx = \frac{1}{a^2\Gamma(p)} \int_0^\infty (ax)^{p+1} e^{-ax}\, d(ax) =$$

$$= \frac{1}{a^2} \cdot \frac{\Gamma(p+2)}{\Gamma(p)} = \frac{1}{a^2}(p+1)p,$$

mivel $M_2(X) - M_1^2(X) = D^2(X)$, ezért:

$$\underline{\frac{1}{a^2}(p^2+p) - \frac{p^2}{a^2} \approx S_n^2(X).}$$

Az aláhúzott két egyenletből kiszámítható a és p becslése:

$$\hat{a} = \frac{\bar{X}}{S_n^2}, \quad \hat{p} = \hat{a}^2 S_n^2 = \frac{\bar{X}^2}{S_n^2}.$$

A kétparaméteres gammaeloszlás a és p paramétere a mintaátlag és tapasztalati szórásnégyzet segítségével becsülhetők.

b) $f(x;\ a, p) = \dfrac{a^p}{\Gamma(p)} x^{p-1} e^{-ax}$, ha $x \geqq 0$, ezért

$$L(a, p) = \prod_{i=1}^{n} f(x_i;\ a, p) = \frac{a^{np}}{\Gamma(p)^n} \prod x_i^{p-1} e^{-ax_i},$$

$$\ln L(a, p) = np \cdot \ln a - n \cdot \ln \Gamma(p) + \sum_{i=1}^{n} ((p-1) \ln x_i - ax_i),$$

$$\frac{\partial \ln L}{\partial p} = n \cdot \ln a - n \cdot \frac{\partial \ln \Gamma(p)}{\partial p} + \Sigma \ln x_i = 0,$$

$$\frac{\partial \ln L}{\partial a} = \frac{np}{a} - \Sigma x_i = 0.$$

A problémát a $\dfrac{\partial \ln \Gamma(p)}{\partial p}$ kifejezés okozza. Numerikus módszerek segítségével az alábbi közelítő megoldás adható meg:

$$\frac{\partial \ln \Gamma(p)}{\partial p} \approx \ln p - \frac{1}{2p} - \frac{1}{12p^2}.$$

Ekkor a $\frac{\partial \ln L}{\partial a} = 0$ egyenletből ezt kapjuk:

$$a = \frac{p}{\bar{X}},$$

ezt a $\frac{\partial \ln L}{\partial p} = 0$ egyenletbe helyettesítve p-re egy másodfokú egyenletet kapunk:

$$\ln p - \ln \bar{X} - \ln p + \frac{1}{2p} + \frac{1}{12\,p^2} + \frac{1}{n}\Sigma \ln x_i = 0,$$

$$12(\ln \bar{X} - \frac{1}{n}\Sigma \ln x_i)p^2 - 6p - 1 = 0.$$

Ebből p kiszámítható (mindig lesz valós gyök és csak az egyik gyök érvényes). Ha így p-re becslést kaptunk, az $a = \frac{p}{\bar{X}}$ egyenlet szolgáltatja a becslését.

A nem részletezett számítások pótlását az Olvasóra bízzuk.

5. $Z = 10\,000$ darab csavarból $n = 20$-elemű mintát veszünk, $N = 5$ alkalommal. A talált selejtesek száma rendre 0, 1, 2, 1, 3. Becsüljük meg a p selejtvalószínűséget, ha *a)* a mintát visszatevéssel vettük és mind az 5 esetet figyelembe vesszük, *b)* a visszatevéses mintákból csak a legrosszabbat (a 3 selejteset) vesszük figyelembe, *c)* a mintát visszatevés nélkül választjuk!

A 2.1. szakasz 3. gyakorló feladatában szereplő becslések segítésével oldhatjuk meg a feladatot.

$$a)\ \hat{p}_1 = \frac{\sum_{i=1}^{N} k_i}{Nn} = \frac{0+1+2+1+3}{5 \cdot 20} = 0{,}07,$$

$$b)\ \hat{p}_2 = \frac{k}{n} = \frac{3}{20} = 0{,}15,$$

$$c)\ \hat{p}_3 = \frac{k}{n} \cdot \frac{z+1}{z} = \frac{3}{20} \cdot \frac{10\,001}{10\,000} = 0{,}150\,015.$$

Legjobb természetesen az első becslés (hiszen ott használtuk fel az egész sokaságról a legtöbb információt). A visszatevés nélküli becslés *c)* alatt alig különbözik a visszatevésestől.

Az *a)* alatti eredmény nagyon szemléletes, egyszerűen azt mondja ki, hogy ha 100-elemű, visszatevéses mintában 7-selejtes csavart találunk, akkor legcélszerűbb a selejtvalószínűséget 7%-nak becsülni.

6. $n = 28$ évre adottak egy folyón az X_i maximális vízhozamok ($m^3 \cdot s^{-1}$-ben).

6597,	9 231,	4813,	8551,	6229,	4757,	4898,
5153,	10 165,	7645,	6654,	4870,	5719,	4360,
4870,	3 029,	3851,	4983,	4672,	4474,	3794,
8155,	8 126,	6456,	4898,	5719,	3437,	7399.

Előzetes tapasztalatok szerint a maximális vízhozam eloszlása jól közelíthető kétparaméteres gammaeloszlással. Becsüljük meg maximum-likelihood módszerrel az a és p paramétereket, írjuk fel a sűrűségfüggvényt!

$$\bar{X} = \frac{\Sigma X_i}{n} = 5839{,}46,$$

$$\ln \bar{X} = 8{,}6723,$$

$$\frac{1}{n}\Sigma \ln X_i = 8{,}62\,694,$$

a p-re megoldandó másodfokú egyenlet:

$$0{,}54\,432p^2 - 6p - 1 = 0,$$

$$p = \frac{6 \pm \sqrt{36 + 2{,}17\,728}}{1{,}08\,864},$$

$$\hat{p} = 11{,}187\,148,$$

(csak a pozitív gyök használható).

$$\hat{a} = 1{,}9158 \cdot 10^{-3}.$$

7. Igazoljuk, hogy az $\bar{x}$ mintaközép az m várható értékre *a)* torzítatlan, a $\sum_{i=1}^{n} p_i x_i$ becslések közt legjobb hatásfokú $\left(\sum_{i=1}^{n} p_i = 1\right)$, *b)* az összes becslések közt nem a legjobb hatásfokú, *c)* erősen konzisztens, *d)* normál eloszlás esetén elégséges becslés.

---

a) A becslés akkor torzítatlan, ha várható értéke éppen a becsült paraméter. Ez teljesül a $\Sigma p_i x_i$ lineáris becslésekre, mert

$$M(\Sigma p_i x_i) = \Sigma p_i M(x_i) = \Sigma p_i m = m\,\Sigma p_i = m.$$

Ez $p_1 = p_2 \ldots = p_m = \frac{1}{n}$ esetén azt jelenti, hogy $\bar{x} = \frac{1}{n}\Sigma x_i$ is torzítatlan becslése m-nek.

Az adott osztályon belül a torzítatlanok közt akkor a legjobb hatásfokú becslés $\bar{x}$, ha szórásnégyzete (minden m esetén) kisebb, mint a $\Sigma x_i p_i$ becslések szórásnégyzete.

Ennek igazolására vezessük be az egyenletes, átlagos súlyozást jelentő $\frac{1}{n}$-től való e_i eltérést minden egyes p_i esetén:

$$p_i = \frac{1}{n} - e_i \quad (i = 1, 2, \ldots, n).$$

Ekkor

$$\Sigma p_i = n \cdot \frac{1}{n} - \Sigma e_i = 1$$

alapján $\sum_{i=1}^{n} e_i = 0$. A becslés szórásnégyzete:

$$D^2(\Sigma p_i x_i) = \Sigma p_i^2 D^2(x_i) = D^2(x)\,\Sigma\left(\frac{1}{n^2} - \frac{2}{n}e_i + e_i^2\right) =$$

$$= D^2(x)\left(n \cdot \frac{1}{n^2} - \frac{2}{n} \cdot \Sigma e_i + \Sigma e_i^2\right) =$$

$$= D^2(x)\left(\frac{1}{n} + \Sigma e_i^2\right) \geqq \frac{D^2(x)}{n} = D^2(\bar{x}).$$

A legkisebb szórású tehát a $\Sigma p_i x_i$ lineáris becslések között a $\Sigma \frac{x_i}{n} = \bar{x}$ becslés, amikor minden $e_i = 0$, minden $p_i = \frac{1}{n}$.

Megjegyzés

Eredményünk a

$$(\Sigma a_i b_i)^2 \leqq \Sigma a_i^2\,\Sigma b_i^2$$

Cauchy–Schwarcz-egyenlőtlenség segítségével is igazolható.

b) Az összes lehetséges becslések közt már nem a számtani közép a legjobb hatásfokú becslés a várható értékre. Bizonyítható ugyanis, hogy ha X egyenletes eloszlású az (a, b) intervallumban, és eloszlásfüggvénye:

$$P(X<x) = \begin{cases} 0, & \text{ha} \quad x<a \\ \dfrac{x-a}{b-a}, & \text{ha} \quad a \leqq x < b, \\ 1, & \text{ha} \quad x \geqq b, \end{cases}$$

akkor a rendezett minta legkisebb és legnagyobb elemének számtani közepe egyrészt torzítatlan becslés $m = \frac{a+b}{2}$-re:

$$M\left(\frac{X_1^* + X_2^*}{2}\right) = \frac{a+b}{2},$$

másrészt szórásnégyzete:

$$D^2\left(\frac{X_1^* + X_2^*}{2}\right) = \frac{(b-a)^2}{2(n+1)(n+2)}$$

kisebb, mint $\bar{X}$ szórásnégyzete (ha $n>2$):

$$D^2(\bar{X}) = \frac{(b-a)^2}{12n}.$$

A két szórásnégyzet aránya:

$$\frac{12n}{2(n+1)(n+2)}$$

monoton csökkenően tart 0-hoz, ha $n \to \infty$. A két szórásnégyzet aránya tetszőleges kicsinnyé tehető.

c) $\bar{X}$ akkor erősen konzisztens becslés m-re, ha torzítatlan becslés, és

$$D^2(\bar{X}) \to 0, \quad \text{ha} \quad n \to \infty.$$

A torzítatlanságot már láttuk. A becslés szórásnégyzete:

$$D^2(\bar{X}) = \frac{D^2(X)}{n} \to 0, \quad \text{ha} \quad n \to \infty,$$

tehát $\bar{X}$ az m-re valóban erősen konzisztens (és akkor egyúttal konzisztens) becslés is.

d) Azt, hogy normál eloszlás esetén $\bar{X}$ elégséges becslés m-re, azt a 2.3. szakasz *e)* pontjában már igazoltuk.

8. Igazoljuk, hogy egy esemény p valószínűségére a $\frac{k}{n}$ relatív gyakoriság *a)* torzítatlan, *b)* legkisebb szórású (hatásos) becslés, vagyis a szórásnégyzete Cramer–Rao-egyenlőtlenségben megadott alsó határt (információhatárt) eléri, *c)* elégséges becslés is!

---

a) Igazoltuk, hogy $\bar{X}$ az $M(X)$ várható értékének torzítatlan becslése. Mivel egy A esemény bekövetkezésekor az 1 értéket p valószínűséggel, a 0 értéket $1-p$ valószínűséggel felvevő X_i indikátorváltozók átlaga:

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{k}{n},$$

várható értéke p, ezért $\frac{k}{n}$ valóban torzítatlan becslése p-nek.

De az előbbi tételre való hivatkozás nélkül, közvetlenül is belátható állításunk. Ugyanis az indikátorváltozókat használva:

$$\bar{X} = \frac{k}{n}, \; M(\bar{X}) = \frac{1}{n} \Sigma\, M(X_i) = \frac{1}{n} np = p,$$

tehát $\frac{k}{n}$ várható értéke valóban p.

b) Egy becslés biztosan a legkisebb szórásnégyzetű (leghatékonyabb), ha szórásnégyzete eléri az alsó információs határt. A Cramer–Rao-egyenlőtlenségben most $g(p) = p$ $\left(\text{magának } p\text{-nek a becslése } \frac{k}{n}\right)$, azt kell igazolnunk, hogy a

$$D^2\left(\frac{k}{n}\right) \geqq \frac{[g'(p)]^2}{I} \geqq \frac{1}{I}$$

egyenlőtlenségben most az egyenlőség érvényes.

Ez pedig igaz. Ugyanis a binomiális eloszlású k gyakoriságra:

$$D^2(k) = np(1-p),$$

így

$$D^2\left(\frac{k}{n}\right) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n},$$

az I információmennyiség az f együttes sűrűségfüggvényből megkapható:

$$P(X_i = 1) = p, f(X_1, X_2, \ldots, X_n; p) = p^k(1-p)^{n-k}.$$

(X_i az A esemény karakterisztikus változója, n független kísérletből a p valószínűségű A esemény k-szor, az ellentéte $(n-k)$-szor következett be.)

$$\ln f = k \ln p + (n-k) \ln (1-p),$$

$$\frac{d \ln f}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k-np}{p(1-p)},$$

$$I = M\left[\left(\frac{d \ln f}{dp}\right)^2\right] = \frac{1}{p^2(1-p)^2} M[(k-np)^2] =$$

$$= \frac{1}{p^2(1-p)^2} D^2(k) = \frac{1}{p^2(1-p)^2} np(1-p) = \frac{n}{p(1-p)}.$$

Mivel $D^2\left(\frac{k}{n}\right) = \frac{p(1-p)}{n}$, ezért $D^2\left(\frac{k}{n}\right) = \frac{1}{I}$ valóban teljesül. $\frac{k}{n}$ a leghatékonyabb, legkisebb szórású becslés.

c) A torzítatlan, leghatékonyabb $\frac{k}{n}$ becslés egyúttal elégséges becslés is p-re.

9. Igazoljuk, hogy az S_n^2 *a)* torzított, *b)* aszimptotikusan torzítatlan, *c)* erősen konzisztens becslés az egész sokaság σ^2 szórásnégyzetére (utóbbi akkor áll fenn, ha az $M(X^4)$ negyedik momentum létezik)!

$$a)\ S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n}\Sigma[(X_i-m)^2-(\bar{X}-m)^2] =$$

$$= \frac{1}{n}\Sigma(X_i-m)^2 - \frac{2}{n}(\bar{X}-m)\,\Sigma(X_i-m) + \frac{1}{n}n(\bar{X}-m)^2 =$$

$$= \frac{1}{n}\Sigma(X_i-m)^2 - 2(\bar{X}-m)\,\Sigma\left(\frac{X_i}{n}-\frac{m}{n}\right) + (\bar{X}-m)^2 =$$

$$= \frac{1}{n}\Sigma(X_i-m)^2 - (\bar{X}-m)^2,$$

$$M[(X_i-m)^2] = D^2(X_i) = \sigma^2,$$

$$M[(\bar{X}-m)^2] = D^2(\bar{X}) = \frac{\sigma^2}{n},$$

ezeket felhasználva kapjuk a már ismert eredményt:

$$M(S_n^2) = \frac{1}{n}n\sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Az S_n^2 véletlen változó megfigyelt értékeinek átlaga nem σ^2, hanem $\frac{n-1}{n}\sigma^2$ körül ingadozik. A torzítást a korrigált szórásnégyzet használatával kiküszöbölhetjük:

$$M(S_n^{*2}) = M\left(\frac{n}{n-1}S_n^2\right) = \sigma^2.$$

b) A mostani levezetésből azonnal megfigyelhető, hogy az S_n^2 aszimptotikusan torzítatlan becslés σ^2-re:

$$\lim_{n\to\infty}\frac{n-1}{n}\sigma^2 = \sigma^2,$$

nagy n-ekre tehát S_n^2 is jó becslés σ^2-hez. Ha $n \geqq 20$, $\frac{n-1}{n} = 1-\frac{1}{n} \geqq 0{,}95$.

Megjegyzés

Bizonyítható, hogy a relatív szórás: $\frac{S_n}{\bar{X}}$ torzított becslés $\frac{\sigma}{m}$-re ($m>0$).

Kimutatható, hogy *a σ szórásra általános esetben már nincs torzítatlan becslés.* Normál eloszlás esetén a

$$\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

statisztika torzítatlan becslése a szórásnak.

c) A bizonyítást (folytonos esetre) csak vázoljuk.

$$D^2(S_n^2) = M(S_n^4) - M^2(S_n^2) = \int_{-\infty}^{\infty} S_n^4 g(x)dx - \left(\frac{n-1}{n}\right)^2\sigma^4 \to 0,$$

ha $n\to\infty$. Itt $g(x)$ az S_n sűrűségfüggvénye. Pongyolán szólva, $M(S_n^4)$ az $\left(\frac{n-1}{n}\right)^2\sigma^4$ érték körül jár; ha $n\to\infty$, akkor tetszőleges pontossággal megközelíti.

Kimutatható, hogy az erős konzisztenciához az $M(X^4)$ negyedik momentum létezése szükséges.

10. Igazoljuk, hogy a Poisson-eloszlás λ paraméterére adott $\bar{x}$ becslés *a)* legjobb hatásfokú (szórásnégyzete eléri az információs határt), *b)* elégséges becslés!

---

a) A mintaelemek együttes sűrűségfüggvénye

$$f(X_1, X_2, \dots, X_n; \lambda) = e^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!},$$

$$\ln f = -n\lambda + \sum_{i=1}^{n} (k_i \ln \lambda - \ln k_i!),$$

$$\frac{d \ln f}{d\lambda} = -n + \frac{1}{\lambda} \Sigma k_i = \frac{1}{\lambda} \Sigma(k_i - \lambda),$$

$$M\left[\left(\frac{d \ln f}{d\lambda}\right)^2\right] = \frac{1}{\lambda^2} M(\Sigma(k_i-\lambda)^2) = \frac{1}{\lambda^2} \Sigma D^2(k_i) = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda} = I,$$

$$d^2(\lambda) = D^2\left(\frac{\Sigma k_i}{n}\right) = \frac{1}{n^2} n\lambda = \frac{\lambda}{n} = \frac{1}{I}.$$

A Poisson-eloszlás λ paraméterére adott $\bar{x}$ becslés szórásnégyzete eléri az alsó információs határt, nincs más, jobb hatásfokú becslés.

b) A fentiekből következik, hogy $\bar{x}$ elégséges becslés is λ-ra. Adunk azonban egy közvetlen bizonyítást is.

Ha $x_1, x_2, \dots, x_n$ egy n-elemű minta, akkor $n\bar{x}$ egy $n\lambda$ paraméterű, Poisson-eloszlású véletlen változó,

$$P\left(x_1 = k_1, \dots, x_n = k_n / \bar{x} = \frac{k}{n}\right) =$$

$$= \frac{P(x_1 = k_1, \dots, x_n = k_n; n\bar{x} = k)}{P(n\bar{x} = k)} =$$

$$= \frac{P(x_1 = k_1, \dots, x_n = k_n)}{P(n\bar{x} = k)} =$$

$$= \frac{e^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!}}{\frac{(n\lambda)^k}{k!} e^{-n\lambda}} = \frac{k!}{n^k \prod k_i!},$$

ha $\sum_{i=1}^{n} k_i = k$, ha pedig $\sum_{i=1}^{n} k_i \neq k$, akkor

$$P\left(x_1 = k_1, \dots, x_n = k_n / \bar{x} = \frac{k}{n}\right) = 0.$$

Ez pedig minden $\frac{k}{n}$ esetén független λ-tól. Az $x_1 = k_1, \dots, x_n = k_n$ minta elégséges a feltételes valószínűség (az eloszlás) meghatározásához, λ nem kell hozzá.

Megjegyzés

Kimutatható, hogy exponenciális eloszlás esetén $g(\lambda) = \frac{1}{\lambda}$-ra az $\bar{X}$ *a)* torzítatlan, *b)* legjobb hatásfokú becslés (a Cramer–Rao-határt elérjük vele), ily módon elégséges becslés is.

---

11. Igaz-e, hogy $F_n(x)$ tapasztalati eloszlásfüggvény az $F(x)$-re *a)* torzítatlan, *b)* erősen konzisztens becslés? *c)* Adjunk megbízhatósági intervallumot $F(x)$-re!

---

a) Mivel $F(x)$ az $\{X < x\}$ esemény valószínűsége, $F_n(x)$ pedig a relatív gyakorisága, ezért a fejezet 8. gyakorló feladata alapján $F_n(x)$ az $F(x)$-re torzítatlan, sőt legkisebb szórású (hatásos) becslés, amelynek szórásnégyzete eléri a Cramer–Rao-egyenlőtlenségben az alsó határt, ezért egyúttal elégséges becslés is.

b) Erősen konzisztens (és akkor egyúttal konzisztens) a becslés, ha szórásnégyzete $\to 0$, midőn $n \to \infty$. Mivel a binomiális eloszlásnál a k gyakoriság szórásnégyzete $np(1-p)$, akkor a $\frac{k}{n}$ relatív gyakoriságé $\frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$, így

$$D^2(F_n(x)) = \frac{F(x)(1-F(x))}{n} \to 0,$$

ha $n \to \infty$. Az erős konzisztencia tehát igaz.

c) A binomiális eloszlás paraméterére adott $(\hat{p}_1; \hat{p}_2)$ megbízhatósági intervallum érvényes $F_n(x) = \hat{p}$-re is, jó közelítésben

$$2\Phi(z) - 1 = 1 - \varepsilon$$

valószínűséggel. (A képletben x helyett most z-t kell használni.) Ez azonban kényelmetlen, mert minden x-re más-más $F_n(x) = \hat{p}$ használandó. Egyszerűbb és minden x-re jó konfidencia-intervallumot (tehát $F_n(x)$ köré egyenletesen rajzolandó *konfidenciasávot*) kapunk, ha a Csebisev-egyenlőtlenséget használjuk:

$$P\left(F_n(x) - \frac{1}{\sqrt{n}} \leqq F(x) \leqq F_n(x) + \frac{1}{\sqrt{n}}\right) =$$

$$= P\left(|F_n(x) - F(x)| \leqq \frac{1}{\sqrt{n}}\right) \geqq 1 - \frac{F(x)(1-F(x))}{n \cdot \frac{1}{n}} \geqq 1 - \frac{1}{4} = 75\%,$$

vagyis $F_n(x)$ *alá is, fölé is* $\frac{1}{\sqrt{n}}$*-t mérve az így kapott sáv legalább 75% valószínűséggel tartalmazza* $F(x)$*-et.*

Alkalmaztuk a mértani–számtani közép közti, ismeretes egyenlőtlenséget:

$$\sqrt{F(x)(1-F(x))} \leqq \frac{F(x)+1-F(x)}{2} = \frac{1}{2}.$$

A kapott sáv fél szélessége például $n=50$ esetén $\frac{1}{\sqrt{50}} \approx 0,14$, $n=100$-nál 0,10.

12. Egy szállítmányból visszatevéssel $n=100$-elemű mintát veszünk, $k=5$ selejtest találunk. Adjunk pont- és intervallumbecslést a p selejtvalószínűségre.

A selejtvalószínűségre torzítatlan, hatásos, elégséges becslés:

$$\hat{p} = \frac{k}{n} = 0,05.$$

a) Közel 90% biztonsági szinthez tartozó megbízhatósági intervallumot keresünk az elméleti összefoglalásban mutatott módszerrel:

$$2\Phi(x) - 1 \approx 0,90,$$

$$x \approx 1,65,$$

a

$$(p-0,05)^2 = 1,65^2 \frac{p-p^2}{100}$$

másodfokú egyenlet megoldása

$$\hat{p}_1 \approx 0,024,$$
$$\hat{p}_2 \approx 0,099,$$

a

$$(p-\hat{p}) \leqq x^2 \frac{p(1-p)}{n}$$

másodfokú egyenlőtlenséghez tartozó parabola egyenes állású, megoldásai a $(\hat{p}_1, \hat{p}_2)$ intervallumban levő p értékek, tehát

$$0,024 < p < 0,100$$

konfidencia-intervallumot kaptuk a p selejtvalószínűségre, jó közelítésben 90% biztonsági szinten.

b) Használjuk a 90% biztonsági szinthez tartozó 18*a* ábrát! $\frac{k}{n} = 0,05$ abszcisszánál az $n=100$ görbékhez tartozó két ordináta kb. 0,02 és 0,10, így

$$0,02 < p < 0,10,$$

kb. 90% biztonsággal.

c) Az alábbiakban megadunk egy nagy n-re jó közelítésben érvényes konfidencia-intervallumot.

Ha k binomiális eloszlású változó esetén:

$$P(k < x) \approx F(x)$$

(ahol $F(x)$ az általános normál eloszlás eloszlásfüggvénye), akkor természetesen:

$$P(-x < k < x) \approx F(x) - F(-x).$$

Viszont $\frac{k-np}{\sqrt{np(1-p)}}$ már 0 várható értékű, 1 szórású binomiális eloszlású változó, hiszen a változóból levontuk a várható értéket, np-t, és a különbségüket osztottuk a szórással, $\sqrt{np(1-p)}$-vel. Ezt az eloszlást a 0 várható értékű, 1 szórású (standardizált) normális eloszlás közelíti jól:

$$P\left(-x < \frac{k-np}{\sqrt{np(1-p)}} < x\right) \approx \Phi(x) - \underbrace{\Phi(-x)}_{1-\Phi(x)} = 2\cdot\Phi(x)-1.$$

$$-x\sqrt{np(1-p)} < k-np < x\sqrt{np(1-p)},$$

osztva n-nel, és rendezve:

$$-\frac{k}{n} - x\sqrt{\frac{p(1-p)}{n}} < -p < -\frac{k}{n} + x\sqrt{\frac{p(1-p)}{n}}, \Big/ \cdot(-1)$$

$$\frac{k}{n} + x\sqrt{\frac{p(1-p)}{n}} > p > \frac{k}{n} - x\sqrt{\frac{p(1-p)}{n}}.$$

A belső egyenlőtlenséget úgy alakítottuk át, hogy p-re alsó és felső határt kapjunk. Ez sikerült is, azonban p-t így önmaga segítségével becsüljük. Azonban p-re torzítatlan becslés a mintabeli $\frac{k}{n}$, írjuk ezt a négyzetgyök alatt a helyére. Így

$$P\left(\frac{k}{n} - x\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}} < p < \frac{k}{n} + x\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}\right) \approx 2\cdot\Phi(x)-1.$$

Ebből $2\cdot\Phi(x)-1 = 0{,}90$, $x\approx 1{,}65$, $\frac{k}{n} = 0{,}05$, $n=100$ esetén a

$$0{,}05 - 1{,}65\sqrt{\frac{0{,}05\cdot 0{,}95}{100}} \approx 0{,}014,$$

$$0{,}05 + 1{,}65\sqrt{\frac{0{,}05\cdot 0{,}95}{100}} \approx 0{,}086$$

határokkal rendelkező megbízhatósági intervallumot kaptuk, közelítőleg 90% biztonsági szinten.

Megjegyzés

A) Érdekes megfigyelni, hogy a *b)* esetben kapott intervallum nem szimmetrikus $\frac{k}{n} = 0{,}05$-re („felfelé" húzódik el az intervallum széle), a *c)* esetben kapott viszont szimmetrikus.

B) A *c)* esetben kapott eredményt úgy is megkaphatjuk, hogy a *b)* alatti másodfokú egyenlet $\hat{p}_1$ és $\hat{p}_2$ gyökében $\frac{1}{n}$-et elhanyagoljuk, 0-nak vesszük $\frac{1}{\sqrt{n}}$-hez képest.

13. Egy fonal szakítószilárdságának átlaga $n=12$ mérésből $\bar{x} = 3{,}25\cdot 10$ N-nak adódott. Az egész sokaság szórása (előzetes mérésekből kialakított szabványelőírás) nem lehet több $\sigma = 0{,}30\cdot 10$ N-nál.

---

a) 95% biztonsággal milyen megbízhatósági intervallumba esik az egész sokaság várható értéke?

A megbízhatósági intervallum fél szélessége

$$q = 1{,}960\cdot\frac{0{,}30}{\sqrt{12}} = 0{,}17,$$

így az egész sokaság átlaga 95% valószínűséggel

$$3{,}25-0{,}17 = 3{,}08 \quad \text{és} \quad 3{,}42 = 3{,}25+0{,}17$$

közé esik. (Az alsó határ legalább 3,08, a felső határ legfeljebb 3,42, mert a szórás is „legfeljebb" 0,30.)

Az intervallumhatárok pontosak, ha a fonál szakítási szilárdsága normál eloszlású (ami gyakran fennáll), közelítőek, ha nem normál eloszlású.

b) A szabványelőírás a megbízhatósági intervallum fél szélességére legfeljebb $q=0{,}20$-ot enged meg. Ha 95% biztonsággal azt akarjuk, hogy a várható érték legalább 3,05, de legfeljebb 3,45 legyen, ennek elérésére hány mérést kell végeznünk?

Azt akarjuk, hogy

$$q \geqq u \frac{\sigma}{\sqrt{n}},$$

$$0{,}20 \geqq 1{,}96 \frac{0{,}30}{\sqrt{n}}$$

legyen, ez teljesül, ha

$$\sqrt{n} \geqq 1{,}96 \frac{0{,}30}{0{,}20},$$

$$n \geqq 8{,}6.$$

Tehát kerekítve legalább 9 mérést kell végeznünk, hogy 95% biztonsággal q maximálisan 0,20 legyen. (Ezzel tulajdonképpen a mintaátlagtól való eltérést maximáljuk, a „maximális" eltérés $3\sigma = 0{,}6$.)

14. Villanyégők élettartamát normál eloszlásúnak találták, $n = 15$ égő élettartam átlaga $\bar{x} = 1200$ óra, korrigált tapasztalati szórása $s^* = 186$ óra, 99%-os biztonsági szinten milyen megbízhatósági intervallumba esik az egész sokaság várható értéke?

Mivel most csak a minta s^* szórása ismert, ezért az

$$\bar{x} - t \frac{s^*}{\sqrt{n}} \leqq m \leqq \bar{x} + t \frac{s^*}{\sqrt{n}}$$

megbízhatósági intervallum határait kell keresnünk. A 8. táblázatot használva, 99% valószínűséggel az egész sokaság várható értékét az

$$1200 - 2{,}977 \cdot \frac{186}{\sqrt{15}} = 1057 \text{ óra}$$

és

$$1200 + 2{,}977 \cdot \frac{186}{\sqrt{15}} = 1343 \text{ óra}$$

határú intervallum lefedi.

15. Az alábbiakban bemutatjuk egy mintavételes munkanap felvételezését a ruhaiparban pillanatfelvétellel (a felmérő szúrópróbaszerűen bejegyzi, hogy milyen tevékenységet végzett a megfigyelt dolgozó, a végén összesíti a tevékenységek előfordulási számát). A mintából következtetéseket vonnak le az egész munkaidő kihasználtságára.

Tervezzük meg a szükséges felvételek számát, N-t!

Vizsgálni kívánjuk az improduktív tevékenységek számának és az összes tevékenység számának az arányát. Ha ez az arány a mintában $\hat{p}$, a valóságos arány jó közelítésben $\hat{p}$ körül $2 \cdot \Phi(x) - 1$ valószínűséggel

$$d = x \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

fél szélességű, $\hat{p}$-re szimmetrikus intervallumba esik. Legyen d kicsi, mondjuk $\hat{p}$-nek a 4%-a, legyen a megbízhatósági szint nagy, 96%-os:

$$2 \cdot \Phi(x) - 1 = 0{,}96, \quad \text{ebből} \quad x \approx 2.$$

Ekkor

$$0{,}04\hat{p} = 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{N}},$$

ebből a szükséges megfigyelések száma:

$$N = \frac{4(1-\hat{p})}{0{,}04^2 \hat{p}}.$$

$\hat{p}$-t voltaképpen a mintavétel után kapjuk meg. Előzetes adatfelvételből vagy normaelőírásból azonban tudjuk, hogy a nem termelő tevékenységek aránya az összeshez viszonyítva $\hat{p} = 0{,}07$. Ekkor a megfigyelések száma:

$$N = \frac{4 \cdot 0{,}93}{0{,}04^2 \cdot 0{,}07} = 33\,124.$$

A felvételhez 720 személyt állítanak be, így a felvételezéshez szükséges utak száma $= \frac{33\,124}{720} = 46$, mivel 1 nap 3 utat tesz meg minden felvételező, a felvételezéshez szükséges napok száma $\frac{46}{3} = 15{,}3$.

A ruhaipari szakemberek megállapították a termelő és nem termelő tevékenységek listáját. A felvételkészítők bejárták a 33 124 utat, bejegyezték az egyes tevékenységek gyakoriságát. Ha a $\hat{p}$ a 0,07-től jelentősen eltér, elvileg újból kell megállapítani az utak számát. A végeredmény: áttekintést kaptak, hogy 1 üzemen belül és a 4 üzem között milyen a nem termelő tevékenység aránya, a munkaidő kihasználtsága. Fontos teendőket tudtak megállapítani ahhoz, hogy minél kevesebb legyen a termelésből különböző okok miatt kieső idő.

A statisztikai követelmény: az improduktív tevékenységek valóságos részaránya a megfigyelt $\hat{p}$-tól felfelé és lefelé is maximum $0{,}04\hat{p}$-re térhessen el nagy (96%-os) biztonsággal. Ez teljesül, ha a megfigyelések száma elég nagy (N legalább 33 124).

16. Egy műszak 8 órája alatt a pillanatfelvételes munkanapelemzés során az eredmények a következők voltak:

Óra	Produktív tevékenységek számaránya, %
6– 7	55
7– 8	76
8– 9	83
9–10	90
10–11	86
11–12	86
12–13	80
13–14	60

A munkaidő-kihasználtságot előzetesen 60%-ra becsülték. 96% biztonsági szinten hány mérés szükséges, ha azt szeretnénk, hogy a tapasztalati érték az egész munkaidőre vonatkozó értéktől annak maximum 10%-ával térhessen el?

$$N = \frac{4(1-0{,}60)}{0{,}10^2 \cdot 0{,}60} = 266{,}7 \approx 267 \text{ mérés.}$$

Az egyes órákban jelentősen eltért a produktív tevékenységek aránya 60%-tól (60%-nak a 10%-ánál, azaz 6%-kal többel). Célszerű korrigálni. Például a műszak negyedik órájában $N = \dfrac{4(1-0{,}90)}{0{,}10^2 \cdot 0{,}90} = 44{,}4$, azaz 45 mérés is elég.

17. A cipők talpfelerősítési szilárdsága véletlentől függő változó. $n=36$ pár cipő vizsgálatánál a mintában az alábbi táblázat szerinti megoszlást észlelték.

Talpfelerősítési szilárdság osztályok, (10 N/cm)	Az adott osztályokba eső adatok száma, f_i
1,5–2,0	1
2,0–2,5	2
2,5–3,0	3
3,0–3,5	6
3,5–4,0	10
4,0–4,5	6
4,5–5,0	5
5,0–5,5	2
5,5–6,0	1
Összesen:	$36=n$

I. Állapítsuk meg osztályba sorolással a talpfelerősítési szilárdság átlagát és korrigált szórását!

Szilárdságosztályok	Osztály-közepek, x_i	Gyakoriságok, f_i	z_i	$f_i z_i$	$f_i z_i^2$
$(k=0{,}5)$ 1,5–2,0	1,75	1	−4	−4	16
2,0–2,5	2,25	2	−3	−6	18
2,5–3,0	2,75	3	−2	−6	12
3,0–3,5	3,25	6	−1	−6	6
3,5–4,0	3,75 = D	10	0	0	0
4,0–4,5	4,25	6	+1	6	6
4,5–5,0	4,75	5	+2	10	20
5,0–5,5	5,25	2	+3	6	18
5,5–6,0	5,75	1	+4	4	16
Összesen:		$36=n$		4	112

$$\bar{x} \approx D+\frac{k}{n}\sum_{i=1}^{9} f_i z_i = 3{,}75+\frac{0{,}5}{36}\cdot 4 = 3{,}8\cdot 10\ \mathrm{N/cm},$$

$$s^{*2} \approx \frac{k^2}{n-1}\left[\sum_{i=1}^{9} f_i z_i^2-\frac{1}{n}\left(\sum_{i=1}^{9} f_i z_i\right)^2\right]=$$

$$=\frac{0{,}5}{35}\left(112-\frac{1}{36}\cdot 4^2\right)=0{,}799\,236,$$

$$s^* \approx 0{,}894\cdot 10\ \mathrm{N/cm}.$$

Ezek torzítatlan becslések az egész sokaság várható értékére és szórására.

II. Szerkesszük meg a minta ún. tapasztalati eloszlásfüggvényét, $F_n(x)$-et!

$$F(x)=P(X<x),$$

vegyük úgy, mintha minden osztályban az osztályközép volna a felvett szilárdságérték! Így

$$F_n(1{,}75)=P(X<1{,}75)=0,$$

de 0 a valószínűsége annak is, hogy a mintában bármely 1,75-nél kisebb x alatt maradjon a szilárdság. Az $F_n(x)$ eloszlásfüggvény tehát 0, ha $x<1{,}75$.

$$F_n(2{,}25)=P(X<2{,}25)=\frac{1}{36},$$

de ugyancsak $\frac{1}{36}$ valószínűségű az, hogy a mintában bármely 1,75 és 2,25 közti x alatt maradjon a szilárdság. Az $F_n(x)=\frac{1}{36}$, ha $1{,}75<x\leqq 2{,}25$.

$$F_n(2{,}75)=P(X<2{,}75)=\frac{1}{36}+\frac{2}{36}=\frac{3}{36},$$

és $F_n(x)=\frac{3}{36}$, ha $2{,}25<x\leqq 2{,}75$.

Nyilván $F_n(6)=P(X<6)=1$, $F_n(6{,}1)=P(X<6{,}1)=1$, az $F_n(x)$ függvény képe $F_n(x)=1$ félegyenes, ha x nagyobb a legnagyobb mintabeli x_i értéknél (20*a* ábra).

Az $F_n(x)$ függvény segítségével megbecsülhetjük annak a valószínűségét, hogy a szilárdság a szabvány szerinti minimális 2,6 alá essék:

$$P(X<2{,}6)=F_n(2{,}6)=\frac{3}{36}\approx 0{,}0833.$$

A selejtvalószínűség becslése tehát kb. 8%.

Annak a valószínűsége, hogy 3 és 4 közé essék a szilárdság:

$$P(3\leqq X\leqq 4)=F_n(4)-F_n(3)=\frac{22}{36}-\frac{6}{36}=0{,}4444,$$

annak a valószínűsége, hogy 4 fölé essék a szilárdság:

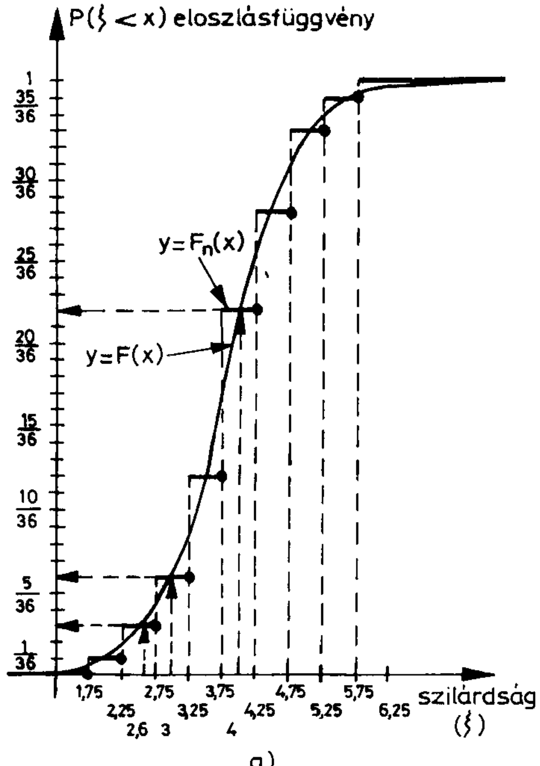
$$P(X\geqq 4)=1-P(X<4)=1-F_n(4)=1-\frac{22}{36}\approx 0{,}3889.$$

A „nagyon jó” (4-nél nem kisebb szilárdságú) cipők várható hányada kb. 39% körül ingadozik.

Ha az osztályok terjedelmét ($k=0{,}5$-et) csökkentjük, a mintaszámot növeljük, sűrűbb ugrású tapasztalati eloszlásfüggvényt kapunk. A már tisztázott értelemben „határesetként” az elméleti $F(x)$ eloszlásfüggvényt kapjuk (közelítő képét l. a 20*a* ábrán).

III. Szerkesszük meg a minta ún. tapasztalati (empirikus) sűrűségfüggvényét, $f_n(x)$-et!

$$f(x) \approx \frac{\text{az intervallumra eső adatok száma}}{(\text{összes adat száma})\cdot(\text{intervallum hossza})},$$



20*a* ábra

így, ha $x<1{,}5$, $f_n(x)=0$,

ha $1{,}5<x<2$, $f_n(x)=\dfrac{1}{36\cdot 0{,}5}=\dfrac{1}{18}$,

ha $2<x<2{,}5$, $f_n(x)=\dfrac{2}{18}$

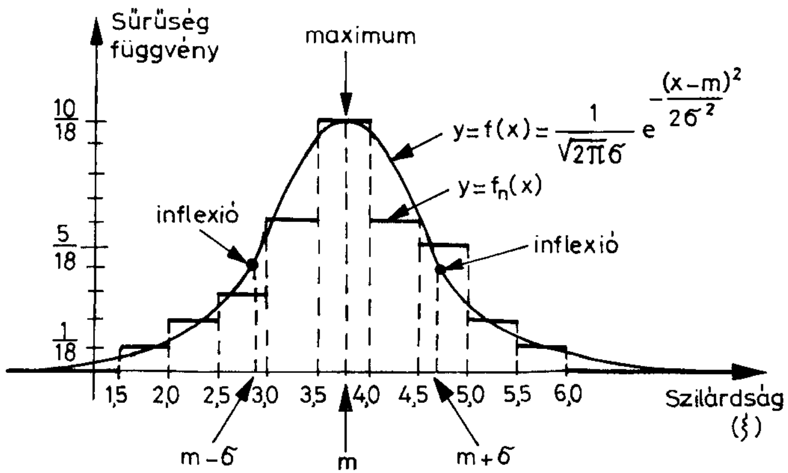
és így tovább, a tapasztalati sűrűségfüggvény egy-egy intervallumon annyi tizennyolcad lesz, ahány adat esett az intervallumra. Nyilván $f_n(x)=0$, ha $x>6$.

Az így megszerkesztett hisztogram a tapasztalati $f_n(x)$ sűrűségfüggvény (20*b* ábra). Ha az intervallumok hossza 0-hoz tart, az adatok mennyiségét minden határon túl növeljük, a már tisztázott értelemben „határesetként” az egész sokaság $f(x)$ sűrűségfüggvényét kapjuk. Ez haranggörbe, *az eloszlás jól közelíthető normálissal.*

IV. A talpfelerősítési szilárdság várható értékeire, m-re torzítatlan becslés $\bar{x}=3{,}8$, szórására torzítatlan becslés $s^*=0{,}894$. Ezek segítségével becsüljük meg annak a valószínűségét, hogy a szilárdság

a) 2,60 alá esik (selejtvalószínűség), *b)* 3 és 4 közé essék, *c)* 4 fölé essék!

$$a)\ P(X<2{,}6)=F(2{,}6)=\Phi\left(\frac{2{,}6-3{,}8}{0{,}894}\right)=\Phi(-1{,}341)=$$
$$=1-\Phi(1{,}341)=1-0{,}9099=0{,}0901.$$



20*b* ábra

b) $P(3 \leqq X < 4) = F(4) - F(3)$,

$$F(4) = \varnothing\left(\frac{4-3{,}8}{0{,}894}\right) = \varnothing(0{,}224) = 0{,}5887,$$

$$F(3) = \varnothing\left(\frac{3-3{,}8}{0{,}894}\right) = \varnothing(-0{,}895) = 1-\varnothing(0{,}895) =$$

$$= 1-0{,}8146 = 0{,}1854.$$

$$P(3 \leqq X < 4) = 0{,}5887-0{,}1854 = 0{,}4043.$$

c) $P(X \geqq 4) = 1-P(X<4) = 1-F(4) = 1-0{,}5887 = 0{,}4113$.

V. A szilárdság sűrűségfüggvénye szimmetrikus, tehát $m-x$ és m, valamint m és $m+x$ közé azonos valószínűséggel esik a változó (hiszen egy intervallumba esés valószínűsége arányos a sűrűségfüggvény alatti területtel).

Számítsuk ki, milyen, a várható értékre szimmetrikus $(m-x;\ m+x)$ intervallumba esnek a szilárdságértékek nagy, 95%-os valószínűséggel!

$$P(m-x \leqq X < m+x) = 0{,}95,$$

$$F(m+x)-F(m-x) = 0{,}95,$$

$$\varnothing\left(\frac{m+x-m}{\sigma}\right) - \varnothing\left(\frac{m-x-m}{\sigma}\right) = 0{,}95,$$

$$\varnothing\left(\frac{x}{\sigma}\right) - \varnothing\left(-\frac{x}{\sigma}\right) = 0{,}95,$$

$$\varnothing\left(\frac{x}{\sigma}\right) - \left[1-\varnothing\left(\frac{x}{\sigma}\right)\right] = 0{,}95,$$

$$2\cdot\varnothing\left(\frac{x}{\sigma}\right) = 1{,}95,$$

$$\varnothing\left(\frac{x}{\sigma}\right) = 0{,}975.$$

Az 5. táblázatból visszakeresve, 1,96-nál lesz az $\varnothing$-függvény értéke 0,975:

x	$\varphi(x)$	$\varnothing(x)$
1,96		0,9750

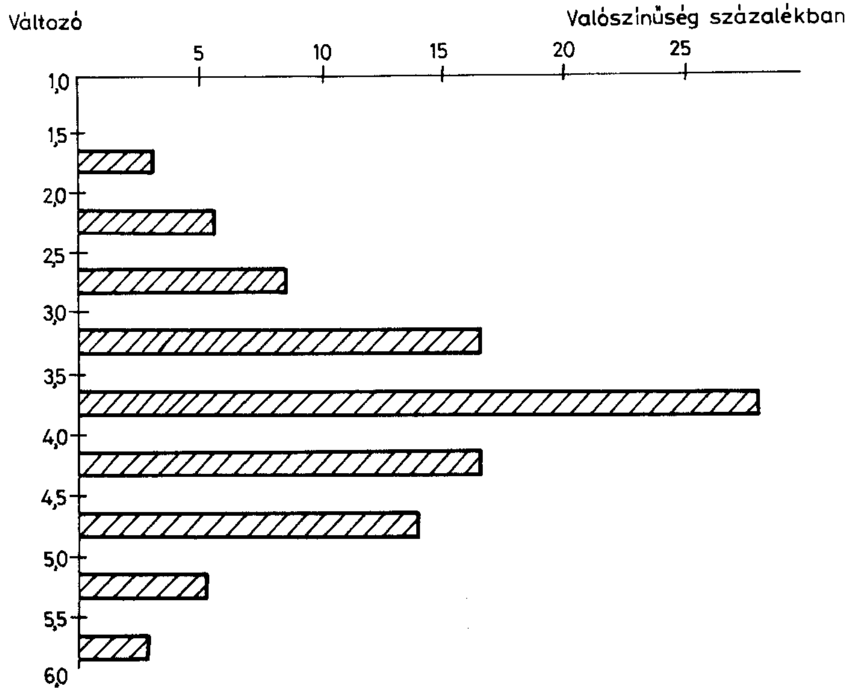
Tehát $\frac{x}{\sigma} = 1{,}96$, $x = 1{,}96\sigma$. A szilárdság értékei 95% valószínűséggel:

$$m-x = m-1{,}96\sigma = 3{,}8-1{,}96\cdot 0{,}894 = 2{,}05,$$

és

$$m+x = m+1{,}96\sigma = 3{,}8+1{,}96\cdot 0{,}894 = 5{,}55$$

közé esnek.



21. ábra

VI. Az adatok osztályokba sorolását, hisztogram készítését, az átlag (MEAN) és a korrigált szórás (STANDARD DEVIATION) kiszámítását végeztük el egy IBM 370/145 gépre írt programcsomag segítségével! Az elkészült eredményt a 21. ábra mutatja. (Az ordinátatengelyen nem $f(x)$ értékei vannak feltüntetve, hanem az intervallumra esés valószínűségei %-ban. Mivel

$$f(x) \approx \frac{\text{intervallumra eső adatok száma}}{(\text{összes adat száma})\cdot(\text{intervallum hossza})} =$$

$$= \frac{\text{intervallumra esés valószínűsége}}{\text{intervallum hossza}},$$

ezért az ordinátaértékek $0{,}5f(x)$ értékét mutatják %-ban.)

VII. A feladatban $\bar{x}=3{,}8$, $s^*=0{,}894$ volt, az $n=36$ elemű minta alapján a szilárdság eloszlását tekintettük normálisnak, a t szorzó 95% biztonsági szinten (interpolációval):

$n=f+1$	Statisztikai biztonság		
	90%	95%	98%
31	1,697	2,042	2,457
41	1,684	2,021	2,423

$$t=2{,}031.$$

$t\frac{s^*}{\sqrt{n}} = 2{,}031\cdot\frac{0{,}894}{\sqrt{36}} = 0{,}3$, ezt $\bar{x}=3{,}8$-ből levonva, ill. 3,8-hez hozzáadva kapjuk, hogy az egész sokaság m várható értéke 3,5 és 4,1 között keresendő 95% valószínűséggel:

$$P(3{,}5<m<4{,}1) = 95\%.$$

18. Az előző feladat alapján adjunk meg konfidenciasávot az elméleti eloszlás $F(x)$ eloszlásfüggvényére!

I. Mint láttuk, legalább 75%-os biztonsági szinten

$$F_n(x)-\frac{1}{\sqrt{36}} \leqq F(x) \leqq F_n(x)+\frac{1}{\sqrt{36}},$$

$$F_n(x)-0{,}17 < F(x) < F_n(x)+0{,}17.$$

A selejtvalószínűségre kapott

$$F_n(2{,}6)\approx 0{,}0833$$

becslés segítségével megbecsülhetjük a selejtvalószínűséget:

$$F_n(2{,}6)-0{,}17 < F(2{,}6) = P(X<2{,}6)<F_n+0{,}17,$$
$$0<F(2{,}6)<0{,}2533$$

legalább 75% valószínűséggel. Ez elég komoly selejtet eredményezhet.

II. Említettük, hogy

$$\bar{x}-t\frac{s^*}{\sqrt{n}} < m < \bar{x}+t\frac{s^*}{\sqrt{n}}$$

(a t értéket a 8. táblázatból az $n-1$ szabadsági fok sorából, és az $1-P_1$ valószínűségi szint oszlopából vesszük), ezenkívül

$$\frac{\sqrt{n}s}{\chi_{\frac{P_2}{2}}} \leqq \sigma \leqq \frac{\sqrt{n}s}{\chi_{1-\frac{P_2}{2}}}$$

$1-P_2$ valószínűséggel. Itt $\chi_{\frac{P_2}{2}}$ a 7. táblázat $\frac{P_2}{2}$ valószínűségi szintjéhez, $n-1$ szabadsági fokhoz tartozó értéke, $\chi_{1-\frac{P_2}{2}}$ pedig ugyanezen szabadsági fokhoz, $1-\frac{P_2}{2}$ valószínűségi szinthez tartozó értéke.

Normál eloszlás esetén:

$$P(X<x) = F(x) = \Phi\left(\frac{x-m}{\sigma}\right).$$

$F(x)$ értékét becsülni tudjuk az m-re és σ-ra adott becslések segítségével.

a) Ha $x < \bar{x} - t\frac{s^*}{\sqrt{n}}$, akkor $x - m < 0$, így

$$F(x) = \Phi\left(\frac{x-m}{\sigma}\right) = 1 - \Phi\left(\frac{m-x}{\sigma}\right).$$

Annak érdekében, hogy $F(x)$-re felső határt kapjunk, $\Phi\left(\frac{m-x}{\sigma}\right)$-ra alsó becslést kell készíteni, és fordítva. Mivel a Φ függvény szigorúan monoton növekedő, így $\frac{m-x}{\sigma}$-ra alsó becslést kell adni $\Phi\left(\frac{m-x}{\sigma}\right)$ alsó becslése érdekében. Az eredmény:

$$1 - \Phi\left(\frac{\bar{x} + t\frac{s^*}{\sqrt{n}} - x}{\frac{s\sqrt{n}}{\chi_{\frac{P_2}{2}}}}\right) < F(x) < 1 - \Phi\left(\frac{\bar{x} - t\frac{s^*}{\sqrt{n}} - x}{\frac{s\sqrt{n}}{\chi_{1-\frac{P_2}{2}}}}\right).$$

Mivel a normál eloszlás szórása és várható értéke független véletlen változók, a fenti megbízhatósági intervallum szintje $(1-P_1)(1-P_2)$.

b) Ugyanígy, ha $x \geqq \bar{x} + t\frac{s^*}{\sqrt{n}}$, akkor $(1-P_1)(1-P_2)$ valószínűséggel:

$$\Phi\left(\frac{x\sqrt{n} - \bar{x}\sqrt{n} - ts^*}{\frac{s\sqrt{n}}{\chi_{1-\frac{P_2}{2}}}}\right) < F(x) < \Phi\left(\frac{x\sqrt{n} - \bar{x}\sqrt{n} + ts^*}{\frac{s\sqrt{n}}{\chi_{\frac{P_2}{2}}}}\right).$$

c) Ha $\bar{x} - t\frac{s^*}{\sqrt{n}} < x < \bar{x} + t\frac{s^*}{\sqrt{n}}$, akkor a t értéket a biztonsági szint csökkentésével annyira kell csökkenteni, hogy x az m-re adott konfidencia intervallumon kívül essék.

Az előbbi módszerrel minden x-hez külön-külön konfidencia-intervallumot szerkesztve, *a normál eloszlás $F(x)$ függvényére egy konfidenciasáv adható meg.*

d) Becsüljük meg a cipőipari feladatban a selejtvalószínűséget (az $\{X<2{,}6\}$ esemény valószínűségét)!

Mint már láttuk, pontbecslésként

$$P(X<2{,}6) = F(2{,}6) \approx 0{,}0901.$$

Az *a)* eredményt használva az intervallumbecslés ugyanerre

$$1 - \Phi\left(\frac{4{,}1-2{,}6}{0{,}752}\right) < P(X<2{,}6) < 1 - \Phi\left(\frac{3{,}5-2{,}6}{1{,}118}\right),$$

$$0{,}0233 < P(X<2{,}6) < 0{,}2105,$$

$0{,}95 \cdot 0{,}90 = 0{,}855$ valószínűséggel. *A selejtvalószínűség tehát rossz esetben 21%-ra is felmehet.*

e) Ugyanilyen módszerrel megbízhatósági intervallum adható a normál eloszlás $f(x)$ sűrűségfüggvényére. Ugyanis

$$f(x) = F'(x) = \frac{d}{dx}\Phi\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sigma}\varphi\left(\frac{x-m}{\sigma}\right).$$

Ha $x \leqq \bar{x} - t\frac{s^*}{\sqrt{n}}$, akkor

$$\frac{\chi_{1-\frac{P_2}{2}}}{s\sqrt{n}} \cdot \varphi\left(\frac{\chi_{\frac{P_2}{2}}(\bar{x}\sqrt{n} + ts^* - x\sqrt{n})}{sn}\right) < f(x) <$$

$$< \frac{\chi_{\frac{P_2}{2}}}{s\sqrt{n}}\varphi\left(\frac{\chi_{1-\frac{P_2}{2}}(\bar{x}\sqrt{n} - ts^* - x\sqrt{n})}{sn}\right).$$

A valószínűségi szint: $(1-P_1)(1-P_2)$.
Hasonló eredmények kaphatók a többi esetekben.

19. *a)* Indokoljuk meg, hogy hogyan jött létre a 10. táblázat; *b)* hasonlítsuk össze a minimális mérések számát, ha az elméleti szórás adott, vagy ha csak a tapasztalati korrigált szórás ismert!

a) A megbízhatósági intervallum fél szélessége (ha a normál eloszlású sokaságból vett n-elemű minta szórása ismert):

$$q = t\frac{s^*}{\sqrt{n}}.$$

Ebből

$$\frac{q}{s^*} = \frac{t}{\sqrt{n}}.$$

Ha 95%-os megbízhatósági szinten dolgozva $q = 0{,}3$, $s^* = 1{,}5$ értékeket engedünk meg, akkor a

$$\frac{t}{\sqrt{n}}\left(= \frac{q}{s^*}\right) = 0{,}20$$

kb. $n = 100$-nál lesz (itt ugyanis $t \approx 2$). Így keletkezett az alábbi táblázatrészlet és a 10. táblázat.

n	$\frac{t}{\sqrt{n}} = \frac{q}{s^*}$ (95% megbízhatósági szinten)
61	0,256
81	0,221
101	0,198
121	0,180
201	0,139

b) Ha $q = 0{,}3$, $\sigma = 1{,}5$, akkor a szükséges minimális mérések száma:

$$n \geqq \left(u_{95}\frac{\sigma}{q}\right)^2 = \left(1{,}96 \cdot \frac{1{,}5}{0{,}3}\right)^2 = 96{,}04.$$

Most minimum 97 mérés szükséges.

Ha viszont csak $q = 0{,}3$, $s^* = 1{,}5$ ismert, akkor minimum 100 mérés szükséges (mint már láttuk). *A mintabeli szórás ismerete kevesebb információ, ugyanazon biztonsági szint esetén több mérés szükséges, mint ha ugyanazon q-nál az elméleti szórás ismert.*

3. STATISZTIKAI HIPOTÉZISEK VIZSGÁLATA

3.1. Bevezetés

Statisztikai hipotézis a megfigyelt valószínűségi változó eloszlásának a típusára vagy az eloszlás paramétereire tett feltevés. Például:

– Egy esemény relatív gyakorisága $\frac{k}{n}$, mit tudunk mondani a valószínűségéről, p-ről (binomiális eloszlás).

– Normál eloszlású változóra vett mintában az átlag $\bar{x}$, igaz-e, hogy az egész sokaság várható értéke $M(X)$ adott m_0-val egyenlő?

– Két sokaságból vett mintában az átlagok: $\bar{x}_1$ és $\bar{x}_2$ milyen közelsége esetén tehetjük fel, hogy a várható értékek egyenlők?

– Két mintánk van két változóra, amelyekből úgy tűnik: a két eloszlásfüggvény megegyezik. Feltehető-e $F_1(x) \equiv F_2(x)$?

– Adott minta alapján feltehető-e, hogy a változó eloszlásfüggvénye adott $F(x)$?

A hipotézisvizsgálat során a feltevést igaznak tételezzük fel, ez a H_0 *nullhipotézis*. Az eloszlásra vagy ismeretlen paraméterre vonatkozó, a nullhipotézissel szembenálló más lehetőséget H_1 *ellenhipotézisnek* nevezzük.

A másodiknak említett problémában korábbi tapasztalatokból ismert az eloszlás típusa (normális), és σ szórása. Fennáll-e a

$$H_0 : M(X) = m_0$$

nullhipotézis (vagyis a minta vagy a mintabeli tapasztalat alátámasztja-e az $M(X) = m_0$ feltevést?) Ekkor a H_0 nullhipotézis *egyértelműen* meghatározza az ismeretlen várható értéket (és így az eloszlást). Ez a fajta hipotézis *egyszerű hipotézis*. Lehet több paraméter is ismeretlen, akkor is lehet egyszerű a hipotézis (pl. normál eloszlásnál

$$H_0 : M(X) = m_0, \qquad D^2(X) = \sigma_0$$

szintén egyértelmű eloszlást definiál).

Ha a H_0 hipotézis teljesülése esetén *több eloszlás* (az eloszlások egyfajta összessége) kerülhet szóba, *összetett a hipotézis*. Például normál eloszlás esetén ismeretlen a σ szórás, a minta alapján igaznak tesszük fel a

$$H_0 : M(X) = m_0$$

hipotézist. Ekkor a megoldás: azonos várható értékű, különböző szórású, normális eloszlások összessége.

A nullhipotézisek fennállását sohasem vizsgálhatjuk 100%-os valószínűséggel (akkor az egész sokaság ismeretére volna szükségünk), mindig csak bizonyos biztonsági szinten döntünk (90%, 95%, 99%, 99,9% stb.).

Az egyszerű és összetett hipotézis definíciója általánosan a következő. Bármely diszkrét vagy folytonos eloszlást egyértelműen meghatároz a

$$P(X < x) = F(x, T)$$

eloszlásfüggvény, ahol a T egyetlen paraméter vagy egy $T = (T_1, T_2, \ldots, T_k)$ paramétervektor.

Legyen a k-dimenziós paramétertér egy résztartománya (részhalmaza) A_0, a nullhipotézis:

$$H_0 : P(X < x) = F(x, T), \qquad T \in A_0.$$

Ekkor H_0 egyszerű hipotézis, ha A_0 a k-dimenziós paramétertér egyetlen pontja, H_0 összetett, ha A_0 több pontot is tartalmaz.

A H_0 hipotézissel szemben vizsgálhatjuk a

$$H_1 : P(X < x) = F(x, T), \qquad T \in A_1$$

alternatívát (ellenhipotézist). Itt A_1 valamilyen ésszerű lehetőséget jelent a paraméterre (nem feltétlenül A_0 komplementerét). Az A_1 hipotézis szintén lehet egyszerű vagy összetett. Természetesen $A_0 A_1 = \emptyset$ (A_0 és A_1 diszjunkt halmazok).

Ha nincs kitüntetett, ésszerű ellenhipotézis, akkor a nullhipotézist minden lehetséges ellenhipotézisre, tehát a

$$H'_1 : P(X < x) = F(x, T), \qquad T \notin A_0$$

alternatívára vizsgáljuk.

Exponenciális eloszlás esetén a paraméter egydimenziós (a számegyenes pozitív fele), és ennek egyetlen $\lambda = \lambda_0$ pontja egyértelműen megadja az eloszlást. Az ilyen feladat *paraméteres probléma*. A paraméterekre vonatkozó hipotézisvizsgálat a *paraméteres próba*.

Gyakran az eloszlásfüggvény típusát sem ismerjük, csak annyit tudunk például, hogy szóba jöhet az összes folytonos eloszlás. *Ez a fajta probléma nemparaméteres,* pl. két valószínűségi változó eloszlása azonos-e. A nemparaméteres problémák eldöntésére szolgáló módszerek közül legismertebbek az *illeszkedésvizsgálati és homogenitásvizsgálati módszerek*. Például adott változóról feltehető-e, hogy normál eloszlású (jól illeszkedik-e a tapasztalati eloszlásfüggvény a feltételezett elméleti eloszlásfüggvényre). Vagy például feltehető-e, hogy két változó eloszlása azonos (homogén-e, egyöntetű-e a két sokaság).

3.2. Statisztikai próbák

Az alábbi próbák jó részében feltételezzük, hogy a változó normál eloszlású.

1. Az egymintás *u*-próba

Vegyünk n-elemű mintát egy $N(m, \sigma)$-eloszlású sokaságból, ahol σ ismert, de m ismeretlen. Az egész sokaság várható értékére szabványelőírás vagy feltételezés, hogy $m = m_0$ adott érték.

A mintabeli $\bar{x}$ nem lesz pontosan m_0, hanem akörül ingadozik. *A mintaátlag mekkora eltérése esetén feltételezhetjük, hogy a várható érték m_0?*

Ha *a*

$$H_0 : m = m_0$$

nullhipotézis teljesül a

$$H_1 : m \neq m_0$$

alternatívával szemben, akkor az

$$u = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}}$$

véletlen változó standard normál eloszlású. Ha ugyanis az $\bar{x}$ véletlen változóból kivonjuk az m_0 várható értékét és osztunk a szórásával, $\sigma/\sqrt{n}$-nel, akkor ezzel a változót standardizáltuk. A mondott lineáris transzformációk nem változtatják meg a normalitást.

Ekkor viszont a $\emptyset(x)$ függvény segítségével kiszámítható, hogy u nagy (pl. $1-\varepsilon = 0,95$) valószínűséggel milyen megbízhatósági intervallumba esik:

$$P\left(-u_\varepsilon \leqq \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} \leqq u_\varepsilon\right) = P\left(\left|\frac{\bar{x} - m_0}{\sigma/\sqrt{n}}\right| \leqq u_\varepsilon\right) = 1 - \varepsilon.$$

Ha tehát a nullhipotézis érvényes, akkor az

$$u_{\text{számított}} = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}}$$

nagy valószínűséggel a $[-u_\varepsilon, u_\varepsilon]$ ún. *elfogadási tartományra* esik, és csak kis (pl. 5%) valószínűséggel esik kívülre, az

$$\{u < -u_\varepsilon, \quad u > u_\varepsilon\}$$

ún. *kritikus tartományra.*

Ezért célszerű, ha így döntünk. *A*

$$H_0 : m = m_0$$

feltevést elfogadjuk $(1-\varepsilon) \cdot 100\%$ *biztonsági szinten, ha*

$$|u_{\text{számított}}| = \left|\sqrt{n}\,\frac{\bar{x} - m_0}{\sigma}\right| \leqq u_\varepsilon = u_{\text{táblázatbeli}};$$

elvetjük, ha

$$|u_{\text{számított}}| > u_{\text{táblázatbeli}}.$$

Ez esetben a H_1 alternatíva mellett döntöttünk. Az u táblázatbeli értéket az 5. vagy a 8. táblázatból kereshetjük. Pl. $1-\varepsilon = 0,95$ esetén $u_\varepsilon = 1,96$:

$n = f+1$	Statisztikai biztonság 95%
∞	1,960

(Lényegében úgy tekinthetjük, mintha σ ismerete miatt lett volna előzetesen végtelen sok mérésünk.)

Döntésünk nem biztos, hiszen az $m = m_0$ fennállását feltételező valószínűség:

$$P\left(-u_\varepsilon \leqq \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} \leqq u_\varepsilon / H_0\right) = 1 - \varepsilon \neq 100\%.$$

Jobb tehát, ha csak annyit mondunk: H_0 *igaz volta esetén a mintabeli tapasztalattal a nullhipotézis nem áll ellentmondásban.* Másként fogalmazva: $\bar{x}$ *eltérése* m_0*-tól nem lényeges, nem szignifikáns az elvetéshez.* Megint másképpen: $\bar{x}$ *eltérése* m_0*-tól nem véletlen.*

Az u-próba lényege tehát: a $H_0 : m = m_0$ hipotézis fennállása esetén ismerjük az

$$u = \sqrt{n}\,\frac{\bar{x} - m_0}{\sigma}$$

eloszlását. Ennek ismeretében olyan $(-u_\varepsilon, u_\varepsilon)$ megbízhatósági intervallumot konstruálunk, amelybe a mintából nyert $u(=u_{\text{számított}})$ nagy, $1-\varepsilon$ valószínűséggel esik. Ha u ezen az intervallumon kívül (a kritikus tartományban) van, a H_0 feltevést elvetjük, ha az intervallumon belül (az elfogadási tartományban), a H_0 feltevést elfogadjuk. A *próba terjedelme* a kritikus tartományra esés valószínűsége (ε).

Minden matematikai statisztikai próba alapelve ugyanaz, csak meg kell találni azt a statisztikai függvényt, amelynek eloszlását H_0 fennállása esetén ismerjük. Ha csak a szóban forgó statisztika határeloszlását ismerjük (midőn $n\to\infty$), akkor megfelelően nagy n esetén alkalmazzuk a próbát (l. később a χ^2-próbát).

Mivel az u-próbánál

$$P(u<-u_\varepsilon) = P(u>u_\varepsilon) = \frac{\varepsilon}{2},$$

ezért ha azt tudjuk, hogy várhatóan érvényes

$$u \geqq -u_\varepsilon,$$

akkor *az előbbi kétoldali próba helyett egyoldali próbát alkalmazunk:* az elfogadási tartomány az $u \geqq -u_\varepsilon$ tartomány lesz $\left(1-\frac{\varepsilon}{2}\right)\cdot 100\%$ biztonsági szinten, a kritikus tartomány az $u<-u_\varepsilon$ tartomány lesz. Egyoldali próba esetén a biztonsági szint növekszik a kétoldalihoz viszonyítva.

2. Első- és másodfajú hiba, a próba erőfüggvénye

Ezeket a fogalmakat az u-próbával kapcsolatosan tisztázzuk. A mondottak viszont *minden próbánál megtárgyalhatók.*

Amikor u beleesik a kritikus tartományba és H_0-t elvetjük, nem biztos, hogy nem tévedünk. Ha H_0 igaz és mégis elvetjük az $M(X)=m_0$ feltevést (mert a számított u érték a kritikus tartományra esett), *elsőfajú hibát* követünk el. Ennek valószínűsége kicsi:

$$P(|u|>u_\varepsilon / H_0) = \varepsilon.$$

Előfordulhat az is, hogy H_0 nem igaz, és az u statisztika értéke mégis az elfogadási tartományba esik. Ilyenkor (hibásan) a $H_0: M(X) = m_0$ hamis feltevést fogadjuk el. Ez a hiba a *másodfajú hiba.*

A lehetséges döntések és hibák az alábbiak:

Nullhipotézis	Döntés – a H_0 hipotézist elfogadjuk	Döntés – a H_0 hipotézist elutasítjuk
Fennáll	helyes döntés	elsőfajú hiba
Nem áll fenn	másodfajú hiba	helyes döntés

A másodfajú hiba esetén a

$$H_1: M(X) = m \neq m_0$$

ellenhipotézis igaz. Az u valószínűségi változó várható értéke:

$$M\left(\frac{\bar{x}-m_0}{\sigma/\sqrt{n}}\right) = \sqrt{n}\,\frac{m-m_0}{\sigma} = d \neq 0$$

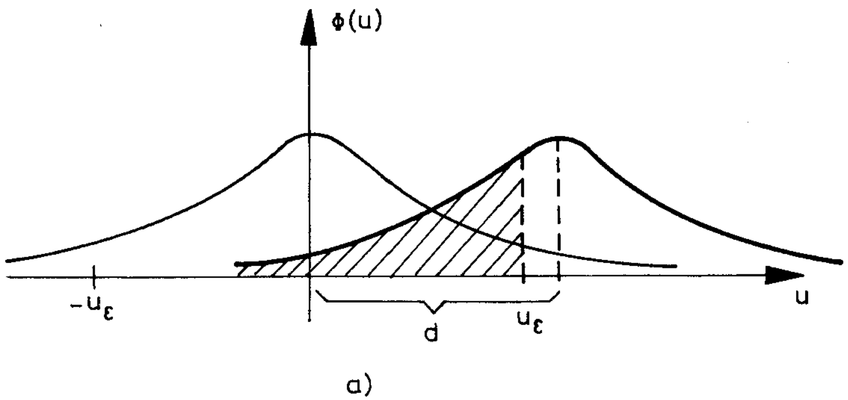
(felhasználtuk, hogy $M(\bar{x})=m$).

A másodfajú hiba elkövetésének β valószínűségét a 22*a* ábrán a vonalkázott terület mutatja. Ez a valószínűség:

$$\beta = P(-u_\varepsilon<u<u_\varepsilon / H_1).$$

Az ábráról látszik, hogy ha d kicsi (m_1 közel esik m_0-hoz), akkor a másodfajú hiba elkövetésének β valószínűsége nagy.

Minél kisebbre választjuk az elsőfajú hiba elkövetésének ε valószínűségét (azaz minél nagyobb a $-u_\varepsilon \leqq u \leqq u_\varepsilon$ intervallum), annál na-



22*a* ábra

gyobb lesz a másodfajú hiba elkövetésének valószínűsége (az ábrán a vonalkázott terület). *Nem célszerű tehát a statisztikai próbáknál a biztonsági szintet túl magasra választani,* ha 99,9%-os szinten fogadunk el vagy vetünk el, túl magas lesz a másodfajú hiba valószínűsége. A próba terjedelmének megállapításakor szokták vizsgálni az első- és másodfajú hiba elkövetése okozta kárt, és ezt is figyelembe venni a biztonsági szint beállításánál.

Ha $M(X) = M(\bar{x}) = m \neq m_0$, akkor *a másodfajú hiba el nem követésének a valószínűsége* (azaz annak a valószínűsége, hogy u a kritikus tartományra esik)

$$E(m) = 1-\beta = P(|u| > u_\varepsilon / H_1).$$

Ez az m értékétől függő mennyiség adott m esetén *a próba ereje, E(m)* pedig *a próba erőfüggvénye* (tehát nem az „expected value”, azaz a várható érték!). Az elnevezés szerencsés: azért a próba „ereje”, mert annál jobb egy próba, minél nagyobb a másodfajú hiba el nem követésének a valószínűsége.

$$E(m_0) = \varepsilon,$$

mivel a kritikus tartományt úgy választjuk meg, hogy ennek valószínűsége $H_0 : M(X) = m_0$ esetén ε (kis érték) legyen.

A másodfajú hiba el nem követésének valószínűsége (mivel most $\sqrt{n}\dfrac{\bar{x}-m}{\sigma}$ lesz standard normál eloszlású):

$$E(m) = P(|u| > u_\varepsilon / H_1) =$$

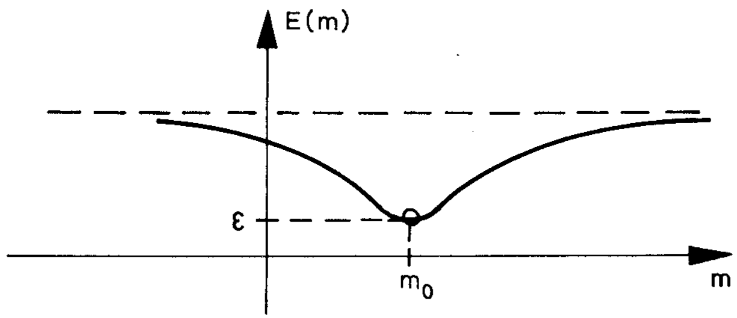
$$= 1-P\left(-u_\varepsilon \leqq \frac{\bar{x}-m_0}{\sigma/\sqrt{n}} \leqq u_\varepsilon / H_1\right) =$$

$$= 1-P\left(-u_\varepsilon \leqq \frac{\bar{x}-m}{\sigma/\sqrt{n}} + \underbrace{\frac{m-m_0}{\sigma/\sqrt{n}}}_{d} \leqq u_\varepsilon / H_1\right) =$$

$$= 1-P\left(-u_\varepsilon - d \leqq \frac{\bar{x}-m}{\sigma/\sqrt{n}} \leqq +u_\varepsilon - d\right) =$$

$$= 1-\Phi(u_\varepsilon - d) + \Phi(-u_\varepsilon - d).$$

Ha $m \to +\infty$, akkor $d = \dfrac{m-m_0}{\sigma/\sqrt{n}} \to +\infty$, $u_\varepsilon - d$ is, $-u_\varepsilon - d$ is mínusz végtelenhez tart, $\Phi(-\infty) = 0$ miatt $E(m) \to 1$. Ugyanígy $m \to -\infty$ esetén is $E(m) \to 1$ ($\Phi(+\infty) = 1$). Bizonyítható, hogy az erőfüggvény szimmetrikus az $m = m_0$ egyenesre. Az erőfüggvény görbéje a 22*b* ábrán látható.



22*b* ábra b)

Az erőfüggvény n-től, m-től, m_0-tól, σ-tól és ε-tól függ.

Látható, hogy a másodfajú hiba el nem követésének a valószínűsége (azaz a próba ereje) annál nagyobb, minél távolabb esik m az m_0-tól.

Hogyan növelhető a próba ereje adott m, σ és ε esetén? Az n növelésével, ugyanis ha $n \to \infty$, akkor

$$E(m) = 1 - \Phi\left(u_\varepsilon - \sqrt{n}\frac{m-m_0}{\sigma}\right) + \Phi\left(-u_\varepsilon - \sqrt{n}\frac{m-m_0}{\sigma}\right) \to 1.$$

Az olyan próbát, amelynek ereje az ellenhipotézis minden elemére 1-hez tart, ha $n \to \infty$, *konzisztens próbának* nevezzük. Az u-próba tehát konzisztens.

Ennek következtében a megfigyelések számának növelésével a másodfajú hiba elkövetését (konzisztens próba esetén) tetszőleges kicsivé tehetjük. A tárgyalandó próbák jó része ilyen természetű.

Figyeljük meg, hogy az u-próba során a kritikus tartomány azon u értékek halmaza a számegyenesen, amelyekre a H_0 hipotézist elutasítjuk, vagyis a

$$\{u > u_\varepsilon\} \cup \{u < -u_\varepsilon\}$$

egyesített halmaz. Ez egyenértékű azzal, hogy a kritikus tartomány mindazon $\bar{x}$ értékek halmaza, amelyekre

$$\bar{x} > m_0 + u_\varepsilon \frac{\sigma}{\sqrt{n}} \quad \text{vagy} \quad \bar{x} < m_0 - u_\varepsilon \frac{\sigma}{\sqrt{n}}.$$

Ez ismét egyenértékű azzal, hogy a kritikus tartomány az n-dimenziós mintatérben mindazon $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ pontok K halmaza, amelyek az előbbi kritikus tartományoknak megfelelnek. Ha a mintából adódó $\mathbf{x} \in K$, a H_0 hipotézist elutasítjuk, ha $\mathbf{x} \notin K$, a nullhipotézist elfogadjuk.

A gyakorlatban az n-dimenziós tér helyett inkább az u vagy az $\bar{x}$ lehetséges értékeit tartalmazó számegyenesen jelöljük meg a kritikus tartományt. Lehet még $\bar{x} - m_0$ értékével skálázott számegyenest is használni, ekkor azt adjuk meg direkt módon, hogy $\bar{x}$ és m_0 között milyen maximális eltérés esetén fogadjuk el vagy vetjük el H_0-t.

3. A statisztikai próbák általános tárgyalása

Minden diszkrét vagy folytonos eloszlást egyértelműen meghatároz az eloszlásfüggvénye. *Megfogalmazhatjuk tehát a statisztikai próbák elvét az* $F(x, T)$ eloszlásfüggvénnyel, ahol a T paraméter (vagy paramétervektor) valamely Θ paraméterhalmaz eleme.

A nullhipotézis:

$$H_0 : P(X < x) = F(x, T),$$

ahol $T \in \omega_0$ (ω_0 az Θ egy részhalmaza).

Az ellenhipotézis:

$$H_1 : P(X < x) = F(x, T),$$

ahol $T \in \omega_1$ (ω_1 az Θ egy másik részhalmaza, ω_0 és ω_1 diszjunkt halmazok).

A próba azt jelenti, hogy az n-dimenziós $(x_1, x_2, \ldots, x_n)$ mintateret egy $g(x_1, x_2, \ldots, x_n)$ statisztikai függvény segítségével (amelynek eloszlását ismerjük) két részre osztjuk: egy K kritikus tartományra és egy A elfogadási tartományra. Ha az $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ mintapont a K tartományba esik, akkor H_0-t elutasítjuk, és úgy tekintjük, hogy a tényleges eloszlás a H_1 feltevésnek megfelelő. Ha viszont a mintapont A halmazba esik, akkor H_0-t és a hozzá tartozó eloszlásfüggvényt fogadjuk el. Az $1-\varepsilon$-szintű kétoldali kritikus tartomány mindazon mintapontok halmaza, amelyekhez tartozó próbastatisztika értéke $g_1(\varepsilon)$-nál kisebb vagy $g_2(\varepsilon)$-nál nagyobb:

$$K: \{(X_1, \ldots, X_n) : g(X_1, \ldots, X_n) < g_1(\varepsilon)$$
$$\text{vagy} \quad g(X_1, \ldots, X_n) > g_2(\varepsilon)\},$$

az elfogadási tartomány:

$$A: \{(X_1, \ldots, X_n) : g_1(\varepsilon) \leqq g(X_1, \ldots, X_n) \leqq g_2(\varepsilon)\}.$$

Az ε-tól függő $g_1(\varepsilon) < g_2(\varepsilon)$ számokat úgy választjuk meg, hogy az elfogadási tartományba esés valószínűsége minden H_0-nak megfelelő eloszlás esetén legalább $(1-\varepsilon)$ legyen:

$$P(A / T) = P[g_1(\varepsilon) \leqq g(X_1, \ldots, X_n) \leqq g_2(\varepsilon)] \geqq 1-\varepsilon$$

mindegyik $T \in \omega_0$ esetén.

Az elsőfajú hiba valószínűsége legfeljebb

$$P(K / T) = 1 - P(A / T) \leqq \varepsilon$$

minden $T \in \omega_0$ esetén. A másodfajú hiba el nem követésének valószínűsége, vagyis a próba ereje a H_1 hipotézisnek megfelelő minden eloszlás esetében:

$$E(T, g) = P(K / T), \quad \text{ahol } T \in \omega_1.$$

Az $E(T, g)$ (pontosabban az $E_n(T, g)$) erőfüggvény n-től, T-től és az alkalmazott g próbastatisztikától függ.

A próba annál jobb (rögzített $T \in \omega_1$ paraméterérték esetén), minél nagyobb az erőfüggvény értéke (a próba ereje) ezen a helyen. Ha minden $T \in \omega_1$ értékre

$$\lim_{n \to \infty} E_n(T, g) = 1,$$

akkor a próbát *konzisztensnek* nevezzük.

4. A kétmintás u-próba

X és Y véletlen változók normál eloszlásúak ismert σ_1 és σ_2 szórásokkal. Az X-re és Y-ra vett n-, ill. m-elemű független minták alapján vizsgálni szeretnénk a

$$H_0: M(X) = M(Y)$$

nullhipotézist. A próbastatisztika:

$$u = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}.$$

Ha H_0 igaz, akkor u $N(0, 1)$-eloszlású, és adott ε-hoz a kritikus tartomány:

$$\{u < -u_\varepsilon \quad \text{vagy} \quad u > u_\varepsilon\},$$

a H_1: $M(X) \neq M(Y)$ alternatívával szemben. Itt u_ε-t a

$$2 \cdot \Phi(\varepsilon) - 1 = 1 - \varepsilon$$

egyenletből állapítjuk meg.

Bizonyítható, hogy ez a próba is konzisztens.

5. A Student-féle egymintás t-próba

Normál eloszlású változóra csak akkor alkalmazható az u-próba, ha az elméleti szórás ismert. A gyakorlatban inkább csak az

$$S_n^{*2} = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$$

korrigált szórás négyzetét ismerjük. Ilyenkor az $N(m, \sigma)$-eloszlású változóra a

$$H_0: M(X) = m_0$$

hipotézis ellenőrzésére a

$$t = \sqrt{n}\,\frac{\bar{X} - m_0}{S_n^*}$$

próbastatisztikát képezzük. Kimutatható, hogy ez a változó $n-1$ szabadságfokú, Student-eloszlású. A 8. táblázatból tetszőleges ε szinthez meghatározható olyan $t_\varepsilon = t_{\text{táblázatbeli}}$ érték, hogy

$$P(-t_\varepsilon \leqq t \leqq t_\varepsilon) = \varepsilon$$

legyen. A kritikus tartomány szimmetrikus az origóra:

$$|t| > t_\varepsilon.$$

Az ellenhipotézis:

$$H_1: M(X) = m \neq m_0.$$

Az erőfüggvény (= a második fajta hiba el nem követésének valószínűsége):

$$E(m) = 1-P\left(-t_\varepsilon \leqq \sqrt{n}\,\frac{\bar{X}-m_0}{S_n^*} \leqq t_\varepsilon \,/\, M(X) = m \neq m_0\right) =$$

$$= 1-P\left(-t_\varepsilon \leqq \frac{1}{\frac{S_n^*}{\sigma}}\left(\frac{\bar{X}-m}{\sigma}\sqrt{n}+\frac{m-m_0}{\sigma}\sqrt{n}\right) \leqq t_\varepsilon \,/\, m\right)=$$

$$= 1-P\left(-t_\varepsilon \leqq \frac{Z+d}{\chi_{n-1}} \leqq t_\varepsilon \,/\, m\right).$$

Itt $\chi_{n-1} = \dfrac{S_n^*}{\sigma}$ $(n-1)$ paraméterű, χ-eloszlású, $Z = \dfrac{\bar{X}-m}{\sigma}\sqrt{n}$ $N(0,1)$-eloszlású, $d = \dfrac{m-m_0}{\sigma}\sqrt{n}$. Az így kapott valószínűségi változó egy $N(d,1)$-eloszlású és egy $n-1$ paraméterű χ-eloszlású változó hányadosa. Ennek eloszlása az ún. *nemcentrális t-eloszlás,* amely csak d-től és n-től függ.

Ha tehát

$$|t_{\text{szám}}| = \left|\sqrt{n}\,\frac{\bar{X}-m_0}{S_n^*}\right| \begin{cases} \leqq t_{\text{tábl}} = t_\varepsilon, \\ > t_{\text{tábl}} = t_\varepsilon, \end{cases}$$

$$\text{az } M(X)=m_0 \text{ hip.} \begin{cases} \text{elfogadjuk} \\ \text{elvetjük} \end{cases}$$

$1-\varepsilon$ valószínűségi szinten. Vigyázat: a 8. táblázatban $(1-\varepsilon)\cdot 100\%$ biztonsági szinthez, adott $n = f+1$ mérésszámhoz tartozó $|t|$ abszolútértéket tartalmaz, nem pedig t értékeket!

Egyoldali próba esetén, ha például az ellenhipotézis

$$H_1'\colon\ M(X)=m>m_0,$$

akkor a kritikus tartomány a számegyenes olyan t_ε-tól jobbra eső része, amelybe t (H_0 fennállása esetén) ε valószínűséggel esik:

$$P(t>t_\varepsilon \,/\, H_0) = \varepsilon.$$

Például $n=30$ mérés, 0,05 elsőfajtájú hibavalószínűség esetén *kétoldali próbánál* az elfogadási tartomány

$$-2{,}045 \leqq t_{\text{szám}} \leqq 2{,}045,$$

95% biztonsággal. *Egyoldali próbánál* (ha azt sejtjük, hogy $m>m_0$ jobb ellenhipotézis) az elfogadási tartomány:

$$t_{\text{szám}} \leqq 2{,}045,$$

de már 97,5% biztonsággal. A 8. táblázat megfelelő részlete:

$n=f+1$	Statisztikai biztonság		
	90%	95%	98%
30	1,699	2,045	2,462

Térjünk vissza a kétoldali próbára! Ha például $t_{\text{számított}}=2{,}4$, akkor:

- 95% biztonsági szinten elutasítunk,
- 98% biztonsági szinten elfogadunk.

Ennek magyarázatát az u-próbánál már megadtuk: kisebb biztonsági szinten szűkebb intervallumba esnek a változó értékei. *Vehetjük azonban a magasabb biztonsági szintet* (és akkor elfogadjuk a H_0 hipotézist), *ezzel az elsőfajú hiba valószínűségét lejjebb szorítottuk* ($\varepsilon=0{,}02$-re). *Végiggondolható, hogy ezen az áron viszont a másodfajú hiba valószínűségét megnöveljük* (mint az u-próbánál).

Ugyanez érvényes minden statisztikai próbára.

Mivel itt *egy* sokaságból vettünk n-elemű mintát, ezt a próbát *egymintás t-próbának* nevezik.

A t-próbához *számítógépes programot* adunk a most következő gyakorló feladatok között.

6. *F*-próba és kétmintás *t*-próba

a) Két független, normál eloszlású változó tapasztalati szórása kissé eltér. Feltehető-e, hogy az egész sokaságban megegyezik a két elméleti szórás, $\sigma_1=\sigma_2$?

Osszuk el a nagyobbik szórásnégyzetet a kisebbel! Az így kapott $F=\frac{S_2^{*2}}{S_1^{*2}}$ számlálójáról bizonyítható, hogy n_2-1 szabadságfokú, χ^2-eloszlású változó, nevezőjéről igazolható, hogy n_1-1 szabadságfokú, χ^2-eloszlású. A két változó független. $\sigma_1=\sigma_2$ esetén az $F=\frac{S_2^{*2}}{S_1^{*2}}$ (n_2-1), (n_1-1) szabadságfokú, F-eloszlású.

Az F-eloszlásból megállapítható két olyan határ, amelybe a változó nagy (95%-os) valószínűséggel esik. Az egyik határ 1-nél kisebb, a másik határ 1-nél nagyobb. A $\sigma_1=\sigma_2$ hipotézis megvizsgálására elegendő csak azt vizsgálni, hogy az 1-nél nem kisebb $F=\frac{S_2^{*2}}{S_1^{*2}}$ a felső határ alatt van-e.

Két független, normál eloszlású változó szórásának az egyezését F-próbával dönthetjük el: osszuk el a nagyobbik tapasztalati korrigált szórásnégyzetet a kisebbikkel, az $F=\frac{S_2^{*2}}{S_1^{*2}}$ *így számított értékét hasonlítsuk össze a 9. táblázatbeli F-eloszlás értékével!*

$$\text{Ha } F_{\text{számított}} \begin{cases} \leqq F_{95} \text{ (táblázatbeli érték)}, \\ > F_{95} \text{ (táblázatbeli érték)}, \end{cases} \text{ akkor } \begin{cases} \text{elfogadjuk} \\ \text{elvetjük} \end{cases}$$

a két szórás egyezését (a $\sigma_1=\sigma_2$ feltevést).

A táblázat értékei csak 95% biztonsági szinthez vannak kiszámítva.

b) Ha $\sigma_1=\sigma_2$, akkor feltehető-e a két független, normál eloszlású sokaságban, hogy a várható értékek megegyeznek; $m_1=m_2$?

A várható értékek egyezésére (az $m_1=m_2$ feltevésre) az alábbi próbát tehetjük (feltéve, hogy $\sigma_1=\sigma_2$):

Kiszámítjuk a

$$|t|=\frac{|\bar{x}_1-\bar{x}_2|}{\sqrt{\frac{(n_1-1)S_1^{*2}+(n_2-1)S_2^{*2}}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

statisztikai függvény (röviden: statisztika) feladatbeli (aktuális) értékét. Bizonyítható, hogy ez a statisztika Student-eloszlású, n_1+n_2-2 szabadságfokkal. *Ezt összehasonlítjuk a Student-eloszlás n_1+n_2-2 szabadságfokú, adott biztonsági szinthez tartozó értékével.*

$$\text{Ha } t_{\text{számított}} \begin{cases} \leqq t_{\text{táblázatbeli}}, \\ > t_{\text{táblázatbeli}}, \end{cases} \text{ akkor az } m_1=m_2 \text{ feltevést } \begin{cases} \text{elfogadjuk} \\ \text{elvetjük} \end{cases}$$

az adott biztonsági szinten.

Ha mind a szórások, mind a várható értékek megegyeznek, akkor a két normál eloszlású változó eloszlásfüggvénye is azonos.

Példa. Két normál eloszlású sokaságból mintát vettünk.

Mintaszám	Mérésszám	Mintaátlag	Mintabeli szórás
I. minta	$n_1=20$	$\bar{x}_1=171{,}6000$	$s_1^*=3{,}2847$
II. minta	$n_2=49$	$\bar{x}_2=169{,}2240$	$s_2^*=4{,}6385$

a) $F_{\text{számított}}=1{,}9941<F_{95}=2{,}0035$, l. a 9. táblázatot:

f_2	f_1: 30	f_1: 50
19	2,07	2,0

Így 95% biztonsági szinten elfogadjuk, hogy a két sokaságban a szórások megegyeznek. Ha ez teljesül, akkor:

b) vizsgálhatjuk, hogy egyeznek-e a várható értékek is!

$|t_{\text{számított}}|=2{,}0829,$

ezt a Student-eloszlás $f = n_1 + n_2 - 2 = 67$ szabadságfokú t értékével összehasonlítva (az $n = f+1 = 68$-nak megfelelő sorban keresve):

$n = f+1$	Statisztikai biztonság	
	95%	98%
61	2,000	2,390
121	1,980	2,358

$|t_{\text{szám}}| < t_{98}$, de $|t_{\text{szám}}| > t_{95}$.

A feltevés elfogadása mellett is, ellene is dönthetünk (első esetben megnövekszik a másodfajtájú hiba valószínűsége, kisebb a próba ereje). *Dönthetünk ilyenkor úgy is, hogy további méréseket kérünk.*

A példa teljes adatait, *a számítógépprogramot* az F- és kétmintás t-próbához l. a soron következő gyakorló feladatok közt.

7. Welch-próba

Legyenek az X és Y valószínűségi változók normál eloszlásúak, és vizsgáljuk *a H_0: $M(X) = M(Y)$ hipotézist, ha az elméleti szórások nem ismeretesek és az F-próba a $D(X) \neq D(Y)$ eredményre vezet!* Ez esetben a kétmintás t-próba nem alkalmazható, sőt egzakt próba nem is adható meg. A H_0 hipotézis helyességének eldöntésére Welch a következő statisztikát adta meg:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{S_n^2}{n} + \dfrac{S_m^2}{m}}},$$

ahol $\bar{X}$ és $\bar{Y}$ a mintaközepek, S_n^2, ill. S_m^2 a tapasztalati korrigálatlan szórásnégyzetek a megfelelő

$$X: x_1, x_2, \ldots, x_n,$$

és

$$Y: y_1, y_2, \ldots, y_m$$

mintákból számítva.

A H_0 hipotézis fennállása esetén a t-statisztika közelítőleg Student-eloszlású f paraméterrel, ahol f értékét a következő módon határozzuk meg:

$$\frac{1}{f} = \frac{1}{m-1}\left[\frac{\dfrac{S_m^2}{m}}{\dfrac{S_n^2}{n} + \dfrac{S_m^2}{m}}\right]^2 + \frac{1}{n-1}\left[\frac{\dfrac{S_n^2}{n}}{\dfrac{S_n^2}{n} + \dfrac{S_m^2}{m}}\right]^2.$$

Ha m és n elég nagy (≥ 40), akkor t közelítően standard normál eloszlású.

8. Bartlett-próba

Két normál eloszlású változó ismeretlen szórásának egyenlőségét az F-próba segítségével döntjük el.

Kettőnél több normál eloszlású változó szórásának megegyezését a Bartlett-próbával vizsgáljuk.

$X^{(1)}, X^{(2)}, \ldots, X^{(k)}$ normál eloszlású valószínűségi változók, amelyekre vett minták alapján az alábbiak ismeretesek:

Változó	Minta elemszám	Minta	Szórás	Szabadságfok
$x^{(1)}$	n_1	$x_{11}, x_{12}, \ldots, x_{1n_1}$	$s_{n_1}^*$	$f_1 = n_1 - 1$
$x^{(k)}$	n_k	$x_{k1}, x_{k2}, \ldots, x_{kn_k}$	$s_{n_k}^*$	$f_k = n_k - 1$

Legyen a szabadságfokok összege f:

$$f = \sum_{i=1}^{k} f_i,$$

ezenkívül legyen a korrigált szórásnégyzetek súlyozott közepe

$$S^{*2} = \frac{1}{f}\sum_{i=1}^{k} f_i S_{n_i}^{*2}.$$

Bartlett eredménye szerint a

$$K^2 = \frac{2{,}3026}{c}\left(f\cdot \lg S^{*2} - \sum_{i=1}^{k} f_i \lg S_{n_i}^{*2}\right)$$

véletlen változó közelítően χ^2-eloszlású, $k-1$ paraméterrel. Itt

$$c = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{f_i} - \frac{1}{f}\right),$$

és az lg a 10-es alapú logaritmust jelenti.

A közelítés akkor megfelelő, ha az egyes minták legalább négyeleműek.

Ha a mintaelemek száma minden változónál azonos n, akkor a formula egyszerűbb.

9. Pitman-próba *(két egyforma, de nem feltétlenül normál eloszlású változó várható értékének egyezésére)*

Fontos kérdés a gyakorlatban, hogy az X, Y véletlen változóra vett mintaátlagok mekkora különbsége esetén célszerű azt feltételezni, hogy az egész sokaságbeli elméleti várható értékek megegyeznek:

$$H_0\colon\ M(X) = M(Y),$$

és mikor célszerű ezt elvetni. Normál eloszlású, azonos szórású sokaságoknál ezt a kétmintás t-próbával ellenőrizzük. Normál eloszlás, különböző szórás, elég nagy n esetén Welch-próbát alkalmazunk.

Ha a normalitás nem tételezhető fel, de a két eloszlás típusa azonos, akkor Pitman-próbát alkalmazunk.

Legyen az X-re vonatkozó minta $X_1, X_2, \ldots, X_n$ (I. minta),

– az Y-ra vonatkozó minta $Y_1, Y_2, \ldots, Y_m$ (II. minta),

– a két minta egyesítéséből nyert minta $Z_1, Z_2, \ldots, Z_{n+m}$ („III. minta”, ahol $Z_1 = X_1$, $Z_2 = X_2, \ldots, Z_{n+m} = Y_m$).

A III. mintát $\binom{n+m}{m}$-féleképpen oszthatjuk be két, n, ill. m elemet tartalmazó csoportba (I., ill. II. típusú mintába). A rendelkezésre álló megfigyeléssorozat ezen csoportosítások egyike.

Kis minták esetén számítsuk ki az I. és II. minta számtani közepei közti különbség abszolút értékét, ily módon képezzük a

$$d_j = |\bar{X} - \bar{Y}| = \frac{n+m}{nm}\left(\sum_{i=1}^{n} X_i - n\bar{Z}\right)$$

próbastatisztikát minden lehetséges módon $\left(j = 1, 2, \ldots, \binom{m+n}{m}\right)$!

Legyen az indexezés olyan, hogy

$$d_1 \geqq d_2 \geqq \ldots \geqq d_j \geqq \ldots \geqq d_{\binom{m+n}{m}}.$$

Válasszuk meg az ε-t (az elsőfajú hiba szintjét)! Legyen k olyan egész szám, hogy

$$\frac{k}{m+n} = \varepsilon.$$

Ha a felsorolt $d_1 \geqq d_2 \geqq \ldots \geqq d_{\binom{m+n}{m}}$ számok között k-nál kevesebb olyan van, amely az adott I. és II. mintához tartozó d-nél nagyobb, azaz ha

$$d_k < d,$$

akkor a H_0 hipotézist elvetjük ($1-\varepsilon$ szinten), különben elfogadjuk.

Nagy minták esetén ez az eljárás a sok számolás miatt nehezen végezhető el.

Az

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

ún. STIRLING-formula alapján:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \approx \frac{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}}{\left(\frac{k}{e}\right)^k \sqrt{2\pi k}\left(\frac{n-k}{e}\right)^{n-k}\sqrt{2\pi(n-k)}} =$$

$$= \frac{n^{n+\frac{1}{2}}}{\sqrt{2\pi} k^{k+\frac{1}{2}}(n-k)^{n-k+\frac{1}{2}}},$$

ez például $n=50$, $k=30$ esetén:

$$\binom{n}{k} \approx 4{,}74 \cdot 10^{13},$$

vagyis 20 és 30 elemből álló I. és II. minta esetén is kb. 10^{13} darab mintastatisztikát kell elkészíteni, és ezeket nagyság szerinti sorrendbe rakni. Ez még nagyteljesítményű számítógéppel sem valósítható meg. Ilyenkor azonban a

$$d = |\bar{X} - \bar{Y}| \frac{1}{\sqrt{\frac{(n-1)S_1^{*2} + (m-1)S_2^{*2}}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

statisztikai függvény aszimptotikusan Student-féle t-eloszlású, $n+m-2$ szabadságfokkal. Így tehát a kétmintás t-próba alkalmazható az elfogadáshoz, ill. elvetéshez.

10. Szórásanalízis *(kettőnél több normál eloszlású véletlen változó várható értékének egyezésére)*

Ha X és Y két, azonos szórású, normál eloszlású változó, akkor a várható értékek egyezéséről a kétmintás t-próbával dönthetünk. A szórások egyezését F-próbával mutathatjuk ki.

Vizsgáljuk most azt, hogy $X_1, X_2, \ldots, X_r$ független, normál eloszlású véletlen változók várható értéke (ismeretlen, de egyforma σ szórás esetén – Bartlett-próba) megegyezik-e. Tehát a nullhipotézis:

$$H_0\colon\ M(X_1) = M(X_2) = \ldots = M(X_r) = m.$$

A megfigyelt minták és a belőlük kiszámítható

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad (i = 1, 2, \ldots, r)$$

mintaátlagok a következők:

Változó	Minta elemszám	Minta	Mintaátlag
x_1	n_1	$x_{11}, x_{12}, \ldots, x_{1n_1}$	$\bar{x}_1$
x_2	n_2	$x_{21}, x_{22}, \ldots, x_{2n_2}$	$\bar{x}_2$
x_r	n_r	$x_{r1}, x_{r2}, \ldots, x_{rn_r}$	$\bar{x}_r$

Ha H_0 igaz, akkor minden x_{ij} mintaelem azonos $N(m, \sigma)$-eloszlású.

Képezzük az összes megfigyelés számtani közepét (ami megegyezik a mintaátlagok súlyozott számtani közepeivel):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{r} n_i\bar{x}_i.$$

Itt $n = \sum_{i=1}^{r} n_i$.

A $\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2$ ún. *teljes négyzetösszeget* két négyzetösszeg összegére bontjuk. Az egyik az ún. *csoportok közötti* $\sum_{i=1}^{r} n_i(\bar{x}_i-\bar{x})^2$

négyzetösszeg, a másik *a csoportokon belüli* $\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$ *négyzet-összege*. Ugyanis minden sorban:

$$\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i) = n_i\bar{x}_i - n_i\bar{x}_i = 0$$

(az x_{ij} mintaelemek mintaátlagtól való eltérésének összege 0), így

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{x}) = \sum_{i=1}^{r}\left[(\bar{x}_i-\bar{x})\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)\right] = 0,$$

ennek következtében a teljes négyzetösszeg:

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i+\bar{x}_i-\bar{x})^2 =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 + 2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{x}) +$$

$$+\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{x})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 + \sum_{i=1}^{r}n_i(\bar{x}_i-\bar{x})^2.$$

Ha a H_0: $M(X_1) = M(X_2) = \ldots = M(X_r) = m$ nullhipotézis igaz, akkor mindegyik x_{ij} véletlen változó egyforma eloszlása és független volta miatt a *teljes négyzetösszeg*,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2 = (n-1)S^{*^2}$$

$n-1$ szabadságfokkal χ^2-eloszlású.

Ugyanúgy a

$$\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

változó is χ^2-eloszlású, n_i-1 szabadságfokkal. Bizonyítható, hogy χ^2-eloszlású változók összege ugyancsak χ^2-eloszlású, szabadságfoka

az összeg tagjai szabadságfokának az összege. Ilyen alapon a *csoporton belüli négyzetösszeg*,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

szintén χ^2-eloszlású, $\sum_{i=1}^{r}(n_i-1) = n-r$ szabadságfokkal.

Igazolható, hogy ha a H_0 hipotézis igaz, akkor a *csoportok közti négyzetösszeg*,

$$\sum_{i=1}^{r}n_i(\bar{x}_i-\bar{x})^2$$

független a csoportokon belüli négyzetösszegtől és χ^2-eloszlású, $r-1$ szabadságfokkal.

Akkor viszont az

$$S_1^2 = \frac{\sum_{i=1}^{r}n_i(\bar{x}_i-\bar{x})^2}{r-1} \quad \text{és} \quad S_2^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{n-r}$$

változók függetlenek, közös várható értékük $M(S_1^2)=M(S_2^2)=\sigma$, *hányadosuk (ha H_0 igaz) F-eloszlású $(r-1,\ n-r)$ paraméterrel. A*

$$H_0\text{: } M(X_1) = M(X_2) = \ldots = M(X_r) = m$$

hipotézis ellenőrzésére F-próba alkalmazható.

A szórásanalízisben használják az ún. szórásfelbontó táblázatot:

Négyzetösszeg	Négyzetösszeg neve	Szabadságfok	Tapasztalati szórásnégyzet
$(r-1)s_1^2 = \sum_{i=1}^{r}n_i(\bar{x}_i-\bar{x})^2$	Csoportok közti	$r-1$	s_1^2
$(n-r)s_2^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$	Csoporton belüli	$n-r$	s_2^2
$(n-1)s^{*^2} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2$	Teljes	$n-1$	s^{*^2}

Érdemes megfigyelni, hogy a szórásnégyzetek a szabadságfokkal szorozva összegeződnek:

$$(n-1)S^{*^2} = (r-1)S_1^2+(n-r)S_2^2,$$

a szabadságfokok is „összegeződnek":

$$n-1 = r-1+n-r.$$

A teljes négyzetösszeget két, egymástól független négyzetösszeg összegére bontottuk: az egyik magyarázza, méri a csoportok közti eltéréseket *(a szisztematikus hibát)*, a másik a csoporton belüli eltéréseket *(a véletlen hibát)*.

Megjegyzés

Az itt közölt anyag a szórásanalízis egyik fejezete, az ún. *egyszeres osztályozás (egyszerű csoportosítás)*. Ha például egyféle anyagból sodratot gyártunk r különböző gépen, akkor azt vizsgáljuk, hogy a fellépő *egyetlen hatás* vagy másként *egyetlen faktor* miatt (ti. a különböző gépeken való gyártás miatt) azonos normál eloszlású-e az r db sokaság. Hétköznapi nyelven szólva azt vizsgáljuk, hogy *egy hatás (egy faktor) esetén egyforma minőségű-e a termék.*

Ha viszont k különböző anyagból, m különböző gépen állítunk elő egy alkatrészt, akkor *két faktor hatása után vizsgáljuk, hogy pl. törőszilárdság szempontjából egyforma minőségű-e a termék,* melyik gép, ill. melyik anyag a kedvezőbb. *Az egyik faktor kölcsönhatását (interakcióját) ki kell küszöbölni. Ez már kétfaktoros (kétszeres osztályozáson alapuló) szórásanalízis.*

Létezik három-, négy- stb. faktoros vizsgálat is. Ha pl. háromszoros osztályozásnál az egyes faktorok n_1-, n_2-, n_3-féleképpen, azaz n_1, n_2, n_3 *szinten* fordulhatnak elő, akkor a lehetséges összes szintkombinációk *(cellák)* száma $n_1n_2n_3$. Ez a szám sokszor olyan nagy, hogy a szükséges idő és költség miatt nem valósítható meg ennyi kísérletsorozat. Ilyenkor megelégedünk olyan kísérletsorozattal, amelynél bizonyos cellák üresen maradnak *(nem teljes kísérleti elrendezések,* latin négyzet, nem teljes blokk, véletlen blokk). Olyan cellákat kell üresen hagyni, amelyeknek a hiánya alapján még megfelelő, megbízható következtetéseket vonhatunk le.

A kvalitatív és kvantitatív faktorokat együtt vizsgáló szórásanalízis a *kovarianciaanalízis.*

A különböző szint például azt jelenti, hogy n_1 különböző anyagból, n_2 különböző esztergán és n_3 különböző polírozógépen gyártjuk az alkatrészt. Ezekkel a témákkal bővebben ismerkedhet az Olvasó például Vincze István: Matematikai statisztika ipari alkalmazásokkal c. könyvében.)

11. Szekvenciális módszer hipotézisvizsgálatra

Az eddigiek során az $x_1, x_2, \ldots, x_n$ mintaelemekből elkészítettük a próbastatisztika aktuális értékét, és egy-egy hipotézisről ennek az egyetlen számadatnak a birtokában döntöttünk.

A szekvenciális módszerrel történő hipotézisvizsgálat ennél alaposabb. Tegyük fel, hogy X valószínűségi változó eloszlását az $F(x, T)$ eloszlásfüggvény jellemzi. A H_0 nullhipotézis az, hogy az ismeretlen paraméter (vagy paramétervektor) értéke $T=T_0$, a H_1 alternatív (ellen) hipotézis az, hogy $T=T_1 \neq T_0$.

Minden egyes mintaelem megfigyelése után háromféle döntés valamelyikét hozzuk:

$$\left.\begin{array}{l} a)\ \text{a } H_0 \text{ hipotézist elfogadjuk} \\ b)\ \text{a } H_0 \text{ hipotézist elvetjük} \end{array}\right\} \text{(végső döntés),}$$

c) el sem fogadjuk, el sem vetjük H_0-t, hanem újabb megfigyelést végzünk (elodázás).

A szekvenciális vizsgálatnál előre megadjuk a lehetséges tévedések valószínűségét. Legyen az elsőfajú hiba valószínűsége α, a másodfajúé β. Számítsuk ki az

$$A = \frac{1-\beta}{\alpha} \quad \text{és} \quad B = \frac{\beta}{1-\alpha}$$

határmennyiségeket! A döntést az ún. L likelihood-hányados segítségével végezzük:

$L \leqq B$	$B < L < A$	$L \geqq A$
H_0-t elfogadjuk (elfogadási tartomány)	Újabb mintaelemet veszünk (indifferens tartomány)	H_0-t elvetjük (visszautasítási, kritikus tartomány)

A középső esetben a mintavételt mindaddig folytatjuk, míg L értéke az A vagy B határokat át nem lépi. Ha az L valószínűséghányados értéke igen nagy mintaelemszám esetén is mindig B és A közé esett, a próbát befejezhetjük úgy, hogy H_0-t vagy H_1-et fogadjuk el attól függően, hogy a valószínűséghányados B-hez vagy A-hoz esik közelebb. Ez az eset csak ritkán következik be, mégpedig akkor, ha T értéke sem T_0-val, sem T_1-gyel nem egyezik.

Az L likelihood-hányados minden egyes mintavétel után annak a valószínűségét adja meg, hogy hányszorta valószínűbb az adott minta előfordulása a H_1 feltevéssel, mint H_0 esetén. Az első, második, ..., k-adik mintavétel során (ha addig a valószínűséghányados értéke az indifferens tartományba esett) rendre kiszámítjuk az

$$L = l_1 = \frac{P(X=x_1 / T_1)}{P(X=x_1 / T_0)},$$

$$L = l_1 l_2 = \frac{P(X=x_1 / T_1)\, P(X=x_2 / T_1)}{P(X=x_1 / T_0)\, P(X=x_2 / T_0)}, \ldots,$$

$$L = l_1 l_2 \ldots l_k = \frac{\prod_{i=1}^{k} P(X=x_i / T_1)}{\prod_{i=1}^{k} P(X=x_i / T_0)}$$

valószínűség-hányados értékeit (a független mintavétel miatt az együttes bekövetkezés valószínűsége az egyes valószínűségek szorzata). Mivel *folytonos esetben* az x_i közelébe esés feltételes valószínűsége a sűrűségfüggvénnyel adható meg (például: $P(x_i \leqq X \leqq x_i + \Delta x_i / T_1) \approx$

$\approx f(x_i, T_1)\, \Delta x_i$), ezért itt a Δx_i értékekkel való egyszerűsítés után a likelihood-hányados a k-adik mintavétel után:

$$L = l_1 l_2 \ldots l_k = \frac{\prod_{i=1}^{k} f(x_i, T_1)}{\prod_{i=1}^{k} f(x_i, T_0)}.$$

Ha logaritmussal dolgozunk, akkor szorzat helyett a könnyebben kezelhető összeadáshoz jutunk. Folytonos esetben tehát:

$$\lg L = \sum_{i=1}^{k} \lg l_i = \sum_{i=1}^{k} \lg \frac{f(x_i, T_1)}{f(x_i, T_0)},$$

ill. diszkrét esetben:

$$\lg L = \sum_{i=1}^{k} \lg \frac{P(X=x_i / T_1)}{P(X=x_i / T_0)}.$$

Természetesen ekkor az elfogadási tartomány határa $\lg B$, a kritikus tartomány határa $\lg A$ lesz.

A szekvenciális módszer becsléselméleti és hipotézisvizsgálati problémákra a magyar származású Abraham Waldtól származik. *Előnyei:*

– a végső döntés meghozatalára lényegesen kisebb elemszámú mintára van szükségünk, mint a kötött mintaelemszámú vizsgálatoknál ugyanolyan első- és másodfajú hibavalószínűségek mellett;

– itt a másod- és elsőfajú hiba valószínűségét előre meg tudjuk szabni.

Gyakorló feladatok

1. Egy automata csővágó gép 1200 mm-es darabok levágására van beállítva. A gép 1200-tól eltérő hosszúságú darabokat is levág. A levágott cső hossza véletlentől függő változó, előzetes adatfelvételből tudjuk, hogy normál eloszlású. Az 1200 mm-től felfelé és lefelé maximálisan 9 mm eltérés még elfogadható. Normál eloszlásnál a várható értéktől felfelé és lefelé 3σ az eltérés, 99,73% valószínűséggel. Ennek alapján úgy tekintjük, hogy az egész normál eloszlású sokaságban a várható csőhossz $m_0 = 1200$ mm, a szórás $\sigma = 3$ mm. Kiválasztunk $n = 16$ legyártott csövet. A mintából kapott méretek: 1193, 1198, 1203, 1191, 1195, 1196, 1199, 1191, 1201, 1196, 1193, 1198, 1204, 1196, 1198, 1200. A minta átlaga $\bar{x} = 1197$ mm.

Elfogadható-e hogy az eltérés nem jelentős (nem szignifikáns), vagyis az egész sokaságban a várható érték $m_0 = 1200$?

A változó normál eloszlású, az elméleti szórás $\sigma = 3$ ismert. u-próbával dönthetünk.

$$|u_{\text{számított}}| = \left|\sqrt{n}\,\frac{\bar{x}-m_0}{\sigma}\right| = \left|\sqrt{16}\,\frac{1197-1200}{3}\right| = 4,$$

az $u = \sqrt{n}\,\dfrac{\bar{x}-m_0}{\sigma}$ statisztikai függvény (röviden statisztika) normál eloszlású, 95% valószínűséggel állíthatjuk, hogy értéke $-1{,}960$ és $+1{,}960$ közé esik (l. az 5. vagy 8. táblázatot):

n	Statisztikai biztonság			
	90%	95%	...	99,9%
∞	1,645	1,960	...	3,291

Mivel $|u_{\text{számított}}| = 4 > u_{\text{táblázatbeli}} = 1{,}960$, ezért az $m_0 = 1200$ feltevést elvetjük, a mintaátlag nem véletlenül tér el tőle, az automatát újból be kell állítani.

Látható, hogy más szinteken döntve (90%, 99,9%) is ugyanehhez a döntéshez jutunk (mindegyik táblázatbeli érték kisebb a számítottnál).

2. Igazoljuk, hogy az u-próba $E_n(m)$ erőfüggvénye *a)* minimális az $m = m_0$ helyen; *b)* szimmetrikus az $m = m_0$ egyenesre; *c)* $E_{n+1}(m) > E_n(m)$, ha $m \neq m_0$!

---

a) Az erőfüggvény a másodfajú hiba el nem követésének valószínűsége. Másodfajú hibát akkor követünk el, ha $m \neq m_0$ és mégis elfogadjuk a $H_0: m = m_0$ nullhipotézist annak alapján, hogy a mintából kapott u az $1-\alpha$ valószínűségű elfogadási tartományra esik:

$$-u_\alpha \leqq \frac{\bar{x}-m_0}{\sigma/\sqrt{n}} \leqq u_\alpha,$$

$$m_0 - u_\alpha \frac{\sigma}{\sqrt{n}} \leqq \bar{x} \leqq m_0 + u_\alpha \frac{\sigma}{\sqrt{n}},$$

$$\frac{m_0-m}{\sigma/\sqrt{n}} - u_\alpha \leqq \frac{\bar{x}-m}{\sigma/\sqrt{n}} \leqq \frac{m_0-m}{\sigma/\sqrt{n}} + u_\alpha,$$

mivel most $\dfrac{\bar{x}-m}{\sigma/\sqrt{n}}$ standard normál eloszlású, ezért a másodfajú hiba el nem követésének valószínűsége (a próba ereje):

$$E_n(m) = 1-\beta = 1 - \Phi\left(\frac{m_0-m}{\sigma/\sqrt{n}} + u_\alpha\right) + \Phi\left(\frac{m_0-m}{\sigma/\sqrt{n}} - u_\alpha\right).$$

Ez a függvény az $m \neq m_0$ feltevés miatt tulajdonképpen nincs értelmezve $m = m_0$-nál. Vegyük $m = m_0$ helyen függvényértéknek a képlettel kapott

$$E_n(m_0) = 1 - \Phi(u_\alpha) + \Phi(-u_\alpha) = 1 - [\Phi(u_\alpha) - \Phi(-u_\alpha)] = 1 - (1-\alpha) = \alpha$$

értéket. Az így kapott $E_n(m)$ erőfüggvény minimuma ott *lehet,* ahol az első deriváltja 0:

$$\frac{dE_n(m)}{dm} = E_n'(m) = \frac{\sqrt{n}}{\sigma}\left[\varphi\underbrace{\left(\frac{m_0-m}{\sigma/\sqrt{n}} + u_\alpha\right)}_{a} - \varphi\underbrace{\left(\frac{m_0-m}{\sigma/\sqrt{n}} - u_\alpha\right)}_{b}\right] = 0,$$

$$\varphi(a) = \varphi(b),$$

$$e^{\frac{-a^2}{2}} = e^{\frac{-b^2}{2}},$$

ez vagy akkor lehetne, mikor $a = b$, azonban

$$a = \frac{m_0-m}{\sigma/\sqrt{n}} + u_\alpha \neq b = \frac{m_0-m}{\sigma/\sqrt{n}} - u_\alpha.$$

Vagy pedig akkor lehet, ha $a=-b$, vagyis

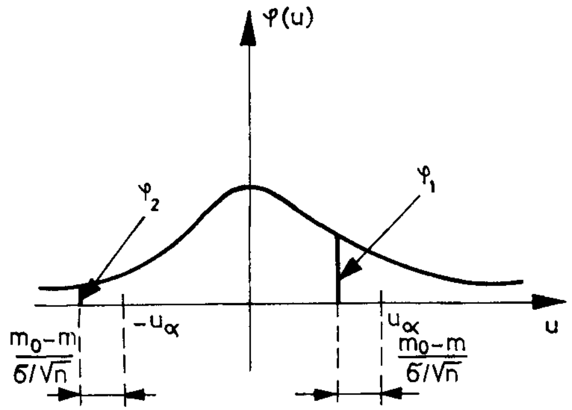
$$\frac{m_0-m}{\sigma/\sqrt{n}}+u_\alpha = -\frac{m_0-m}{\sigma/\sqrt{n}}+u_\alpha,$$

$$m=m_0.$$

Igazolni kell még, hogy $m=m_0$-nál valóban minimum *van*.

Ha $m>m_0$, akkor $\frac{m_0-m}{\sigma/\sqrt{n}}<0$, így a 23*a* ábrán jelölt $\varphi_1>\varphi_2$, tehát

$$E_n'(m) = \frac{\sqrt{n}}{\sigma}(\varphi_1-\varphi_2) > 0,$$



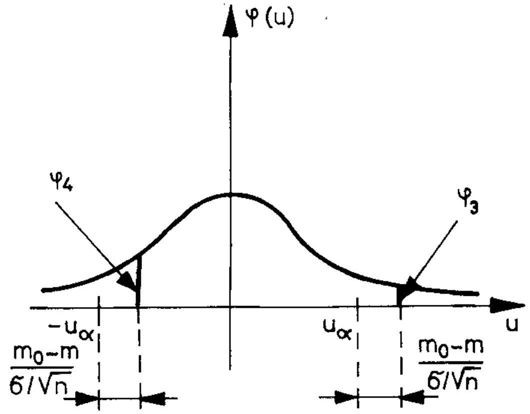
23*a* ábra a)

ezért az erőfüggvény ezen a szakaszon szigorúan monoton növekedő.

Ha $m<m_0$, akkor itt $\frac{m_0-m}{\sigma/\sqrt{n}}>0$, a 23*b* ábrán jelölt $\varphi_3<\varphi_4$, vagyis $m<m_0$ esetén

$$E_n'(m) = \frac{\sqrt{n}}{\sigma}(\varphi_3-\varphi_4) < 0,$$

az erőfüggvény ezen a szakaszon szigorúan monoton csökkenő.



23*b* ábra b)

Az erőfüggvény tehát $m=m_0$-nál minimális, a minimális függvényérték α (az elsőfajú próba valószínűsége).

b) $$E_n(m_0-x) = 1-\Phi\left(\frac{x}{\sigma/\sqrt{n}}+u_\alpha\right)+\Phi\left(\frac{x}{\sigma/\sqrt{n}}-u_\alpha\right),$$

$$E_n(m_0+x) = 1-\Phi\left(-\frac{x}{\sigma/\sqrt{n}}+u_\alpha\right)+\Phi\left(-\frac{x}{\sigma/\sqrt{n}}-u_\alpha\right) =$$

$$= 1-\left[1-\Phi\left(\frac{x}{\sigma/\sqrt{n}}-u_\alpha\right)\right]+\left[1-\Phi\left(\frac{x}{\sigma/\sqrt{n}}+u_\alpha\right)\right] =$$

$$= 1-\Phi\left(\frac{x}{\sigma/\sqrt{n}}+u_\alpha\right)+\Phi\left(\frac{x}{\sigma/\sqrt{n}}-u_\alpha\right),$$

$$E_n(m_0-x) = E_n(m_0+x),$$

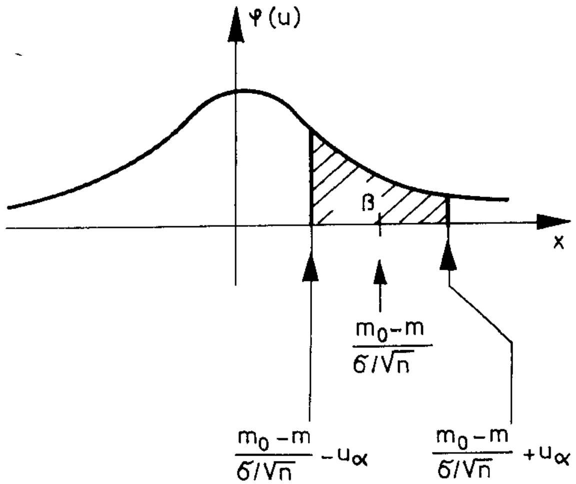
az erőfüggvény valóban szimmetrikus az $m=m_0$ egyenesre.

Mindezeket hozzávéve az elméleti tudnivalókban közölt anyaghoz, látható, hogy az erőfüggvény képe valóban olyan, mint azt a 22*b* ábrán bemutattuk.

c) Láttuk, hogy n-elemű minta esetén az erőfüggvény:

$$E_n(m) = 1-\beta = 1-\left[\Phi\Bigg(\underbrace{\frac{m_0-m}{\sigma/\sqrt{n}}+u_\alpha}_{a}\Bigg)-\Phi\Bigg(\underbrace{\frac{m_0-m}{\sigma/\sqrt{n}}-u_\alpha}_{b}\Bigg)\right].$$

$\Phi(a)-\Phi(b)$ a $\varphi(x)$ haranggörbe alatti terület az (a, b) intervallumon (ez adja meg β értékét, lásd a 24. ábrát az $m<m_0$ esetben). Ha n helyett $(n+1)$-et veszünk, akkor a $2u_\alpha$ széles intervallum jobbra tolódik, a $\varphi(x)$ haranggörbe alatti, *ugyanolyan széles* intervallum fölötti új terület kisebb lesz, mint az előbb volt.



24. ábra

Az $m>m_0$ esetben $\dfrac{m_0-m}{\sigma/\sqrt{n}}$ negatív, a $2u_\alpha$ széles terület balra tolódik, kisebbedik, az erőfüggvény n növelésével nő. Tehát

$$E_{n+1}(m) > E_n(m),$$

ha $m \neq m_0$ ($m=m_0$ esetén mindkét erőfüggvény értéke α). Még inkább így van, ha $n+2$, $n+3$ stb. elemszámú mintát veszünk. *A másodfajú hiba valószínűsége csökken, a próba ereje nő, ha n-elemű minta helyett nálánál nagyobb elemszámú mintát veszünk.*

Láttuk már, hogy ha $n \to \infty$, akkor $E_n(m) \to 1$, $\beta \to 0$ (azaz *a próba konzisztens*).

3. Egy próbával szemben szokás még azt a követelményt is támasztani, hogy torzítatlan legyen, vagyis ha a tényleges eloszlás az ellenhipotézishez tartozik, akkor H_0-t nagyobb valószínűséggel utasítsuk el, mint ha H_0 állna fenn. Azt kívánjuk tehát, hogy a kritikus tartományba esés valószínűsége $>\alpha$

(= a kritikus tartományba esés valószínűsége H_0 fennállása esetén) H_1 ellenhipotézis igaz volta mellett.

Mutassuk meg, hogy *az u-próba torzítatlan is!*

Ha H_0: $M(X) = m_0$ igaz, akkor a kritikus tartományba esés valószínűsége:

$$P(u<-u_\alpha)+P(u>u_\alpha) = \Phi(-u_\alpha)+1-\Phi(u_\alpha) = \alpha.$$

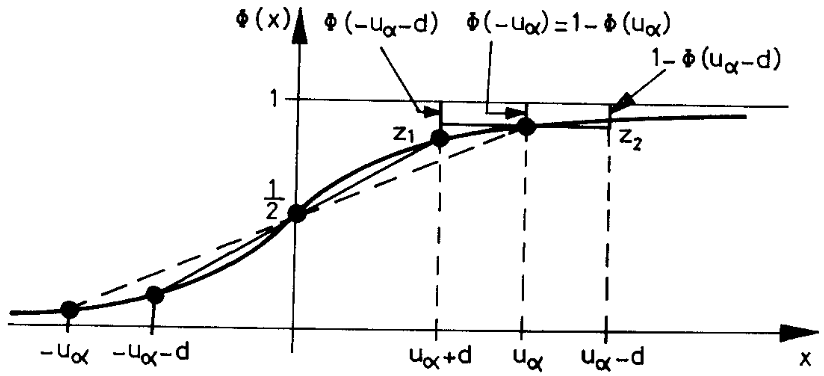
Ha H_1: $M(X) = m \neq m_0$ igaz, akkor az

$$u = \frac{\bar{x}-m_0}{\sigma/\sqrt{n}} = \frac{\bar{x}-m}{\sigma/\sqrt{n}} + \frac{m-m_0}{\sigma/\sqrt{n}} = \frac{\bar{x}-m}{\sigma/\sqrt{n}} + d$$

változó $N(d, 1)$-eloszlású, tehát ez esetben a kritikus tartományra esés valószínűsége:

$$P(u<-u_\alpha)+P(u>u_\alpha) = F(-u_\alpha)+1-F(u_\alpha) = \Phi(-u_\alpha-d)+ \\ +1-\Phi(u_\alpha-d).$$

Legyen $d<0$. Figyeljük most a 25. ábrát (ahol kihasználtuk a $\Phi(x)$ függvény $\left(0; \frac{1}{2}\right)$ pontra való szimmetriáját)! Mivel a $\Phi(x)$ függvény az $(u_\alpha+d; u_\alpha)$



25. ábra

intervallumon „meredekebben” emelkedik, mint az ugyanolyan hosszúságú $(u_\alpha; u_\alpha - d)$ intervallumon, ezért az ábrán feltüntetett jelölésekkel:

$$z_1 > z_2$$

(ezt az érintő iránytangens változásával, a függvény alulról konkáv voltával pontosabban is be lehet látni). Így a

$$\varnothing(-u_\alpha - d) = \varnothing(-u_\alpha) + z_1,$$

$$1 - \varnothing(u_\alpha - d) = 1 - \varnothing(u_\alpha) - z_2$$

egyenletek összeadásával állításunkat igazoltuk:

$$\varnothing(-u_\alpha - d) + 1 - \varnothing(u_\alpha - d) = \varnothing(-u_\alpha) + 1 - \varnothing(u_\alpha) + z_1 - z_2 > \varnothing(-u_\alpha) +$$

$$+ 1 - \varnothing(u_\alpha) = \alpha.$$

Ha $d > 0$, ugyanígy végigvihető a bizonyítás (az ábra ekkor az ordinátatengelyre szimmetrikus).

Megjegyzés

a) Az, hogy a kritikus tartományba esés valószínűsége a H_0 hipotézis esetén a legkisebb, egyenértékű azzal, hogy *az elfogadási tartományba esés valószínűsége akkor a legnagyobb,* $(1-\alpha)$, *mikor* H_0 *áll fenn.* A próba *torzítatlan* voltát ezzel is megfogalmazhatjuk. *Ezért fogadjuk el* $|u_{\text{számított}}| \leqq u_{\text{táblázatbeli}}$ *esetén inkább* H_0*-t, mint* H_1*-et.*

Ez az eredmény az erőfüggvényről is leolvasható: $1 - E_n(m)$ megadja az elfogadási tartományra esés valószínűségét. Ez a legnagyobb, ha az $m = m_0$ feltétel teljesül, ekkor $P(|u_{\text{számított}}| \leqq u_\alpha) = 1 - \alpha$.

b) Bizonyítható, hogy *a kétmintás u-próba is konzisztens és torzítatlan.*

c) Az, hogy az elfogadási tartományba esés valószínűsége akkor a legnagyobb, mikor H_0 áll fenn, tulajdonképpen a maximum-likelihood módszer kiterjesztése a próbákra. Ott olyan becslést fogadunk el, amely esetén az adott minta bekövetkezési valószínűsége a legnagyobb. Itt olyan hipotézist, amely esetén az adott mintából számított u a legnagyobb valószínűséggel esik az elfogadási tartományra.

4. Szerkesszük meg az 1. gyakorló feladatban szereplő próba erőfüggvényét!

Az $\varepsilon = 0{,}05$-hez tartozó $u_\varepsilon = 1{,}96$ (a $2 \cdot \varnothing(u_\varepsilon) - 1 = 0{,}95$ egyenletből). A másodfajú hiba valószínűsége:

$$\beta = \varnothing\left(\frac{1200-m}{3/\sqrt{16}} + 1{,}96\right) - \varnothing\left(\frac{1200-m}{3/\sqrt{16}} - 1{,}96\right).$$

$m = 1199$ esetén:

$$\beta(1199) = \varnothing\left(\frac{4}{3} + 1{,}96\right) - \varnothing\left(\frac{4}{3} - 1{,}96\right) = \varnothing(3{,}29) - \varnothing(-0{,}63) =$$

$$= 0{,}999\,49 - 1 + 0{,}7357 = 0{,}73519.$$

Annak a valószínűsége, hogy a mintából számolt u érték $m = 1199$ esetén az elfogadási tartományra esik, elég nagy (kb. 74%).

$m = 1198$ esetén:

$$\beta(1198) = \varnothing\left(\frac{8}{3} + 1{,}96\right) - \varnothing\left(\frac{8}{3} - 1{,}96\right) = \varnothing(4{,}63) - \varnothing(0{,}71) =$$
$$= 0{,}238\,897.$$

Most tehát sokkal kisebb (kb. 24%) az elfogadási tartományra esés valószínűsége.

Tudjuk, hogy az $E(m) = 1 - \beta$ erőfüggvény $m = m_0 = 1200$ helyen $\varepsilon = 0{,}05$, és hogy az erőfüggvény az $m = m_0$ egyenesre szimmetrikus. Az erőfüggvény $m = 1200$-tól jobbra-balra rohamosan növekszik, értékei:

$$E(1198) = E(1202) \approx 0{,}76,$$
$$E(1199) = E(1201) \approx 0{,}26,$$
$$E(1200) = 0{,}05 \text{ (itt a legkisebb).}$$

Görbéje a 26. ábrán, ahol

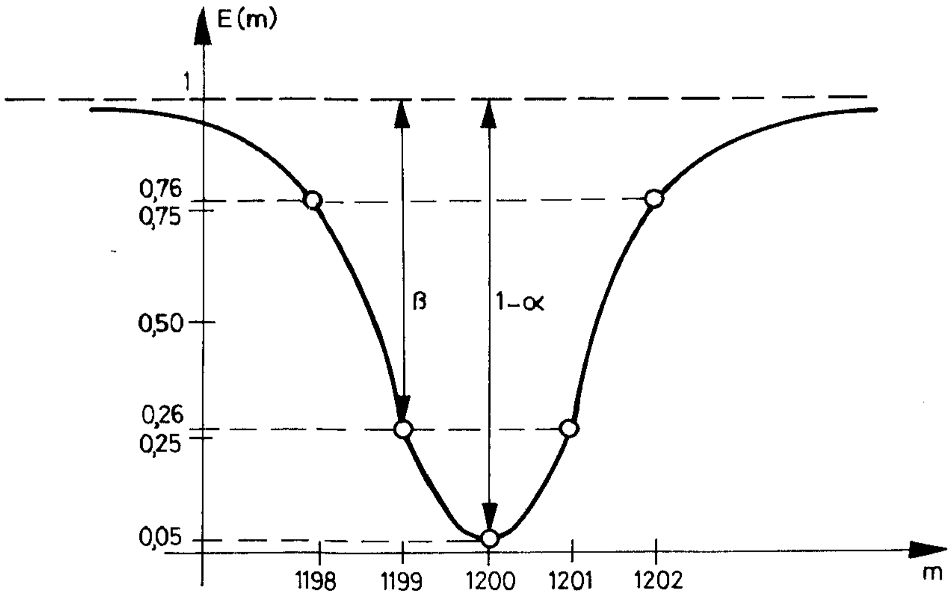
$$\beta < 1 - \alpha,$$

és az $|u_{\text{számított}}| < u_{\text{táblázatbeli}}$ egyenlőtlenség fennállásának a valószínűsége $m = m_0 = 1200$ esetén a legnagyobb.

5. Az előző feladatban a kritikus tartomány,

$$X_k = \{u > u_\varepsilon \quad \text{vagy} \quad u < -u_\varepsilon\},$$

szimmetrikus volt az origóra.



26. ábra

Az 1. gyakorló feladat adataival *keressünk most origóra nemszimmetrikus kritikus tartományokat!* Tegyük fel, hogy 1207 mm átmérőnél nagyobb műszaki okokból már nem kívánatos. 0,9544 valószínűséggel:

$$m_1 - 2\sigma \leqq X \leqq m_1 + 2\sigma,$$

$m_1 + 2\sigma$ tehát maximálisan 1207 lehet, vagyis m_1 maximum 1201 lehet. Legyen a nullhipotézis:

$$H_0\colon\ M(X) = m_0 = 1200,$$

az ellenhipotézis pedig:

$$H_1\colon\ M(X) = m_1 = 1201.$$

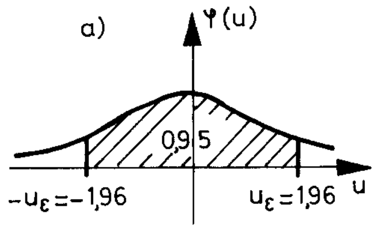
A követelmény az, hogy *azonos* $\varepsilon = 0{,}05$ *elsőfajú hiba mellett olyan próbát válasszunk, amelynél a másodfajú hiba a legkisebb, vagyis az erőfüggvény a legnagyobb* (voltaképpen azt keressük, hogy melyik a *jobb próba* $M(X) = m_1 = 1201$ esetén).

I. Láttuk, hogy a

$$0{,}95 = 1 - \varepsilon = 2\Phi(u_\varepsilon) - 1$$

egyenletből (amely az elsőfajú hiba szintjét szabályozza) a kritikus tartomány:

$$X_k = \{u > 1{,}96 \quad \text{vagy} \quad u < -1{,}96\}$$



27*a* ábra

(27*a* ábra). Láttuk, hogy az erőfüggvény (a másodfajú hiba el nem követésének valószínűsége a H_1 hipotézis teljesülése esetén):

$$E_1(m_1) = 1 - \beta_1 = 1 - P(\overbrace{-u_\varepsilon \leqq u \leqq u_\varepsilon}^{\text{elfogadási tartomány}}/H_1),$$

itt az

$$u = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - m_1}{\sigma/\sqrt{n}} + \underbrace{\frac{m_1 - m_0}{\sigma/\sqrt{n}}}_{d}$$

statisztika $N(d, 1)$-eloszlású $F(u)$ eloszlásfüggvénnyel, tehát

$$E_1(m_1) = 1 - F(u_\varepsilon) + F(-u_\varepsilon) = 1 - \Phi(u_\varepsilon - d) + \Phi(-u_\varepsilon - d) =$$

$$= 1 - \Phi\left(1{,}96 - \frac{1}{3/\sqrt{16}}\right) + \Phi\left(-1{,}96 - \frac{1}{3/\sqrt{16}}\right) =$$

$$= 0{,}264\,81 \approx 26\%,$$

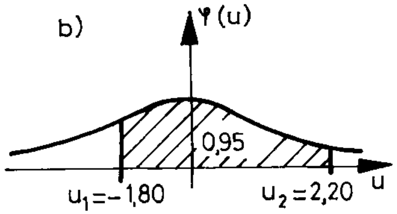
amint azt már kiszámítottuk.

II. Ha most olyan $u_1 = -1{,}80$ és u_2 határokat keresünk, amelyre a

$$P(u_1 \leqq u \leqq u_2) = \Phi(u_2) - \Phi(u_1) = 0{,}95 = 1 - \varepsilon$$

egyenletből u_2-t megkeressük, akkor azt kapjuk, hogy a kritikus tartomány:

$$X_k = \{u < -1{,}80 \quad \text{vagy} \quad u > 2{,}20\}.$$



27*b* ábra

(27*b* ábra). Az erőfüggvény az m_1 helyen:

$$E_2(m_1) = 1 - P(\overbrace{u_1 \leqq u \leqq u_2}^{\text{elfogadási tartomány}}/H_1) = 1 - F(u_2) + F(u_1) =$$

$$= 1 - \Phi(u_2 - d) + \Phi(u_1 - d) =$$

$$= 1 - \Phi\left(2{,}20 - \frac{4}{3}\right) + \Phi\left(-1{,}80 - \frac{4}{3}\right) =$$

$$= 1 - \Phi(0{,}87) + \Phi(-3{,}13) = 1 - 0{,}8078 + 1 - 0{,}99909 =$$

$$= 0{,}193\,31 \approx 19\%,$$

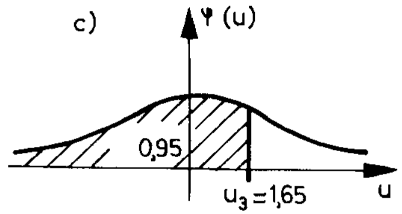
vagyis a próba ereje kisebb, mint előbb volt.

III. Keressünk most egyoldali kritikus tartományt! Ha

$$P(u < u_3) = \Phi(u_3) = 1 - \varepsilon = 0{,}95,$$

akkor ebből $u_3 = 1{,}65$, és így a kritikus (visszautasitási) tartomány:

$$X_k = \{u > 1{,}65\}$$



27*c* ábra

(27*c* ábra). Az erőfüggvény most az m_1 helyen:

$$E_3(m_1) = 1 - P(u \leqq u_3/H_1) = 1 - F(u_3) = 1 - \Phi(u_3 - d) =$$

$$= 1 - \Phi\left(1{,}65 - \frac{4}{3}\right) = 1 - \Phi(0{,}32) = 1 - 0{,}6265 =$$

$$= 0{,}373\,5 \approx 37\%,$$

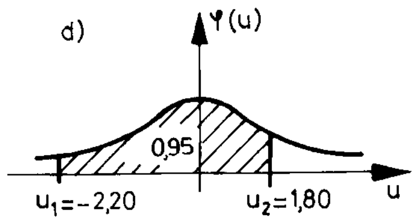
ami az eddigi legjobb próba.

IV. Tükrözzük a II. eset határait az origóra! A

$$P(u_1 = -2{,}20 \leqq u \leqq u_2) = \Phi(u_2) - \Phi(u_1) = 1 - \varepsilon = 0{,}95$$

egyenletből nyilván $u_2 = 1{,}80$, a kritikus tartomány:

$$X_k = \{u < -2{,}20 \quad \text{vagy} \quad u > 1{,}80\}$$



27*d* ábra

(27*d* ábra). Az erőfüggvény az m_1 helyen:

$$E_4(m_1) = 1-P(u_1 \leqq u \leqq u_2/H_1) = 1-F(u_2)+F(u_1) =$$
$$= 1-\varnothing(u_2-d)+\varnothing(u_1-d) =$$
$$= 1-\varnothing\left(1{,}80-\frac{4}{3}\right)+\varnothing\left(-2{,}20-\frac{4}{3}\right) =$$
$$= 1-\varnothing(0{,}47)+\varnothing(-3{,}53) = 1-0{,}6808+1-0{,}999\,79 =$$
$$= 0{,}319\,41 \approx 32\%.$$

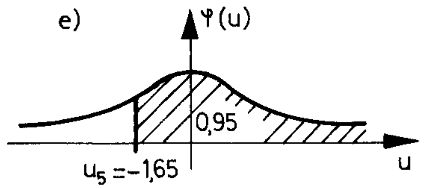
Valamivel gyengébb a próba, mint a III. esetben.

V. Ha most a III. esetben tükrözzük a határokat az origóra, a

$$P(u>u_5) = 1-P(u \leqq u_5/H_1) = 1-\varnothing(u_5) = 1-\varepsilon = 0{,}95$$

egyenletből nyilván $u_5 = -1{,}65$-öt kapunk, és így a kritikus tartomány:

$$X_k = \{u < -1{,}65\}$$



27*e* ábra

(27*e* ábra). Az erőfüggvény az m_1 helyen:

$$E_5(m_1) = 1-P(u \geqq u_5/H_1) = P(u<u_5/H_1) = F(u_5) =$$
$$= \varnothing(u_5-d) = \varnothing\left(-1{,}65-\frac{4}{3}\right) = \varnothing(-2{,}98) =$$
$$= 0{,}0014 \approx 0{,}14\%.$$

Ez az eddigi legrosszabb próba.

A legjobb próba $M(X)=m_1=1201$ esetén a III. alatti, ahol *a „kritikus tartomány valószínűsége":*

H_0 teljesülése esetén $P(X_k/H_0)=\varepsilon=0{,}95$,
H_1 teljesülése esetén $P(X_k/H_1)=0{,}3735$.

legjobb tehát, ha a

$$H_0\colon\ M(X)=m_0=1200$$

nullhipotézis és a

$$H_1\colon\ M(X)=m_1=1201$$

ellenhipotézis esetén így döntünk. Ha

$u_{\text{számított}}$ $\leqq u_3 = 1{,}65$, akkor a H_0 hipotézist elfogadjuk
$u_{\text{számított}}$ $> u_3 = 1{,}65$, akkor a H_0 hipotézist elvetjük,

elvetés esetén H_1-et fogadjuk el.

6. A mintatér olyan X_k és X_k^* kritikus tartománya, amelyek esetén az elsőfajú hiba valószínűsége azonos:

$$P(X_k/H_0) = \varepsilon \quad \text{és} \quad P(X_k^*/H_0) = \varepsilon,$$

megad $1-\varepsilon$-szintű próbát. *Az X_k-próba „jobb", „erősebb", mint az X_k^*-próba,* ha a próba ereje X_k esetén nagyobb, mint X_k^* esetén, azaz ha

$$P(X_k/H_1) > P(X_k^*/H_1).$$

Legyen a nullhipotézis:

$$H_0 : M(X) = m_0,$$

az ellenhipotézis pedig:

$$H_1 : M(X) = m_1 > m_0.$$

Az X_k-próba egyenletesen jobb próba, mint az X_k^* próba, ha minden $m_1 > m_0$-ra nagyobb az ereje, mint a másodiknak. *Az első próba erőfüggvénye tehát minden $m_1 > m_0$ esetén a második próba erőfüggvénye fölött halad.*

A legerősebb próba konstrukcióját adja meg a *Neyman–Pearson-lemma:* A

$$H_0 : P(X<x) = F(x) = \int_{-\infty}^{x} f(t)\,dt$$

nullhipotézissel szemben támasszuk a

$$H_1 : P(X<x) = G(x) = \int_{-\infty}^{x} g(t)\,dt$$

alternatív hipotézist. A nullhipotézis elfogadásához vagy elvetéséhez a legerősebb próba az azonos $1-\varepsilon$-szintű próbák között a következő. H_0-t elutasítjuk, ha $\frac{g(x)}{f(x)} > c$, és elfogadjuk, ha $\frac{g(x)}{f(x)} \leqq c$. Itt a c állandót a

$$P\left(\frac{g(x)}{f(x)} > c/H_0\right) = \varepsilon$$

összefüggésből határozzuk meg. $F(x)$ és $G(x)$ folytonos eloszlásfüggvények.

Mutassuk meg a Neyman–Pearson-lemma segítségével, hogy az előbb kapott (I–V.) próbák között a legerősebb a III. próba, mégpedig minden $m_1>m_0$-ra (tehát az $m_1>m_0$ tartományban egyenletesen a legjobb)!

A $H_0 : M(X) = m_0$ hipotézis egyenértékű a következővel:

$$H_0 : P\left(u = \frac{\bar{x}-m_0}{\sigma/\sqrt{n}} < x\right) = \Phi(x).$$

A $H_1 : M(X) = m_1>m_0$ hipotézis pedig a következővel:

$$H_1 : P\left(u = \frac{\bar{x}-m_0}{\sigma/\sqrt{n}} < x\right) = \Phi\left(x-\overbrace{\frac{m_1-m_0}{\sigma/\sqrt{n}}}^{d}\right).$$

A Neyman–Pearson-lemmában szereplő c állandót a sűrűségfüggvények hányadosából kaphatjuk meg (felhasználjuk, hogy a második eloszlás $N(d;\,1)$-eloszlás, az első pedig $N(0;\,1)$-eloszlás). Ha

$$\frac{g(u)}{f(u)} = \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{(u-d)^2}{2}}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}} > c,$$

akkor H_0-t elutasítjuk, ha pedig $\frac{g(u)}{f(u)} \leqq c$, akkor elfogadjuk. A $\frac{g(u)}{f(u)} > c$ egyenlőtlenség akkor és csak akkor áll fenn, ha

$$e^{\frac{u^2-(u-d)^2}{2}} > c,$$

$$u^2-(u-d)^2 > \ln c^2,$$
$$2ud > \ln c^2+d^2,$$
$$u > \frac{\ln c^2+d^2}{2d}.$$

H_0 esetén ez utóbbi egyenlőtlenség fennállásának a valószínűsége ε, tehát

$$P(u > \frac{\ln c^2+d^2}{2d} = u_\varepsilon/H_0) = \varepsilon,$$
$$P(u \leqq u_\varepsilon/H_0) = 1-\varepsilon,$$
$$\Phi(u_\varepsilon) = 1-\varepsilon.$$

Az utóbbi egyenletből u_ε-t meghatározva, az $u>u_\varepsilon$ esetben H_0-t elutasítjuk, $u\leqq u_\varepsilon$ esetben H_0-t elfogadjuk. Így *a III. alatti próbát kaptuk meg, amely a lemma értelmében* a legerősebb próba minden $m_1>m_0$-ra, tehát *egyenletesen a legjobb próba az $m_1>m_0$ tartományon.*

Ha tehát olyan sejtésünk – műszaki okokból történő kikötésünk – van, hogy az $M(X)=m_0$ nem teljesülése esetén jobb feltenni az $M(X)=m_1>m_0$ hipotézist, akkor ennek a tesztelésére a III. alatti próba a legjobb.

7. Vizsgáljuk meg az előző gyakorló feladatban mutatott II. pont alatti próba erőfüggvényét, $E_2(m)$-et!

A próbához tartozó kritikus tartomány

$$x_k=\{u<u_1 \quad \text{vagy} \quad u>u_2\}$$

valószínűsége H_0 fennállása esetén:

$$P(X_k/H_0) = \varepsilon = 0{,}05,$$

ahol $u_1=-1{,}80$, $u_2=2{,}20$. Az erőfüggvény:

$$E_2(m) = 1-P(u_1\leq u\leq u_2/H_1) = 1-\varnothing\left(u_2-\frac{m-m_0}{\sigma/\sqrt{n}}\right)+$$
$$+\varnothing\left(u_1-\frac{m-m_0}{\sigma/\sqrt{n}}\right).$$

Az erőfüggvény m_0 helyen ugyanannyi, mint az I. próba erőfüggvénye m_0 helyen:

$$E_2(m_0) = 1-P(u_1\leq u\leq u_2/H_0) = 1-(1-\varepsilon) = \varepsilon = 0{,}05.$$

– *Keressük deriválással az erőfüggvény minimumát!*

$$\frac{dE_2(m)}{dm} = +\frac{\sqrt{n}}{\sigma}\varphi\left(u_2-\frac{m-m_0}{\sigma/\sqrt{n}}\right)-\frac{\sqrt{n}}{\sigma}\varphi\left(u_1-\frac{m-m_0}{\sigma/\sqrt{n}}\right) = 0,$$

$$e^{\frac{-\left(u_2-\frac{m-m_0}{\sigma/\sqrt{n}}\right)^2}{2}}-e^{-\frac{\left(u_1-\frac{m-m_0}{\sigma/\sqrt{n}}\right)^2}{2}} = 0,$$

$$\left(u_2-\frac{m-m_0}{\sigma/\sqrt{n}}\right)^2 = \left(u_1-\frac{m-m_0}{\sigma/\sqrt{n}}\right)^2,$$

a) $u_2-\dfrac{m-m_0}{\sigma/\sqrt{n}} = u_1-\dfrac{m-m_0}{\sigma/\sqrt{n}}$,

ez $u_1\neq u_2$ miatt lehetetlen,

b) $u_2-\dfrac{m-m_0}{\sigma/\sqrt{n}} = \dfrac{m-m_0}{\sigma/\sqrt{n}}-u_1$,

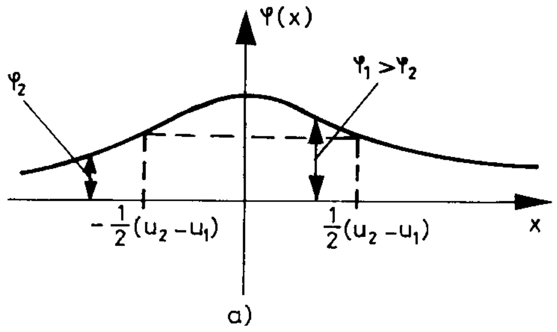
$$m_{\min} = (u_1+u_2)\frac{\sigma}{2\sqrt{n}}+m_0.$$

Itt *lehet* minimum. Igazoljuk, hogy valóban *van* minimum.

c) Ha $m>m_{\min} = (u_1+u_2)\dfrac{\sigma}{2\sqrt{n}}+m_0,$ akkor

$$\frac{m_0-m}{\sigma/\sqrt{n}} < -\frac{1}{2}(u_1+u_2),$$
$$\frac{m_0-m}{\sigma/\sqrt{n}}+u_2 < \frac{1}{2}(u_2-u_1),$$
$$\frac{m_0-m}{\sigma/\sqrt{n}}+u_1 < -\frac{1}{2}(u_2-u_1),$$
$$\frac{dE_2(m)}{dm} = \frac{\sqrt{n}}{\sigma}\left[\underbrace{\varphi\left(u_2+\frac{m_0-m}{\sigma/\sqrt{n}}\right)}_{\varphi_1}-\underbrace{\varphi\left(u_1+\frac{m_0-m}{\sigma/\sqrt{n}}\right)}_{\varphi_2}\right] > 0$$

(l. a 28*a* ábrát; azt is kihasználjuk, hogy $u_2+\dfrac{m_0-m}{\sigma/\sqrt{n}} > u_1+\dfrac{m_0-m}{\sigma/\sqrt{n}}$). Ha a derivált pozitív az $m>m_{\min}$ szakaszon, akkor itt $E_2(m)$ nő.

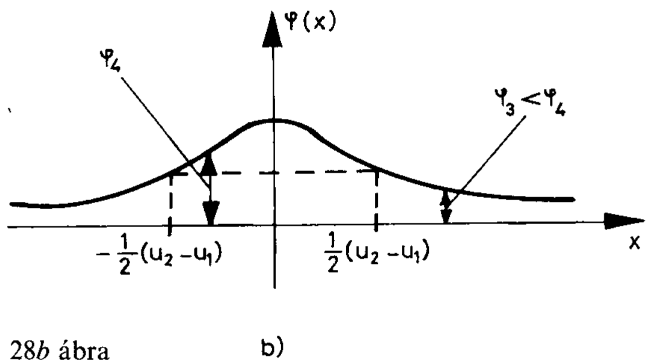


28*a* ábra

d) Ha $m<m_{\min} = (u_1+u_2)\dfrac{\sigma}{2\sqrt{n}}+m_0$, akkor

$$\frac{m_0-m}{\sigma/\sqrt{n}}+u_2 > \frac{1}{2}(u_2-u_1),$$
$$\frac{m_0-m}{\sigma/\sqrt{n}}+u_1 > -\frac{1}{2}(u_2-u_1),$$
$$\frac{dE_2(m)}{dm} = \frac{\sqrt{n}}{\sigma}\left[\underbrace{\varphi\left(u_2+\frac{m_0-m}{\sigma/\sqrt{n}}\right)}_{\varphi_3}-\underbrace{\varphi\left(u_1+\frac{m_0-m}{\sigma/\sqrt{n}}\right)}_{\varphi_4}\right] < 0$$

(l. a 28*b* ábrát; felhasználjuk, hogy $u_2+\frac{m_0-m}{\sigma/\sqrt{n}} > u_1+\frac{m_0-m}{\sigma/\sqrt{n}}$). A derivált negatív az $m<m_{\min}$ szakaszon, így itt $E_2(m)$ fogy.



28*b* ábra

A fogyóból növőbe átmenő $E_2(m)$ erőfüggvénynek $m_{\min}$ abszcisszánál valóban minimuma van. Megjegyezzük, hogy

$$m_{\min} = (u_1+u_2)\frac{\sigma}{2\sqrt{n}}+m_0 > m_0,$$

ha $|u_1| < u_2$ (amint ez számpéldánkban teljesül). Mivel m_0-nál $E_1(m_0) = E_2(m_0) = \varepsilon$, ezért

$$\min E_2(m) < \min E_1(m).$$

– *Igazoljuk, hogy $E_2(m)$ szimmetrikus* az

$$m = m_{\min} = (u_1+u_2)\frac{\sigma}{2\sqrt{n}}+m_0$$

egyenesre!

$$E_2(m) = 1-\Phi\left(\frac{m_0-m}{\sigma/\sqrt{n}}+u_2\right)+\Phi\left(\frac{m_0-m}{\sigma/\sqrt{n}}+u_1\right),$$

$$E_2(m_{\min}-x) = 1-\Phi\left(-\frac{1}{2}(u_1+u_2)+u_2+\frac{x}{\sigma/\sqrt{n}}\right)+$$

$$+\Phi\left(-\frac{1}{2}(u_1+u_2)+u_1+\frac{x}{\sigma/\sqrt{n}}\right) =$$

$$= 1-\Phi\left(\frac{1}{2}(u_2-u_1)+\frac{x}{\sigma/\sqrt{n}}\right)+\Phi\left(\frac{1}{2}(u_1-u_2)+\frac{x}{\sigma/\sqrt{n}}\right),$$

$$E_2(m_{\min}+x) = 1-\Phi\left(\frac{1}{2}(u_2-u_1)-\frac{x}{\sigma/\sqrt{n}}\right)+$$

$$+\Phi\left(\frac{1}{2}(u_1-u_2)-\frac{x}{\sigma/\sqrt{n}}\right) =$$

$$= 1-\left[1-\Phi\left(-\frac{1}{2}(u_2-u_1)+\frac{x}{\sigma/\sqrt{n}}\right)\right]+1-$$

$$-\Phi\left(-\frac{1}{2}(u_1-u_2)+\frac{x}{\sigma/\sqrt{n}}\right) = E_2(m_{\min}-x);$$

a szimmetria valóban fennáll.

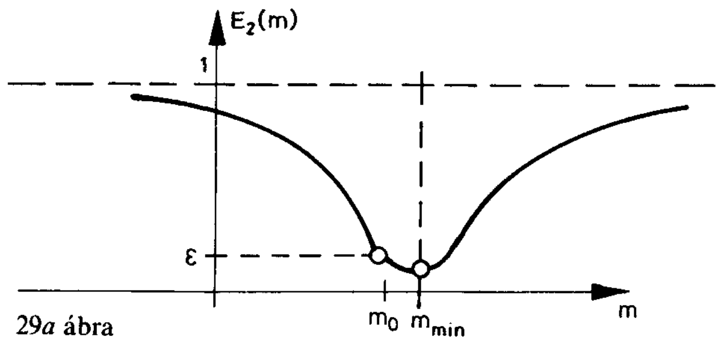
– *Igazoljuk, hogy az erőfüggvény határértéke végtelenben 1:*

$$E_2(m) \to 1, \quad \text{ha} \quad m \to \pm\infty.$$

A szimmetria miatt elég csak $m\to\infty$-re bizonyítani.

$$E_2(m) = 1-\Phi\left(\frac{m_0-m}{\sigma/\sqrt{n}}+u_2\right)+\Phi\left(\frac{m_0-m}{\sigma/\sqrt{n}}+u_1\right)\to$$
$$\to 1-\Phi(-\infty)+\Phi(-\infty) = 1-0+0 = 1.$$

– Az $E_2(m)$ *erőfüggvény görbéje* látható a 29*a* ábrán.



29*a* ábra

8. Az n, σ, m_0 és ε változatlanul hagyásával változtassuk az u-próbát úgy, hogy különböző (u_1, u_2) elfogadási intervallumokat veszünk! Vizsgáljuk az így létrejövő erőfüggvényeket!

– *Először megadjuk az elfogadási tartomány határait úgy,* hogy

$$P(u_1 \leqq u \leqq u_2/H_0) = 1-\varepsilon$$

legyen.

Az erőfüggvény

$$E(m) = 1-P(u_1 \leqq u \leqq u_2/H_1) = 1-\Phi(u_2-d)+\Phi(u_1-d),$$

ahol

$$d = \frac{m-m_0}{\sigma/\sqrt{n}}.$$

Az egyes próbáknál az elfogadási intervallum határára az alábbi értékeket kaptuk (felhasználva, hogy $\Phi(-\infty)=0$, $\Phi(\infty)=1$):

Próba	u_1	u_2
I.	$-1{,}96$	$1{,}96$
II.	$-1{,}80$	$2{,}20$
III.	$-\infty$	$1{,}65$
IV.	$-2{,}20$	$1{,}80$
V.	$-1{,}65$	∞

Az I–II–IV. próbákra az előző gyakorló feladat levezetésének minden pontja érvényes:

$E_i(m_0)=\varepsilon,$ (közös pont),

$m_{\min} = \frac{1}{2}(u_1+u_2)\frac{\sigma}{\sqrt{n}}+m_0$ (szélsőérték),

$E_i(m_{\min}-x) = E_i(m_{\min}+x)$ (szimmetria),

$E_i(m) \to 1$, ha $m \to \pm\infty$ (határérték a végtelenben),

$i = 1, 2, 4.$

A minimum nagysága az I–II–IV. próbáknál (mivel itt $d = \frac{m_{\min}-m_0}{\sigma/\sqrt{n}} = \frac{1}{2}(u_1+u_2)$):

$$\min E_i(m) = 1-\Phi\left(u_2-\frac{1}{2}(u_1+u_2)\right)+\Phi\left(u_1-\frac{1}{2}(u_1+u_2)\right) =$$

$$= 1-\Phi\left(\frac{1}{2}(u_2-u_1)\right)+\underbrace{\Phi\left(\frac{1}{2}(u_1-u_2)\right)}_{1-\Phi\left(\frac{1}{2}(u_2-u_1)\right)} = 2\cdot\Phi\left(\frac{u_1-u_2}{2}\right).$$

– *Igazoljuk, hogy $E_2(m)$ tükörképe az $m=m_0$ egyenesre az $E_4(m)$ függvény!*

$$E_2(m_0+x) = 1-\Phi(u_2-d)+\Phi(u_1-d) =$$

$$= 1-\Phi\left(u_2-\frac{x}{\sigma/\sqrt{n}}\right)+\Phi\left(u_1-\frac{x}{\sigma/\sqrt{n}}\right).$$

A IV. próbánál $u_3=-u_2$ és $u_4=-u_1$, így

$$E_4(m_0-x) = 1-\Phi(u_4-d)+\Phi(u_3-d) = 1-\Phi\left(-u_1+\frac{x}{\sigma/\sqrt{n}}\right)+$$

$$+\Phi\left(-u_2+\frac{x}{\sigma/\sqrt{n}}\right) = 1-\left[1-\Phi\left(u_1-\frac{x}{\sigma/\sqrt{n}}\right)\right]+$$

$$+1-\Phi\left(u_2-\frac{x}{\sigma/\sqrt{n}}\right) = 1-\Phi\left(u_2-\frac{x}{\sigma/\sqrt{n}}\right)+$$

$$+\Phi\left(u_1-\frac{x}{\sigma/\sqrt{n}}\right),$$

$$E_2(m_0+x) = E_4(m_0-x).$$

A II. próba erőfüggvényének tükörképe az $m=m_0$ egyenesre valóban a IV. próba erőfüggvénye.

Az I., II., IV. próbáknál $u_2 \to \infty$ esetén $u_1 \to -1{,}65$,

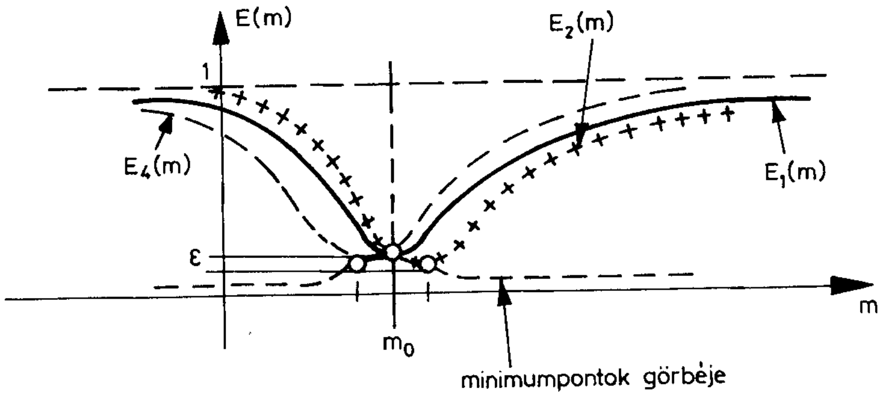
$$\min E_i(m) = 2 \cdot \Phi\left(\frac{u_1-u_2}{2}\right) \to 0.$$

Ha viszont ugyanezen próbáknál $u_1 \to -\infty$, akkor $u_2 \to 1{,}65$,

$$\min E_i(m) = 2 \cdot \Phi\left(\frac{u_1-u_2}{2}\right) \to 0.$$

A 29*b* ábrán vázoltuk az I–II–IV. próbák erőfüggvényeit, a szaggatott görbén sorakoznak a minimumpontok. A minimumok maximuma az $E_1(m)$ görbéhez tartozó minimumpont:

$$\max \min E_i(m) = E_1(m_0) = \varepsilon, \qquad (i=1, 2, 4).$$



29*b* ábra b)

– *Vizsgáljuk most a III. és V. próbák erőfüggvényét!*

A III. próbánál $u_1 = -\infty$, $u_2 = 1{,}65$, az erőfüggvény:

$$E_3(m) = 1 - \Phi(u_2-d) + \Phi(u_1-d) = 1 - \Phi(u_2-d) + \Phi(-\infty) =$$
$$= 1 - \Phi(u_2-d),$$

ahol $d = \dfrac{m-m_0}{\sigma/\sqrt{n}}$.

Könnyen leolvashatók az $E_3(m)$ függvény alábbi tulajdonságai:

$E_3(m_0) = \varepsilon$, mert $E_3(m_0) = 1-\Phi(u_2) = \Phi(-u_2) = \varepsilon$,
$0 < E_3(m) < 1$, mert $0 < \Phi(u_2-d) < 1$,
ha $m \to \infty$, akkor $E_3(m) = 1-\Phi(u_2-d) \to 1-\Phi(-\infty) = 1$,
ha $m \to -\infty$, akkor $E_3(m) = 1-\Phi(u_2-d) \to 1-\Phi(\infty) = 0$.

– *Igazolni fogjuk, hogy $E_3(m)$ mindenütt szigorúan monoton nő.*

$$\frac{dE_3(m)}{dm} = \frac{\sqrt{n}}{\sigma}\varphi\left(u_2 - \frac{m-m_0}{\sigma/\sqrt{n}}\right) > 0$$

minden m-re, tehát $E_3(m)$ minden m-re szigorúan monoton növekedő.

– *Keressük meg $E_3(m)$ inflexiós pontját!*

$$\frac{d^2E_3(m)}{dm^2} = \frac{\sqrt{n}}{\sigma}\frac{1}{\sqrt{2\pi}}e^{-\frac{(u_2-d)^2}{2}}[-(u_2-d)]\left(-\frac{\sqrt{n}}{\sigma}\right) = 0,$$

ha $d = \dfrac{m-m_0}{\sigma/\sqrt{n}} = u_2$, vagyis $m_{\text{infl}} = \dfrac{\sigma}{\sqrt{n}}u_2 + m_0$. Mivel a második derivált előjele $u_2 - d$ előjelétől függ (a többi tényező ugyanis pozitív), ezért a második derivált:

$$>0, \quad \text{ha} \quad u_2 > d, \quad \text{vagyis ha} \quad m < m_{\text{infl}} = \frac{\sigma}{\sqrt{n}}u_2 + m_0,$$

$$<0, \quad \text{ha} \quad u_2 < d, \quad \text{vagyis ha} \quad m > m_{\text{infl}} = \frac{\sigma}{\sqrt{n}}u_2 + m_0.$$

Az m_{infl} abszcissza előtt a függvény alulról domború, utána alulról homorú. Az

$$m_{\text{infl}} = \frac{\sigma}{\sqrt{n}}u_2 + m_0 \qquad (>m_0)$$

helyen tehát valóban inflexiós pont van. Az inflexiós pontbeli függvényérték (mivel most $u_2 - d = 0$)

$$E_3(m_{\text{infl}}) = 1-\Phi(0) = \frac{1}{2}.$$

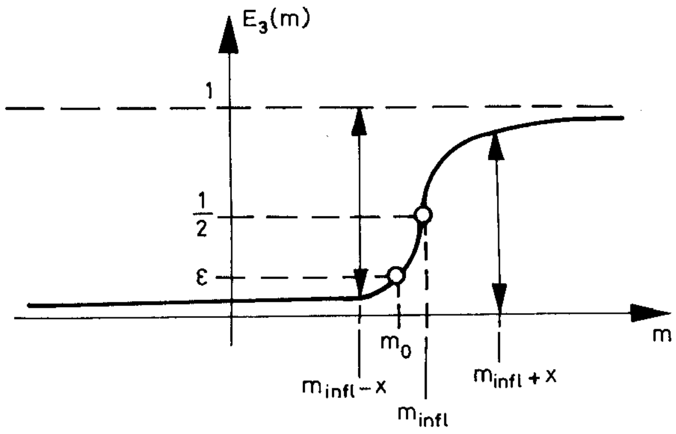
– *Vizsgáljuk meg, szimmetrikus-e az erőfüggvény az inflexiós pontjára!*

$$E_3(m) = 1-\Phi\left(u_2 - \frac{m-m_0}{\sigma/\sqrt{n}}\right),$$

$$E_3(m_{\text{infl}}-x) = 1-\Phi\left(u_2-u_2+\frac{\sqrt{n}x}{\sigma}\right) = 1-\Phi\left(\frac{\sqrt{n}x}{\sigma}\right),$$

$$E_3(m_{\text{infl}}+x) = 1-\Phi\left(-\frac{\sqrt{n}x}{\sigma}\right) = 1-1+\Phi\left(\frac{\sqrt{n}x}{\sigma}\right) = \Phi\left(\frac{\sqrt{n}x}{\sigma}\right).$$

Amennyivel az utolsónak kapott függvényérték eltér a 0-tól (az alsó aszimptotától), ugyanannyival tér el az előtte levő függvényérték az 1-től (a felső aszimptotától). Ez pedig azt bizonyítja, hogy *az $E_3(m)$ erőfüggvény éppen az inflexiós pontjára szimmetrikus* (29*c* ábra).



c)

29*c* ábra

– Az V. próbánál $u_1 = -1{,}65$, $u_2 = \infty$, az erőfüggvény:

$$E_5(m) = 1-\Phi(u_2-d)+\Phi(u_1-d) = 1-1+\Phi(u_1-d) = \Phi(u_1-d),$$

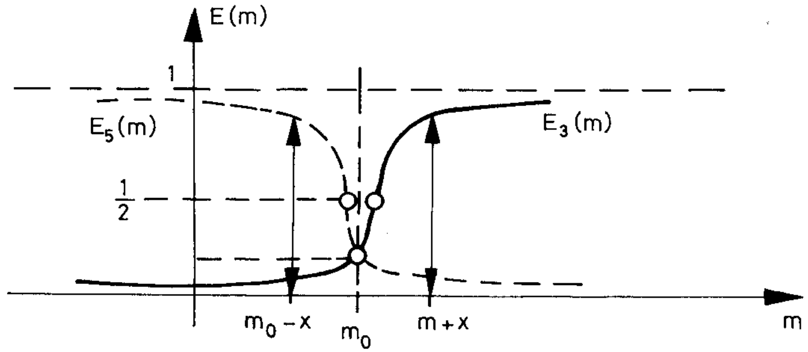
$$E_5(m) = \Phi(-1{,}65-d) = \Phi\left(-1{,}65-\frac{m-m_0}{\sigma/\sqrt{n}}\right),$$

$$E_5(m_0-x) = \Phi\left(-1{,}65+\frac{x}{\sigma/\sqrt{n}}\right).$$

Mivel $E_3(m) = 1-\Phi(1{,}65-d) = 1-\Phi\left(1{,}65-\frac{m-m_0}{\sigma/\sqrt{n}}\right)$,

$$E_3(m_0+x) = 1-\Phi\left(1{,}65-\frac{x}{\sigma/\sqrt{n}}\right) = \Phi\left(-1{,}65+\frac{x}{\sigma/\sqrt{n}}\right) = E_5(m_0-x),$$

ez azt igazolja, hogy $E_5(m)$ *az* $E_3(m)$ *erőfüggvény tükörképe az* $m = m_0$ *egyenesre* (l. a 29*d* ábrát).



29*d* ábra

Megjegyzés

a) Ha $n \to \infty$, az I–II–IV. próbáknál:

$$m_{\min} = (u_1+u_2)\frac{\sigma}{2\sqrt{n}}+m_0 \to m_0,$$

b) ha $n \to \infty$, a III. próbánál:

$$m_{\text{infl}} = \frac{\sigma}{\sqrt{n}}u_2+m_0 \to m_0,$$

ill. az V. próbánál:

$$m_{\inf} = \frac{\sigma}{\sqrt{n}} u_1 + m_0 \to m_0.$$

c) $E_3(m)$ és $E_5(m)$ tetszőleges $1-\varepsilon$-szintű próba esetén szimmetrikusak az $m = m_0$ egyenesre (hiszen mindig érvényes az, hogy az V. próbánál alkalmazott u_1 a III. próbánál alkalmazott u_2 (-1)-szerese).

9. Vizsgáljuk meg az I–V. próbákat *a)* konzisztencia, *b)* torzítatlanság szempontjából! *c)* Mit tudunk mondani az egyes erőfüggvények n-től való függéséről?

---

a) Egy próba konzisztens, ha $n \to \infty$ esetén a második fajta hiba valószínűsége (minden alternativ hipotézis esetén) 0-hoz tart.

Az I. próba konzisztens voltát már igazoltuk. Az I–II–IV. próbánál a másodfajú hiba valószínűsége:

$$\beta = \Phi(u_2 - d) - \Phi(u_1 - d),$$

ahol

$$d = \frac{m - m_0}{\sigma} \sqrt{n}.$$

Ha $n \to \infty$, akkor $d \to \pm\infty$,

$$\lim \beta = \begin{cases} \Phi(\infty) - \Phi(\infty) = 1 - 1 = 0, \\ \Phi(-\infty) - \Phi(-\infty) = 0 - 0 = 0. \end{cases}$$

A III. próbánál a másodfajú hiba valószínűsége $n \to \infty$ esetén:

$$\beta = \Phi(u_2 - d) \begin{cases} \Phi(-\infty) = 0, & \text{ha} \quad m > m_0, \\ \Phi(\infty) = 1, & \text{ha} \quad m < m_0. \end{cases}$$

Az V. próbánál $n \to \infty$ esetén:

$$\beta = 1 - \Phi(u_2 - d) \begin{cases} 1 - \Phi(-\infty) = 0, & \text{ha} \quad m > m_0, \\ 1 - \Phi(\infty) = 1, & \text{ha} \quad m < m_0. \end{cases}$$

Mindez az egyes próbák erőfüggvényéről is leolvasható, hiszen $\beta = 1 - E(m)$. *Az I–II–IV. próbák tehát konzisztensek, a III. és V. próbák nem.* Mégis, ajánlható a III. próba a H_1: $m > m_0$ ellenhipotézis esetén, mint ez esetben egyenletesen legjobb próba, amelynél az $m > m_0$ szakaszon $\beta \to 0$. – Ugyanígy vagy az erőfüggvények $m = m_0$ egyenesre való szimmetriája alapján bizonyítható, hogy *az V. próba a H_1: $m < m_0$ esetben egyenletesen legjobb próba, amelynél az $m < m_0$ esetben a másodfajú hiba valószínűsége is 0-hoz tart.*

b) Egy próba torzítatlan voltát így is megfogalmazhatjuk: a próba torzítatlan, ha az elfogadási tartomány valószínűsége akkor a legnagyobb, mikor H_0 áll fenn (vagyis, ha H_1 áll fenn, akkor nagyobb valószínűséggel utasítsuk el H_0-t, mint ha H_0 állna fenn).

– Számítsuk ki tehát először azt, hogy az I., II. és III. próbáknál milyen m esetén maximális valószínűségű a $P(u_1 \leqq u \leqq u_2)$ valószínűség (= az elfogadási tartomány valószínűsége)!

$$P(u_1 \leqq u \leqq u_2) = \Phi(u_2 - d) - \Phi(u_1 - d),$$

$$\frac{dP}{dm} = -\frac{\sqrt{n}}{\sigma}(\varphi(u_2 - d) - \varphi(u_1 - d)) = 0,$$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{(u_2 - d)^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(u_1 - d)^2}{2}},$$

$$(u_2 - d)^2 = (u_1 - d)^2.$$

Mivel $u_2 - d = u_1 - d$ lehetetlen $(u_1 \neq u_2)$, ezért

$$u_2 - d = -(u_1 - d),$$

$$u_2 - \frac{m - m_0}{\sigma/\sqrt{n}} = -u_1 + \frac{m - m_0}{\sigma/\sqrt{n}},$$

$$m = \frac{1}{2}(u_1 + u_2)\frac{\sigma}{\sqrt{n}} + m_0 = m_{\min}.$$

Itt lehet szélsőérték.

Ha $m > m_{\min} = \frac{1}{2}(u_1 + u_2)\frac{\sigma}{\sqrt{n}} + m_0$, akkor

$$\frac{m - m_0}{\sigma/\sqrt{n}} > \frac{1}{2}(u_1 + u_2),$$

$$u_2 - \frac{m-m_0}{\sigma/\sqrt{n}} < u_2 - \frac{1}{2}(u_1+u_2) = \frac{1}{2}(u_2-u_1),$$

$$u_1 - \frac{m-m_0}{\sigma/\sqrt{n}} < u_1 - \frac{1}{2}(u_1+u_2) = -\frac{1}{2}(u_2-u_1),$$

ugyanakkor

$$u_2 - \frac{m-m_0}{\sigma/\sqrt{n}} > u_1 - \frac{m-m_0}{\sigma/\sqrt{n}},$$

így (l. a 28*b* ábrát):

$$\varphi_1 = \varphi(u_2-d) > \varphi_2 = \varphi(u_1-d),$$

ennek következtében $\frac{dP}{dm} < 0$, vagyis P fogy az $m > m_{\min}$ szakaszon.

Ugyanígy az $m < m_{\min}$ szakaszon:

$$u_2 - d > \frac{1}{2}(u_2-u_1),$$

$$u_1 - d > -\frac{1}{2}(u_2-u_1),$$

de $u_2-d > u_1-d$, és ezért a 28*b* ábra alapján:

$$\varphi_3 = \varphi(u_2-d) < \varphi_4 = \varphi(u_1-d),$$

$\frac{dP}{dm} > 0$, tehát P nő az $m < m_{\min}$ szakaszon. Akkor viszont *az elfogadási tartomány valószínűsége $m_{\min}$-nál maximális. A H_0: $m=m_0$ hipotézis fennállása esetén az I. próba torzítatlan* (ott ugyanis $m_{\min}=m_0$), *a II. és IV. próba nem torzítatlan* (ott $m_{\min} \neq m_0$). Ezért van, hogy a gyakorlatban többször alkalmazzák az I. próbát, mint a II. és IV. próbákat. Utóbbiak akkor lennének torzitatlanok, ha H_0: $m=m_{\min}$ volna.

– Számítsuk ki másodszor azt, hogy a III. és V. próbáknál milyen m esetén maximális a $P(u_1 \leqq u \leqq u_2)$ valószínűség!

A III. próba esetében $u_1 = -\infty$, így

$$P(u_1 \leqq u \leqq u_2) = \Phi(u_2-d),$$

$$\frac{dP}{dm} = -\frac{\sqrt{n}}{\sigma}\varphi\left(u_2 - \frac{m-m_0}{\sigma/\sqrt{n}}\right) < 0$$

minden m-re, P szigorúan monoton fogyó mindenütt, *a III. próba nem torzítatlan.*

Az V. próbánál $u_2 = \infty$, így

$$P(u_1 \leqq u \leqq u_2) = 1 - \Phi(u_1-d),$$

$$\frac{dP}{dm} = \frac{\sqrt{n}}{\sigma}\varphi(u_1-d) > 0$$

minden m-re, P mindenütt szigorúan monoton növő, *az V. próba sem torzítatlan.*

Megjegyzés

Az elfogadási tartomány valószínűségét H_1 fennállása esetén a másodfajú hiba valószínűsége, $\beta = 1 - E(m)$ adja meg. Ezért az előbbi eredmények az erőfüggvények görbéjéről is leolvashatók.

Az I. próba torzítatlan voltát már másképpen is igazoltuk (a $\Phi(x)$ függvény közvetlen vizsgálatával).

c) A 24. ábrán azt igazoltuk, hogy *az I. próbánál a mintaelemszámot növelve, a másodfajú hiba valószínűsége csökken, az erőfüggvény nő:*

$$E_{1,n+1}(m) > E_{1,n}(m),$$

ha $m \neq m_0$ (az m_0 helyen egyenlőek az erőfüggvényértékek).

Az I., II. és IV. próbáknál

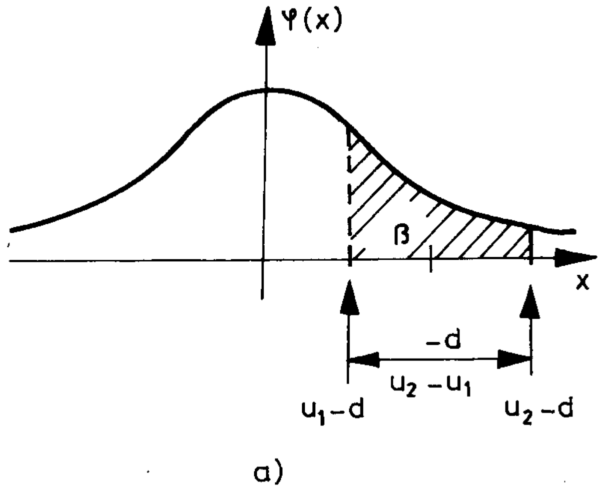
$$\beta = \Phi(u_2-d) - \Phi(u_1-d),$$

itt d az, ami az n-től való függést megadja:

$$d = \frac{m-m_0}{\sigma}\sqrt{n}.$$

Adott $m < m_0$ és adott σ esetén β a 30*a* ábrán látható területtel reprezentálható. Az u_2-u_1 hosszúsághoz tartozó terület n növelésével jobbra csúszik, β kisebbedik, az erőfüggvény nagyobbodik,

$$E_{i,n+1}(m) > E_{i,n}(m) \quad (i = 1, 2, 4).$$



30*a* ábra

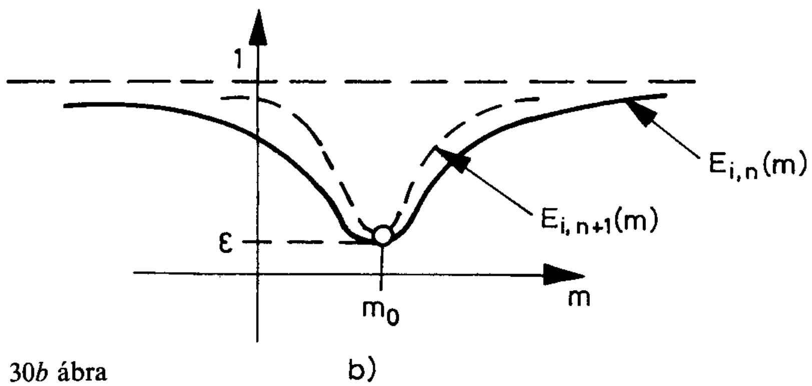
Ha $m > m_0$, akkor $-d$ negatív, a $\varphi(x)$ függvény alatt, $u_2 - u_1$ hosszúsághoz tartozó terület balra csúszik, β csökken, az erőfüggvény nő.

Az $m = m_0$ helyen a két erőfüggvény közös értéke ε.

Az I., II. és IV. próbáknál a mintaelemszám növelése jót tesz a próbának, *növeli a próba erejét:*

$$E_{i,n+1}(m) \geqq E_{i,n}(m) \quad (i = 1, 2, 4),$$

egyenlőség csak az $m = m_0$ helyen van (30*b* ábra).



30*b* ábra

A III. próbánál a másodfajú hiba valószínűsége:

$$\beta = \Phi(u_2 - d),$$

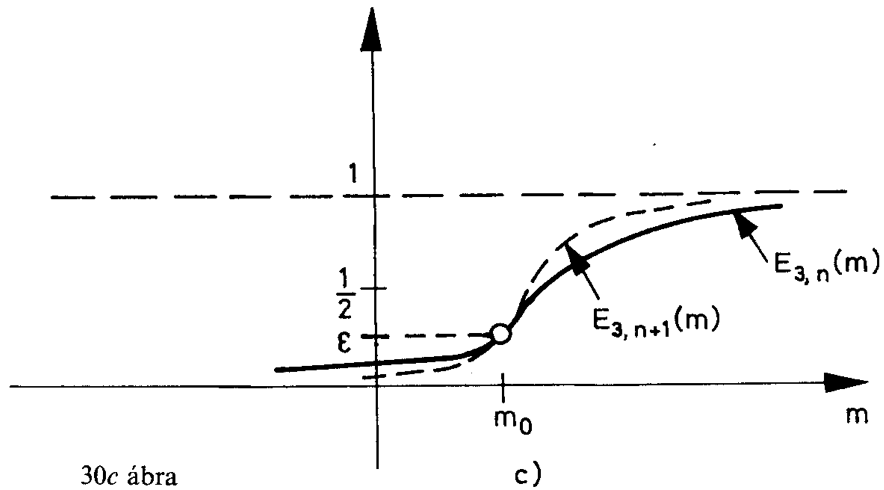
ha $m > m_0$ (d pozitív), n-et $(n+1)$-re növelve $u_2 - d$ csökken, Φ is és β is csökken,

$$E_{3,n+1}(m) > E_{3,n}.$$

Ha viszont $m < m_0$ (d negatív), n-et $(n+1)$-re növelve $u_2 - d$ nő, Φ és β is nő,

$$E_{3,n+1}(m) < E_{3,n}.$$

A III. próbánál n növelése az $m > m_0$ szakaszon az erőfüggvényt növeli, az $m < m_0$ szakaszon csökkenti, $m = m_0$-nál változatlanul hagyja (30*c* ábra).



30*c* ábra

Az V. próba erőfüggvényét a III. próbáéból az $m = m_0$ egyenesre való tükrözéssel kapjuk, tehát *az V. próbánál a mintaelemszám növelése az $m > m_0$ szakaszon az erőfüggvényt csökkenti, az $m < m_0$ szakaszon növeli, az $m = m_0$ helyen változatlanul hagyja.*

10. Vizsgáljuk az I. típusú u-próbánál az erőfüggvény σ szórástól való függését (vagyis tekintsünk különböző, de azonos $1-\varepsilon$-szintű, azonos m_0-val rendelkező, azonos n mérésszámú u-próbákat)!

Tüntessük fel, hogy most az erőfüggvény a szórástól is függ, legyen tehát az erőfüggvény:

$$E(m, \sigma) = 1 - \Phi(u_\varepsilon - d) + \Phi(-u_\varepsilon - d),$$

ahol

$$d = \frac{m - m_0}{\sigma/\sqrt{n}}.$$

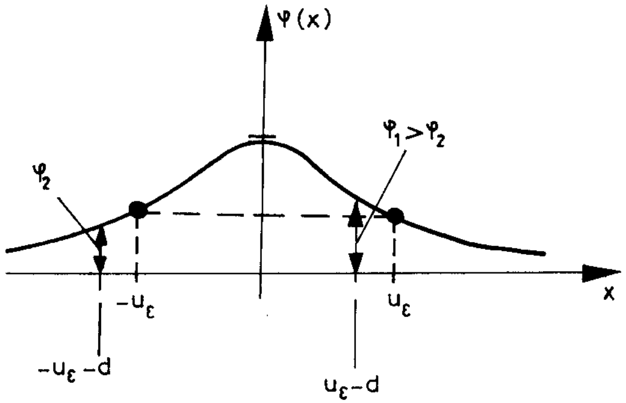
Vizsgáljuk az m = állandó síkmetszeteket! Nézzük $E(m, \sigma)$-t *növés, ill. fogyás szempontjából!*

$$\frac{\partial E}{\partial \sigma} = \frac{1}{\sigma^2}\sqrt{n}(m - m_0)\,[-\varphi(u_\varepsilon - d) + \varphi(-u_\varepsilon - d)],$$

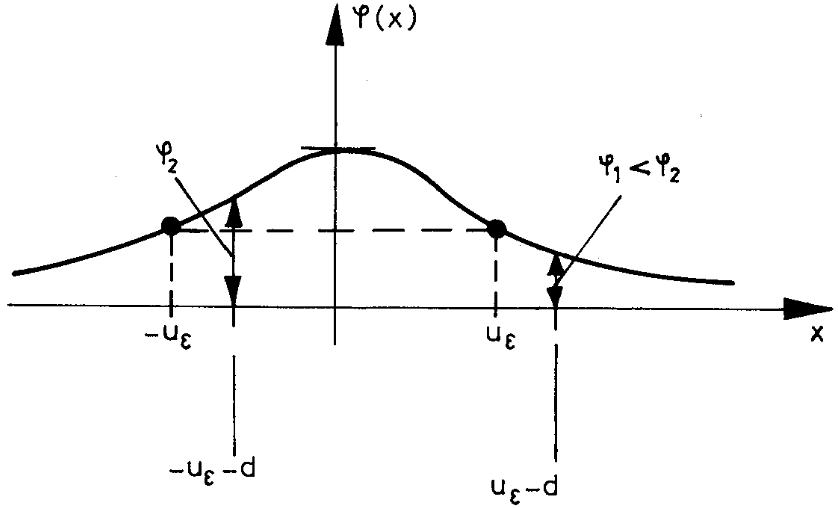
$m > m_0$ esetén $m - m_0 > 0$, $-\varphi(u_\varepsilon - d) + \varphi(-u_\varepsilon - d) = -\varphi_1 + \varphi_2 < 0$ (31. ábra), tehát ez esetben:

$$\frac{\partial E}{\partial \sigma} < 0,$$

az erőfüggvény szigorúan monoton fogyó, $m < m_0$ esetén $m - m_0 < 0$, $-\varphi(u_\varepsilon - d) + \varphi(-u_\varepsilon - d) = -\varphi_1 + \varphi_2 > 0$ (32. ábra), tehát a parciális derivált ez esetben is negatív, az erőfüggvény most is szigorúan monoton fogyó.



31. ábra



32. ábra

Az $m = m_0$ síkmetszetnél a parciális derivált minden σ-ra 0,

$$E(m, \sigma) = 1 - [\Phi(u_\varepsilon) - \Phi(-u_\varepsilon)] = 1 - (1 - \varepsilon) = \varepsilon$$

pedig állandó.

A függvény értelmezési tartománya $\sigma > 0$. Vizsgáljuk *a függvény határértékét az értelmezési tartomány szélein!*

Ha $\sigma \to 0$ és $m \neq m_0$,

$$\lim E(m, \sigma) = 1 - \Phi(\mp\infty) + \Phi(\mp\infty) = \begin{cases} 1 - 0 + 0 = 1, \\ 1 - 1 + 1 = 1. \end{cases}$$

Ha $\sigma \to \infty$ és $m \neq m_0$,

$$\lim E(m, \sigma) = 1 - [\Phi(u_\varepsilon) - \Phi(-u_\varepsilon)] = 1 - (1 - \varepsilon) = \varepsilon.$$

Az $E(m, \sigma)$ függvény tehát 1-től ε-ig süllyed, vagy végig alulról domború, vagy először alulról homorú, aztán domború. Ennek vizsgálata hosszadalmas, ezért itt elhagyjuk.

11. Láttuk, hogy normál eloszlású változó m várható értéke nagy, $1 - \varepsilon$ valószínűséggel esik az

$$\left(\bar{x} - u_\varepsilon \frac{\sigma}{\sqrt{n}};\quad \bar{x} + u_\varepsilon \frac{\sigma}{\sqrt{n}}\right)$$

megbízhatósági intervallumra, u_ε megkeresése az

$$\Phi(u_\varepsilon)-\Phi(-u_\varepsilon) = 2\cdot\Phi(u_\varepsilon)-1 = 1-\varepsilon$$

egyenletből történik.

Vezessük le ebből az u-próbára adott képletet!

$1-\varepsilon$ valószínűséggel állíthatjuk, hogy

$$\bar{x}-u_\varepsilon\frac{\sigma}{\sqrt{n}} \leqq m \leqq \bar{x}+u_\varepsilon\frac{\sigma}{\sqrt{n}}.$$

Az egyenlőtlenségeket átrendezve ugyancsak nagy, $1-\varepsilon$ valószínűségű, hogy

$$-u_\varepsilon \leqq \frac{\bar{x}-m}{\sigma/\sqrt{n}} \leqq u_\varepsilon,$$

vagyis

$$|u_{\text{számított}}| = \left|\frac{\bar{x}-m}{\sigma/\sqrt{n}}\right| \leqq u_\varepsilon = u_{\text{táblázatbeli}}.$$

Ha tehát H_0: $m=m_0$ hipotézis igaz, akkor nagy, $1-\varepsilon$ valószínűségű, hogy

$$|u_{\text{számított}}| = \left|\frac{\bar{x}-m_0}{\sigma/\sqrt{n}}\right| \leqq u_{\text{táblázatbeli}},$$

és csak kis ε valószínűségű, hogy

$$|u_{\text{számított}}| > u_{\text{táblázatbeli}}.$$

Az első esetben a nullhipotézis elfogadása mellett döntünk, a második esetben elvetjük. Ez éppen az u-próba.

Megjegyzés

Minden megbízhatósági intervallumból megadhatunk egy statisztikai próbát. Így például a normál eloszlású változó ismeretlen m valószínűségére (ismeretlen elméleti szórás esetén) adott, $1-\varepsilon$-szintű

$$\bar{x}-t_\varepsilon\frac{s^*}{\sqrt{n}} \leqq m \leqq \bar{x}+t_\varepsilon\frac{s^*}{\sqrt{n}}$$

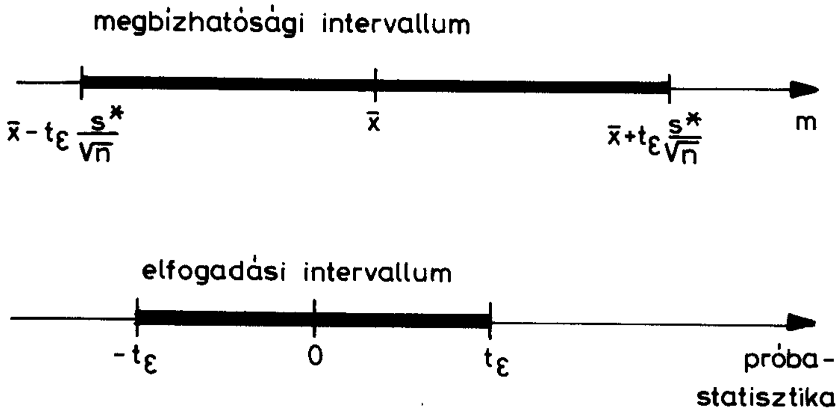
megbízhatósági intervallum alapján következik az egymintás t-próba:

ha $$|t_{\text{szám}}| = \left|\frac{\bar{x}-m_0}{s^*/\sqrt{n}}\right| \leqq t_\varepsilon = t_{\text{tábl}},$$

akkor a H_0: $m=m_0$ hipotézist elfogadjuk, ha

$$|t_{\text{szám}}| > t_{\text{tábl}},$$

akkor a feltevést elvetjük (mert a nullhipotézis alapján alig valószínű, hogy $|t_{\text{szám}}| > t_{\text{tábl}}$ legyen) (33. ábra).



33. ábra

12. Laboratóriumi mérleghez kétféle tárasúlyt azonos súlyúra kell méretezni. Az egyikre végzett $n=9$ független mérés során $\bar{x}=0{,}1672$ (század N), a másikra $m=16$ független mérésből $\bar{y}=0{,}1683$ (század N) mintaátlagot kaptunk. A mérési eredmények igen jó közelítésben normál eloszlást mutatnak. A mérőeszköz szórása (egyúttal a két sokaság elméleti szórása) $\sigma=0{,}0012$. $1-\varepsilon = 0{,}95$ megbízhatósági szinten elfogadható-e, hogy a két sokaságban a várható érték megegyezik?

Kétmintás u-próbával dönthetünk. Az aktuális (a mintából számított) u érték:

$$u = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} = \frac{0{,}1672-0{,}1683}{\sqrt{\frac{0{,}0012^2}{9}+\frac{0{,}0012^2}{16}}} = -2{,}2.$$

A 8. táblázat végtelennek megfelelő utolsó sorából 95% biztonsági szinthez $u_{\text{táblázatbeli}} = 1{,}96$ tartozik.

Mivel a számított u a $(-1{,}96;\ 1{,}96)$ elfogadási tartományon kívül esik, a kétféle tárasúly-sokaság várható értékének azonosságát elutasítjuk.

13. Árammérőket úgy igazítanak be, hogy a mérőket együtt működtetik egy standard árammérővel. Beigazítás után $n=10$ árammérőt választunk ki, és precíziós wattmérő, stoppermérő segítségével mérjük a szóban forgó árammérő egy jellemző paraméterét. A standard árammérő paramétere 1, ettől felfelé-lefelé eltéréseket tapasztalunk. A kapott értékek:

0,895; 1,003; 0,996; 0,994; 1,002;
0,987; 0,993; 0,991; 1,004; 0,985.

Kérdés: véletlen-e az eltérés vagy szisztematikus?

A minta átlaga: $\bar{x}=0{,}994$, a korrigált tapasztalati szórás: $s^*=0{,}007\,225\,8$.

Mivel most csak a minta szórása áll rendelkezésre, t-próbával dönthetünk.

$$\left|\sqrt{n}\,\frac{\bar{x}-m_0}{s^*}\right| = \left|\sqrt{10}\,\frac{0{,}994-1}{0{,}007\,225\,8}\right| = |-2{,}63| = 2{,}63.$$

Hasonlítsuk össze a számított értéket a Student-eloszlás $n=10$-hez ($f=9$ szabadságfokhoz), 95% biztonsági szinthez tartozó $t_{95}=2{,}262$ értékével! $2{,}63 > t_{95} = 2{,}262$, tehát 95%-os biztonsági szinten a tapasztalt eltérés szignifikáns.

De ha a 99%-os szinthez tartozó táblázatbeli értékkel hasonlítunk össze, akkor nincs szignifikáns eltérés, az eltérés nem szisztematikus, csak véletlen okozza:

$$2{,}63 < t_{99} = 3{,}250.$$

Így a válasz: lehet, hogy véletlen eltérés, de lehet, hogy nem; a kis mintaelemszámot is figyelembe véve további méréseket kell végeznünk a végső döntéshez.

A fentieket figyelembe véve a gyakorlatban a 95, a 99 és a 99,9% biztonsági szintekhez tartozó t_{95}, t_{99} és $t_{99,9}$ szorzószámokkal szokták összehasonlítani a $|\sqrt{n}\,(\bar{x}-m_0)/s^*|$ értéket.

	Ha a fenti abszolút érték			
	kisebb t_{95}-nél	t_{95} és t_{99} közé esik	nagyobb t_{99}-nél	nagyobb $t_{99,9}$-nél
Akkor a várható érték $= m_0$ feltevés elbírálása	$\bar{x}$ és m_0 eltérése véletlen, a feltevés elfogadható	további vizsgálat szükséges	$\bar{x}$ és m_0 eltérése nem véletlen, a feltevés nem fogadható el	$\bar{x}$ és m_0 eltérése fokozottan nem véletlen, a feltevés fokozottan nem fogadható el

---

14. Adjunk programot a szükséges próbastatisztika kiszámítására HT-PTK-1050-es zsebszámológépre *a)* egymintás t-próbához, *b)* kétmintás t-próbához!

Emlékeztetünk rá, hogy a gépen az adatbevitel és a szükséges értékek kiszámítása így történik:

x_1 2nd Σ^+
x_2 2nd Σ^+
– – – – – –
x_n 2nd Σ^+,

ezután a 2nd $\bar{x}$ billentyűzés hatására megjelenik a mintaátlag, 2nd σ^2 hatására pedig a korrigálatlan szórás négyzete. A tárak és tartalmuk:

Tár	0	1	2	6	7
Tartalom	n	Σx_i	Σx_i^2	0	x_{n+1}

a) A t próbastatisztikát először átalakítjuk úgy, hogy benne a korrigálatlan σ_n szórás szerepeljen:

$$|t| = \left|\sqrt{n}\,\frac{\bar{x}-m_0}{s^*}\right| = \left|\sqrt{n-1}\,\frac{\bar{x}-m_0}{\sigma_n}\right|.$$

Először m_0-t bevisszük a számítás során üresen maradó 6. tárba. A programot itt megállítjuk, és bevisszük az $x_1, x_2, \ldots, x_n$ adatokat. Ezek után kiszámíttatjuk a szükséges $|t|$ értéket. A program az LRN utasítás után:

STO 6 R/S 2nd $\bar{x}$ − RCL 6 = ÷ 2nd σ^2 $\sqrt{x}$=
x (RCL Ø − 1) $\sqrt{x}$ = 2nd |x| R/S RST.

Példa. m_0 = 3,2, x_i = 1, 2, 2, 3, 3, 3, 3, 4, 4, 5. Az m_0 és az $n=10$ adat bevitele:

3,2 R/S
1 2nd Σ^+
2 2nd Σ^+ 2nd Σ^+
3 2nd Σ^+ 2nd Σ^+ 2nd Σ^+ 2nd Σ^+
4 2nd Σ^+ 2nd Σ^+
5 2nd Σ^+

Újabb R/S után a program végigfut, eredmény: $|t|$ = 0,547 722 6 $< t_{95} = 2{,}262$, $m = m_0$ (= 3,2) elfogadható.

Ha most új m_0-t akarunk kipróbálni, a többit nem is kell billentyűzni. Például $m_0 = 3{,}4$, 3,5 vagy 3,8 elfogadható, $m_0 = 3{,}9$ vagy 4 el is fogadható, el is vethető, $m_0 = 4{,}8$ 99,9%-os szinten is elvetendő.

b) A kétmintás t-próba az adatbeviteltől kezdve 50 lépésnél hosszabb programhoz vezet, ami megengedhetetlen. Ezért az adatokból először billentyűzés-

sel (a beépített program segítségével) kiszámítjuk a mintaátlagokat, a mintabeli korrigált szórásokat. Ezután a 0–1–2–3–4–5 tárakat megtöltjük:

n_1 STO 0 n_2 STO 1 $\bar{x}_1$ STO 2 $\bar{x}_2$ STO 3 s_1^* x^2 STO 4 s_2^* x^2 STO 5.

A 34. ábráról leolvashatjuk, hogy mit melyik tárba tároltunk, áttekinthetjük a programozási lépéseket. *A program:*

RCL Ø − 1 = x RCL 4 = STO 7 RCL 1 − 1 = x RCL 5 = SUM 7 RCL Ø + RCL 1 − 2 = INV 2nd Prd 7 RCL Ø $\frac{1}{x}$ + RCL 1 $\frac{1}{x}$ = 2nd Prd 7 RCL 7 $\sqrt{x}$ STO 7 RCL 2 − RCL 3 = 2nd |x| INV 2nd Prd 7 RCL 7 $\frac{1}{x}$ R/S RST.

Elejére tehető még az F-próbastatisztika kiszámítása is: RCL 4 ÷ RCL 5 = R/S. Ha az így kapott érték < 1, akkor $\frac{1}{x}$ billentyűzéssel a reciprokát vesszük. Ha > 1, akkor megkaptuk F értékét.

Példa. $n_1 = 20$, $n_2 = 49$, $\bar{x}_1 = 171{,}660$, $\bar{x}_2 = 169{,}224$, $s_1^* = 3{,}2847$, $s_2^* = 4{,}6385$ (az adatok eredetét illetően l. a következő gyakorló feladatot).

Eredmény: $F = \dfrac{s_2^{*2}}{s_1^{*2}}$ = 1,594 177 $< F_{95}$, tehát a két sokaságban elfogadható a szórások egyezése, $|t|$ = 2,083 320 6 $> t_{95}$, de $< t_{98}$, $m_1 = m_2$ el is vethető, el is fogadható.

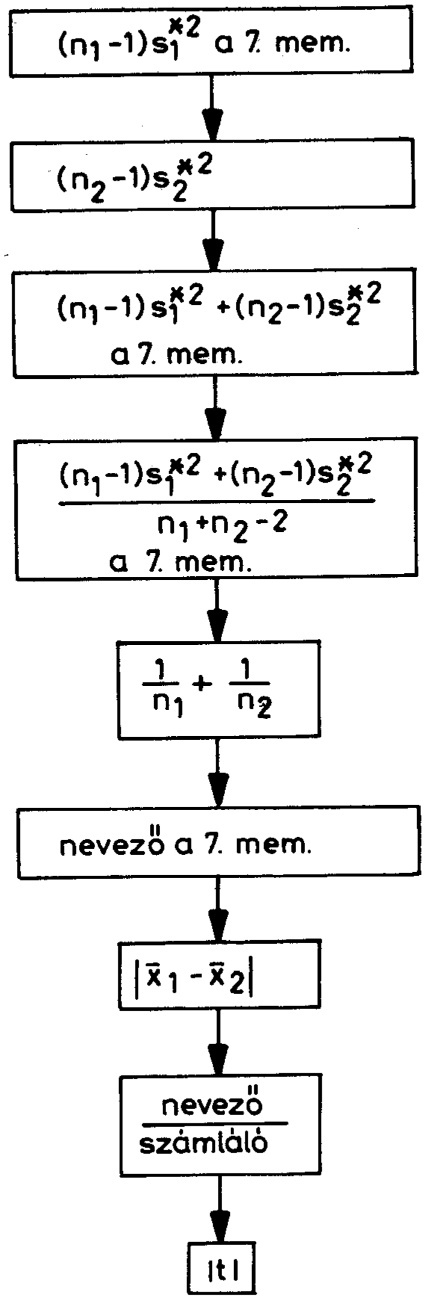
15. Végezzük el az alábbi minőségellenőrző vizsgálatot, adjunk közben *folyamatábrát és programot a szükséges egy- és kétmintás t-próbákhoz, F-próbához!*

A motor fajlagos fogyasztási vizsgálatának értékelését számítógéppel végeztük. Motorok fajlagos fogyasztását vizsgálták főjavítás után.

Az energiatakarékosság szempontjából a motorok fajlagos fogyasztásának ellenőrzése fontos tényező. Minél kevesebb benzin felhasználásával végez ugyanannyi munkát a motor, annál takarékosabb. A fajlagos fogyasztás mértékegysége gramm/735,5 wattóra = gramm/735,5 · 3,6 · 10^3 joule.

Az azonos típusú motorok két műhelyben végzett javításából először az I. műhely nagyjavítási munkáját értékeljük.

a) Az I. műhely munkájának ellenőrzése; egymintás t-próba. Motorfőjavításnál azt tapasztalták, hogy a gramm/735,3 wattórában mért fajlagos fogyasztás normális eloszlású. Az $n = 20$-elemű mintában mért x_i értékeket számítógépre vitték. A mintaátlag 171,60. Az előírás az, hogy az egész sokaságban a várható



34. ábra

érték $m_0=170$ kell, hogy legyen. Véletlen okozza-e az eltérést, feltehető-e, hogy a nagyjavítás megfelelő volt ($m_0=170$ feltehető-e)?

A TPA/i számítógépen rendelkezésre álló program bemenetként kéri az n, az m_0 és az x_i értékeket, és kimenetként megadja a mintaátlagot, a tapasztalati korrigált szórást, az ezekből számított t értéket.

INPUT

N=20 MO=170,00

X(I)=172.00
180.00
176.00
169.00
169.00
173.00
170.00
168.00
169.00
170.00
173.00
175.00
174.00
172.00
176.00
168.00
169.00
171.00
168.00
170.00

OUTPUT

TAPASZTALATI AATLAG
M=171.00
TAPASZTALATI
KORRIGAALT
SZOORAAS

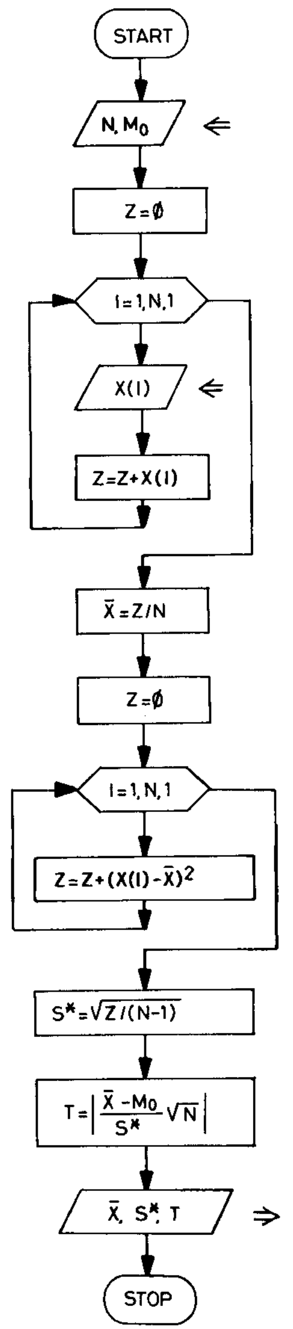
S=3.285

T SZAAMITOTT

T=2.180

A Student-eloszlás táblázatát használva:

n	Statisztikai biztonság	
	95%	99%
20	2,093	2,861



35*a* ábra a)

A számított t abszolút értéke 2,180; ez nagyobb, mint t_{95} és kisebb, mint t_{99}. Lehet, hogy véletlen az eltérés, lehet, hogy nem, kicsi a mintaelemszám, további méréseket kell végezni a végső döntéshez.

A számítógépes program folyamatábráját közöljük a 35*a* ábrán.

b) További vizsgálat az egymintás t-próbával; intervallumbecslés a szabványtól való eltérésre. Az egymintás t-próbát az előzőekben vázolt számítás után sokszor be szokták fejezni. Pedig érdemes a számításokat kiterjeszteni az alábbiakra.

Az $n=20$ mérés eredményéből azt kaptuk, hogy a mintaátlag $\bar{x}=171{,}6$, a tapasztalati korrigált szórás $s^*=3{,}285$. Tegyük fel, hogy az egész sokaságban a várható érték $170+k$. Milyen becslést tudunk adni k-ra (a szabványelőírástól való eltérésre)?

Tudjuk, hogy ha a várható érték $170+k$, akkor 95% biztonsági szinten:

$$\left|\sqrt{n}\,\frac{\bar{x}-m}{s^*}\right| \leqq t_{95},$$

az adatokat behelyettesítve:

$$\left|\sqrt{20}\,\frac{171{,}6-170-k}{3{,}285}\right| \leqq 2{,}093,$$

$$|1{,}6-k| \leqq 1{,}537,$$

$$0{,}063 \leqq k \leqq 3{,}137.$$

Így az egész sokaságban a várható érték a szabványelőírástól minimálisan 0,063-del, maximálisan 3,137-del térhet el felfelé (95% biztonsági szinten).

Az egymintás t-próba BASIC programja:

```
1000 REM * EGYMINTAS T PROBA *
1010 INPUT "N="; N
1020 INPUT "M0="; M0
1025 DIM X(N)
1030 Z = 0
1040 FOR I=1 TO N
1050 PRINT "X(";I;")=": INPUT X(I)
1060 Z = Z + X(I)
1070 NEXT I
1080 XA = Z/N: Z = 0
1090 FOR I=1 TO N
1100 Z = Z + (X(I)-XA)↑2
1110 NEXT I
1120 SC = SQR(Z/(N-1))
1130 T = ABS ((XA-M0)*SQR(N)/SC)
1140 PRINT "ATLAG="; XA, "SZORAS="; SC, "T ERTEK="; T: END
```

c) Az I. és II. műhely munkájának ellenőrzése; kétmintás F-próba és t-próba. Az előző motornagyjavítást két műhelyben végezték. Utána mérték a fajlagos fogyasztást (ami korábbi mérések során normális eloszlásúnak mutatkozott). Az első műhelyben végzett $n_1=20$ mérés eredményét, a második műhelyben végzett $n_2=49$ mérés eredményét mutatja az alábbi, TPA/i számítógépen készült lista: az első négy sorban az I. műhelyben mért értékek, a következő 10 sorban a II. műhelyben mért értékek láthatók.

172.0000	176.0000	180.0000	169.0000	169.0000
173.0000	170.0000	168.0000	169.0000	170.0000
173.0000	175.0000	174.0000	172.0000	176.0000
168.0000	169.0000	171.0000	168.0000	170.0000

160.0000	160.0000	161.0000	162.0000	163.0000
163.0000	164.0000	164.0000	165.0000	165.0000
165.0000	166.0000	166.0000	167.0000	167.0000
167.0000	168.0000	168.0000	168.0000	168.0000
169.0000	169.0000	169.0000	169.0000	169.0000
169.0000	170.0000	170.0000	170.0000	170.0000
171.0000	171.0000	171.0000	171.0000	171.0000
172.0000	172.0000	172.0000	173.0000	173.0000
174.0000	174.0000	175.0000	175.0000	175.0000
176.0000	177.0000	178.0000	178.0000	

```
AZ N1 ELEMUE MINTA AATLAGA                       XA=171.6000
AZ N2 ELEMUE MINTA AATLAGA                       YA=169.2240
AZ N1 ELEMUE MINTA KORRIGAALT SZOORAASA          S1=  3.2847
AZ N2 ELEMUE MINTA KORRIGAALT SZOORAASA          S2=  4.6385
AZ F EERTEEK                                      F=  1.9941
A STUDENT-FEELE T EERTEEK                         T=  2.0829
```

A számítógép kiszámította az I. és a II. műhelybeli minta átlagát és szórását:

$$\bar{x}=171{,}6000; \quad \bar{y}=169{,}2240;$$
$$s_1^*=\ \ 3{,}2847; \quad s_2^*=\ \ 4{,}6385.$$

Az átlagok alapján úgy tűnik, hogy a II. műhely munkája jobb (kisebb a motor fajlagos fogyasztása). A szórásoknál fordítva van: $s_2^*>s_1^*$, ez egy kicsit ellentmond az előző megállapításnak.

A kérdés tehát a következő:

I. Két független, normális eloszlású változó tapasztalati szórása kissé eltér. Feltehető-e, hogy az egész sokaságban megegyezik a két elméleti szórás, $\sigma_1=\sigma_2$?

Osszuk el a nagyobbik szórásnégyzetet a kisebbel! Az így kapott $F=\dfrac{s_2^{*2}}{s_1^{*2}}$ számlálójáról bizonyítható, hogy n_2-1 szabadságfokú, χ^2-eloszlású változó; nevezőjéről igazolható, hogy n_1-1 szabadságfokú, χ^2-eloszlású. A két változó független. $\sigma_1=\sigma_2$ esetén az $F=\dfrac{s_2^{*2}}{s_1^{*2}}$ n_2-1, n_1-1 szabadságfokú, F-eloszlású.

Az F-eloszlásból megállapítható két olyan határ, amelybe a változó nagy (95%-os) valószínűséggel esik. Az egyik határ 1-nél kisebb, a másik határ 1-nél nagyobb. A $\sigma_1=\sigma_2$ hipotézis megvizsgálására elegendő csak azt vizsgálni, hogy az 1-nél nem kisebb $F=\dfrac{s_2^{*2}}{s_1^{*2}}$ a felső határ alatt van-e.

Két független, normális eloszlású változó szórásának az egyezését F-próbával dönthetjük el: osszuk el a nagyobbik tapasztalati korrigált szórásnégyzetet a kisebbikkel, az $F=\dfrac{s_2^{*2}}{s_1^{*2}}$ így számított értékét hasonlítsuk össze az előbbi táblázatbeli F-eloszlás értékével.

Ha $F_{\text{számított}}$ $\leqq F_{95}$ (táblázatbeli érték), akkor elfogadjuk
Ha $F_{\text{számított}}$ $> F_{95}$ (táblázatbeli érték), akkor elvetjük

a két szórás egyezését (a $\sigma_1=\sigma_2$ feltevést).

A táblázat értékei csak 95% biztonsági szinthez vannak kiszámítva.

II. Ha $\sigma_1=\sigma_2$ feltehető, akkor feltehető-e a két független, normál eloszlású sokaságban, hogy a várható értékek megegyeznek, $m_1=m_2$?

A várható értékek egyezésére (az $m_1=m_2$ feltevésre) az alábbi próbát tehetjük (feltéve, hogy $\sigma_1=\sigma_2$):

Kiszámítjuk a

$$|t|=\frac{|\bar{x}_1-\bar{x}_2|}{\sqrt{\dfrac{(n_1-1)s_1^{*2}+(n_2-1)s_2^{*2}}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}.$$

statisztikai függvény (röviden: statisztika) feladatbeli (aktuális) értékét. Bizonyítható, hogy ez a statisztika Student-eloszlású, n_1+n_2-2 szabadságfokkal. Ezt összehasonlítjuk a Student-eloszlás n_1+n_2-2 szabadságfokú, adott biztonsági szinthez tartozó értékével.

Ha $|t_{\text{számított}}|$ $\leqq t_{\text{táblázatbeli}}$ az $m_1=m_2$ feltevést elfogadjuk
Ha $|t_{\text{számított}}|$ $> t_{\text{táblázatbeli}}$ az $m_1=m_2$ feltevést elvetjük

az adott biztonsági szinten.

Esetünkben a számítógép kiszámította az $F=1{,}9941$ értéket. Hasonlítsuk ezt össze a táblázatbeli F értékkel (a nagyobbik szórásnégyzetű számláló szabadságfoka 48, a nevezőé 19)!

f_2	f_1	
	30	50
19	2,07	2,0

Mivel

$$F_{\text{számított}}=1{,}9941 < F_{95}=2{,}0035,$$

ezért 95%-os biztonsági szinten elfogadjuk, hogy a két sokaságban a szórások megegyeznek (a mintabeli szórások különbsége ellenére).

Ha ez utóbbi teljesül, akkor vizsgálhatjuk, hogy egyeznek-e a várható értékek is.

$$t_{\text{számított}} = 2{,}0829,$$

hasonlítsuk ezt össze a Student-eloszlás $f = n_1+n_2-2 = 20+49-2 = 67$ szabadságfokú t értékével ($n = f+1 = 68$-nak megfelelő sorban kell keresni)!

$n = f+1$	Statisztikai biztonság	
	95%	98%
61	2,000	2,390
121	1,980	2,358

Látszik, hogy $|t_{\text{számított}}| < t_{98}$, de $|t_{\text{számított}}| > t_{95}$. A feltevés elfogadása mellett is, ellene is lehet dönteni. Dönthetünk úgy is, hogy további méréseket kérünk.

A szükséges F- és kétmintás t-próba elvégzéséhez a számítógépes program blokkdiagramját közöljük a 35*b* ábrán.

d) További vizsgálat a kétmintás t-próbánál; intervallumbecslés a két várható érték eltérésére. Az elméleti szórások egyezését tételezzük fel. A kétmintás t-próbát ezek után gyakran be szokták fejezni, az előbb közölt eredménnyel, bár a vizsgálat tovább folytatható.

Tegyük fel, hogy a két sokaság várható értéke között k különbség van:

$$m_1 = m_2 + k.$$

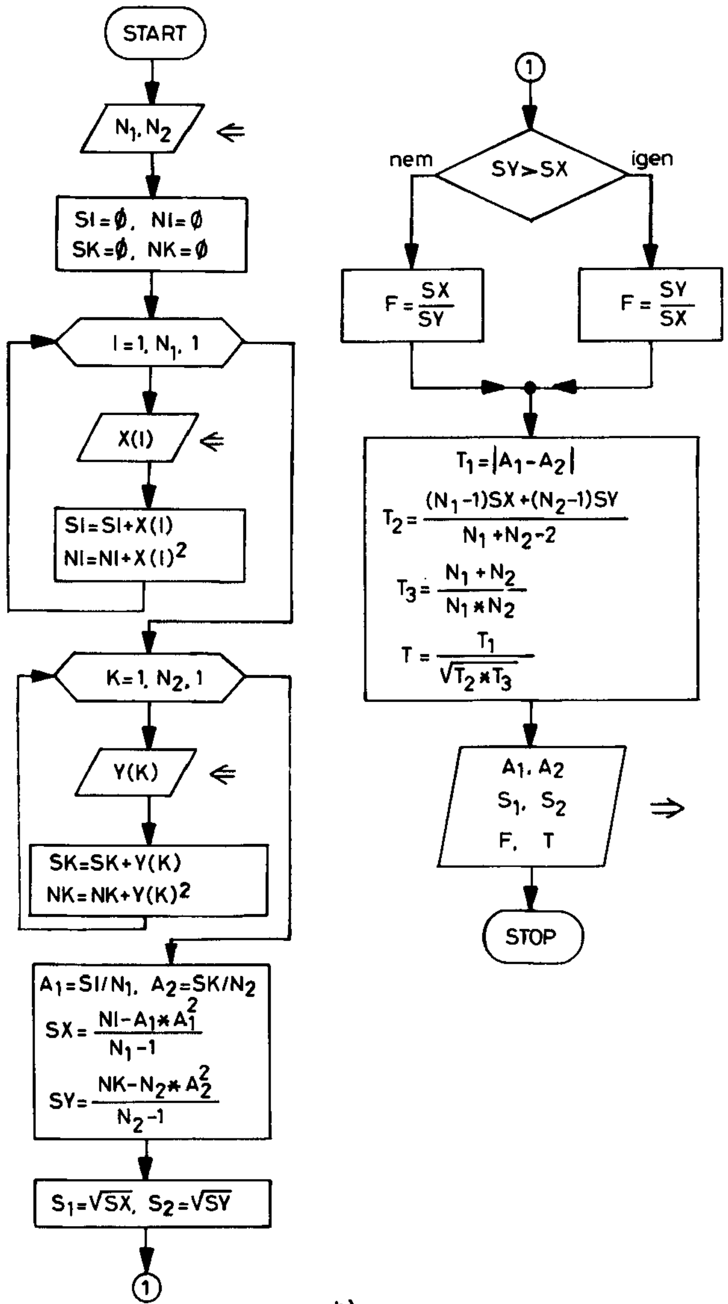
95% biztonsági szinten milyen becslést (megbízhatósági intervallumot) tudunk megadni k-ra?

Itt voltaképpen az előző $m_1 - m_2 = 0$ feltevés helyett

$$(m_1-k)-m_2 = 0$$

feltevéssel élünk. Így az ugyancsak normál eloszlású $X-k$ és Y változók várható értékének egyezésére végezhetünk kétmintás t-próbát. Felhasználva, hogy $X-k$ várható értéke $E(X-k) = E(X)-k$, k-ra az alábbi egyenlőtlenséget kapjuk:

$$|\bar{x}-k-\bar{y}| \leq t_{95}\sqrt{\frac{(n_1-1)s_1^{*2}+(n_2-1)s_2^{*2}}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$



b)

35*b* ábra

Az előző fejezet mérési adatait felhasználva:

$$|2{,}376-k| \leqq 1{,}998 \cdot \sqrt{\frac{19 \cdot 3{,}2847^2+48 \cdot 4{,}6385^2}{20+49-2}} \sqrt{\frac{1}{20}+\frac{1}{49}} = 2{,}279,$$

ami azt jelenti, hogy k-ra 95%-os biztonsági szinten az alábbi megbízhatósági intervallumot nyerjük:

$$0{,}098 \leqq k \leqq 4{,}655.$$

A minták alapján tehát a következőt látjuk. Az I. műhelyben javított motorok fajlagos fogyasztásának várható értéke a II. műhelyben javított motorok fajlagos fogyasztásának várható értékénél minimálisan 0,098-del, maximálisan 4,655-del nagyobb lehet (95% biztonsági szinten).

e) Újabb vizsgálat a két műhely munkájának ellenőrzésére. Mivel az előző méréssorozat értékelésekor azt kaptuk, hogy a döntéshez további méréseket kérünk, ezt meg is tettük. A további mérési adatokból a program az alábbi statisztikákat számítja ki, és az alábbi következtetések vonhatók le:

Műhelyek	n	$\bar{x}$	s^*
I. műhely	60	168,10	4,70
II. műhely	49	169,22	4,64

$$F_{\text{számított}} = 1{,}0260 < F_{95} = 1{,}607;$$

az elméleti sokaságok szórásának megegyezése feltehető,

$$|t_{\text{számított}}| = 1{,}2447 < t_{95} = 1{,}984.$$

95%-os biztonsági szinten elfogadható, hogy a két műhely munkája közt nincs szignifikáns eltérés, $m_1 = m_2$.

Az F-próba és a kétmintás t-próba BASIC programja:

```
10 REM KETMINTAS F ES T PROBA
20 INPUT N1,N2
30 DIM X(100)
40 DIM Y(100)
50 SI=0: NI=0
60 SK=0: NK=0
70 FOR I=1 TO N1 STEP 1
80 PRINT "X(";I;") =": INPUT X(I)
90 SI=SI+X(I): NI=NI+X(I)↑2
100 NEXT I
110 FOR K=1 TO N2
120 PRINT "Y(";K;")=": INPUT Y(K)
130 SK=SK+Y(K): NK=NK+Y(K)↑2
135 NEXT K
140 A1=SI/N1: A2=SK/N2
150 SX=(NI-N1*A1↑2)/(N1-1)
160 SY=(NK-N2*A2↑2)/(N2-1)
170 S1=SQR(SX): S2=SQR(SY)
180 IF SY>SX THEN F=SY/SX:GOTO 210
200 F=SX/SY
210 T1=ABS(A1-A2)
220 T2=(N1-1)*SX+(N2-1)*SY
225 TS=T2/(N1+N2-2)
230 T3=(N1+N2)/(N1*N2)
240 T=T1/SQR(TS*T3)
250 ? "X1 ATLAG="; A1; "X2 ATLAG="; A2
260 ? "S1 SZORAS="; S1; "S2 SZORAS="; S2
270 ? "F ERTEK="; F; "T ERTEK="; T
280 END
```

Egyszerű *próbapélda:*

$N1=3,\quad X(I)=1,\ 2,\ 3,$
$N2=3,\quad Y(I)=4,\ 5,\ 6,$
$\bar{x}=2,\quad \bar{y}=5,\quad s_1^{*2}=1,\quad s_2^{*2}=1,$
$F=1,\quad t=3{,}6742.$

– Az egymintás t-próbára, a kétmintás F- és t-próbára bemutatunk néhány alkalmazási területet. Mindegyiknél előzetesen ki kell mutatni, hogy a vizsgált valószínűségi változó eloszlása jó közelítésben normális.

	Ellenállásértékek									
1. Működés kezdetén	130,2	128,4	131,0	131,3	128,0	126,8	129,9	130,3	132,1	128,6
2. 1000 óra működés után	131,4	128,5	130,1	131,5	127,8	127,1	130,2	130,5	131,3	129,6
3. Különbségek	1,2	0,1	−0,9	0,2	−0,2	0,3	0,3	0,2	−0,8	1,0

	Ellenállásértékek									
1. Működés kezdetén	128,9	131,0	130,5	129,1	127,6	128,7	128,5	129,2	127,3	130,5
2. 1000 óra működés után	128,2	131,5	130,6	127,2	129,3	128,5	129,3	129,4	127,1	129,2
3. Különbségek	−0,7	0,5	0,1	−1,9	1,7	−0,2	0,8	0,2	−0,2	−1,3

16. Egy konzervgyárban adagolóautomata tölti a dobozokat. Az egy dobozba töltendő anyag tömegének várható értékére az előírás 500 g. Mintavétel során az alábbi értékeket kapták: 483, 502, 498, 496, 502, 483, 494, 491, 505, 486. Döntsünk 95%-os biztonsági szinten, teljesül-e a várható értékre az $m_0 = 500$ előírás?

Az adatokból $\bar{x} = 494$; $s^{*2} = 64{,}9$; $t_{\text{számított}} = -2{,}36$; mivel $|t_{\text{számított}}| = 2{,}36 > t_{95} = 2{,}262$, ezért az automata nagy valószínűséggel nem működik jól, újból be kell állítani.

17. Elektromos berendezéshez szükséges alkatrészek ellenállását mérjük az üzembe helyezés kezdetén és 1000 óra működés után. A mért értékeket és különbségüket mutatja az alábbi táblázat ($n = 20$ mérés esetén).

Arra szeretnénk választ kapni, hogy az ellenállás várható értéke szignifikánsan változott-e a működés kezdetén és végén.

Bármennyire csábít, kétmintás F- és t-próbát itt nem végezhetünk, mert a működés kezdetén és végén mért ellenállások nem függetlenek. Ha viszont a két változó várható értéke megegyezik, akkor a különbségük várható értéke 0. Mivel kimutatható, hogy a különbségek jól közelíthetők normál eloszlással (soroljuk a különbségeket pl. -2-től 2-ig 0,8 terjedelmű osztályokba), ezért egymintás t-próbával dönthetünk, fennáll-e az a hipotézis, hogy a különbségek várható értéke 0.

$$\bar{x} = 0{,}02; \quad s^* = 0{,}8551; \quad t_{\text{szám}} = 0{,}1046 < t_{95} = 2{,}093,$$

elfogadható az a feltevés, hogy nem változott lényegesen az ellenállás várható értéke a működés kezdetén és végén.

18. Vizsgáljuk meg, hogy egy új készítési eljárás növeli-e a beton normális eloszlású törőszilárdságát! Az egyik és a másik eljárással készített ($n_1 = 6$ és $n_2 = 6$) próbakocka törőszilárdsága:

I. eljárás. Törőszilárdság, 10 N/cm²		II. eljárás. Törőszilárdság, 10 N/cm²	
	300		305
	301		317
	303		308
	288		300
	294		314
	296		316
Átlag	$\bar{x}_1 = 297$	Átlag	$\bar{x}_2 = 310$
Szórásnégyzet	$s_1^{*2} = 30{,}4$	Szórásnégyzet	$s_2^{*2} = 46$

A mintaátlag növekedést mutat. A szórásnégyzetek növekedése azonban kétségessé teszi, vajon valóban nőtt-e a II. eljárás szerinti gyártásnál a vasbeton törőszilárdsága (hiszen ez az átlagtól való eltérések növekedését jelenti). Két kérdést kell eldöntenünk:

1. Elég jelentős-e a szórások növekedése, hogy az elméleti szórások különböző voltára következtessünk, vagy feltehetjük, hogy az egész sokaságban a két elméleti szórás megegyezik?

2. Statisztikailag elég jelentős (szignifikáns)-e a két mintaátlag közti eltérés, hogy az „elméleti átlagok" különbözőségére következtethessünk, vagy feltehetjük, hogy az egész sokaságban a két „elméleti átlag" (várható érték) megegyezik?

$F = \frac{46}{30{,}4} < F_{95} = 5{,}05$, az egész sokaságban a két szórás egyezése elfogadható.

$|t| = 3{,}33 > t_{99} = 3{,}169$, elvetjük, hogy a két sokaságban a várható érték egyenlő volna, a második eljárás feltehetően javítja a beton törőszilárdságát. (A döntés értékét kissé csökkenti a kis mintaelemszám.)

19. Ki akarjuk mutatni, hogy egy kezelés, amelyet állatokon végeztek, hatást gyakorol a testsúlynövekedésre. Az alábbi táblázat mutatja a gyarapodást dekagrammban:

	Gyarapodás											
Kezelteknél	53	59	63	67	60	57	73	65	58	68	62	71
Nem kezelteknél	61	52	47	51	58	64	60	55	49	53		

A kezelteknél $\bar{x}_1 = 63$; $s_1^{*2} = 36$; $n_1 = 12$; a nem kezelteknél $\bar{x}_2 = 55$; $s_2^{*2} = 31{,}1$; $n_2 = 10$. $F = 1{,}157 < F_{95} = 3{,}10$, a két szórás egyezése feltehető. Az $m_1 = m_2$ nullhipotézissel szemben itt most nem azt feltételezzük, hogy $m_1 \neq m_2$ (tehát hogy $m_1 < m_2$ vagy $m_1 > m_2$), mint az eddigi feladatokban (kétoldali ellenhipotézis), hanem azt, hogy $m_1 > m_2$ (egyoldali ellenhipotézis). Ekkor a szignifikancia szintjét 99%-nak választva, a számított t értéket a 98%-os, kétoldali ellenhipotézisnek megfelelő t értékkel kell összehasonlítani (az $n_1 + n_2 - 2 = 20$ szabadságfoknak megfelelően):

$$t_{\text{számított}} = 3{,}214 > t_{98} = 2{,}528,$$

elvetjük az $m_1 = m_2$ hipotézist, elfogadjuk az $m_1 > m_2$ hipotézist (tehát azt, hogy a kezelt állatoknál nagyobb a gyarapodás várható értéke, mint a kontrollcsoportban).

20. A Fűzfői Papírgyár 1-es papírgépén folyamatosan gyártott 50 tekercsből $31{,}3 \times 32$ cm^2 szakítósablonnal keresztirányban tekercsenként 6–6 helyről mintát vettünk, és a *papír négyzetmétersúlyát* lemértük:

a) tekercsváltáskor azonnal,

b) 65% relatív légnedvességű térben az egyensúlyi nedvességtartalom elérése után (röviden: klimatizálása után), és

c) abszolút száraz állapotban.

Mértük ezenkívül a papír:

d) tényleges *nedvességtartalmát* a gépről lekerülve (tekercsváltáskor),

e) a klimatizálás után beállott egyensúlyi nedvességtartalmat.

Az adatok száma tehát $5 \times 50 \times 6 = 1500$ db, ami kb. 60 t papír mennyiségre vonatkozik.

Mindkét véletlen változó (négyzetmétersúly és nedvességtartalom) eloszlása és ingadozása jelentősen befolyásolja a papír minőségét, fizikai tulajdonságait (vastagság, átnézet, simaság, felületi vízfelvétel, szakítás stb.).

A vizsgálat során először gyakorisági hisztogrammal és χ^2-próbával megállapítottuk, hogy *a vizsgált változók eloszlása igen jól közelíthető normális eloszlással.*

Két kérdés merülhet fel a minőségellenőrző vizsgálattal kapcsolatban:

I. Nem felesleges-e a tekercseken keresztirányban 6 helyen mérni, nem volna-e elegendő csak 1 helyen mérni (vagyis: gyártásirányban *nem egyezik-e meg az így kapott 6 változó szórása és várható értéke,* tehát nem azonos-e a 6 db eloszlásfüggvény)?

II. Nem felesleges-e gyártásirányban 50 tekercsen mérni, nem volna-e elegendő kevesebb tekercset ellenőrizni 6 helyen (vagyis: hosszirányban tekintve 50 változót, *nem egyezik-e meg ezek szórása és várható értéke*)?

A szórások egyezésének vizsgálatát kell először vizsgálni, mégpedig *Bartlett-próbával.* Ha a szórások egyezése feltehető (nem mond ellent a mintának), akkor másodszor *szórásanalízissel, F-próbával* ellenőrizzük, *megegyeznek-e a várható értékek.*

A számítógéppel lefuttatott teljes számítási anyagból néhány részletet mutatunk be az alábbiakban.

A *b)* eset vizsgálatához szükséges adatmátrixot a következő táblázat mutatja.

Famentes illusztrációs nyomópapír négyzetmétersúlya klimatizálás után mérve (század N/m^2)

Tekercsszám	1.	2.	3.	4.	5.	6.
1.	98,0	102,2	99,5	101,0	104,0	99,8
2.	98,2	103,7	97,7	97,7	100,5	99,3
3.	96,5	100,3	96,2	98,7	100,3	99,0
4.	97,8	106,2	100,5	100,5	102,5	101,2
5.	97,5	101,0	97,2	98,2	100,8	99,3
6.	95,7	100,0	97,3	99,0	101,0	98,5
7.	99,0	102,0	99,5	98,5	102,3	100,7
8.	98,5	101,8	98,0	100,7	101,5	99,8
9.	97,4	101,1	97,5	101,0	101,2	99,9
10.	98,0	101,0	98,5	98,5	100,0	98,0
11.	93,2	97,0	93,5	95,5	98,2	95,0
12.	107,0	107,1	107,2	106,5	108,5	104,0
13.	103,0	106,5	104,5	103,5	107,0	100,8
14.	103,3	107,6	104,0	104,2	108,5	102,7
15.	99,5	104,2	103,2	103,0	104,7	101,8
16.	99,2	102,2	101,0	103,0	103,5	101,2
17.	101,8	105,5	102,8	101,8	107,3	101,0
18.	97,5	102,7	97,3	99,0	100,5	97,5
19.	99,5	103,0	99,7	101,5	105,5	100,0
20.	100,5	103,2	100,4	102,7	103,0	102,7
21.	101,8	103,2	100,2	101,9	103,9	99,5
22.	97,3	105,0	101,5	102,8	102,3	97,2
23.	101,5	104,8	104,5	106,7	106,6	101,4
24.	98,5	99,4	97,6	100,4	99,0	95,5
25.	104,5	103,8	104,0	104,0	105,2	102,3
26.	104,5	104,0	102,3	104,7	106,3	103,2
27.	98,0	101,0	97,9	100,3	99,7	97,3
28.	97,8	99,0	96,8	97,5	97,6	93,2
29.	103,5	105,5	102,0	104,0	105,6	100,2
30.	103,0	103,5	103,2	103,5	105,0	102,5
31.	99,8	106,0	101,0	102,0	103,5	101,2
32.	98,3	102,5	99,5	100,9	103,7	102,3
33.	100,3	102,5	101,7	100,7	104,0	104,0
34.	98,2	101,0	100,5	100,0	102,8	103,1
35.	97,0	99,2	99,5	99,0	100,6	100,2
36.	99,2	101,4	99,0	100,8	102,5	100,5

A táblázat folytatása

Tekercsszám	1.	2.	3.	4.	5.	6.
37.	99,5	101,5	98,4	101,0	101,4	100,8
38.	99,7	100,7	97,7	99,5	101,4	99,0
39.	100,7	99,2	99,3	101,7	102,0	100,0
40.	100,4	102,7	101,5	101,3	103,0	100,0
41.	102,2	102,0	99,8	100,7	102,3	102,7
42.	102,0	103,5	101,5	101,0	102,0	100,4
43.	99,0	99,0	97,0	99,0	99,8	101,4
44.	100,0	98,3	100,0	101,0	101,2	101,2
45.	99,0	102,0	98,8	100,9	101,5	100,5
46.	98,4	100,8	100,0	100,0	101,5	99,0
47.	95,5	94,2	95,5	96,4	97,8	96,5
48.	97,9	100,0	97,8	100,9	102,2	100,3
49.	99,2	98,0	97,5	100,0	99,5	98,2
50.	103,0	102,8	102,6	104,2	104,6	104,0
Változó	X_1	X_2	X_3	X_4	X_5	X_6

Vizsgáljuk meg, hogy a 6 db X_i oszlopváltozó nem azonos normál eloszlású-e!

A szórások egyezésének vizsgálata (Bartlett-próba):

Oszlopok	s_i^*	s_i^{*2}	$\lg s_i^*$	f_i	$\frac{1}{f_i}$
1	2,60	6,77	0,83059	49	0,0204
2	2,68	7,17	0,85552	49	0,0204
3	2,67	7,14	0,85370	49	0,0204
4	2,33	5,46	0,73719	49	0,0204
5	2,60	6,78	0,83123	49	0,0204
6	2,31	5,31	0,71509	49	0,0204
Összesen:		38,63	4,82332	294	0,1224

$$k=6, \quad f=\sum_{i=1}^{k} f_i = 294,$$

$$S^{*2} = \frac{1}{f}\sum_{i=1}^{k} f_i S_i^{*2} = \frac{1}{294}\cdot 49\cdot 38{,}63 = 6{,}4383,$$

$$c = 1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{f_i}-\frac{1}{f}\right) = 1+\frac{1}{15}\left(0{,}1224-\frac{1}{294}\right) = 1{,}0079,$$

$$K^2 = \frac{2{,}3026}{c}\left(f\lg S^{*2}-\sum_{i=1}^{k} f_i \lg S_i^{*2}\right) =$$

$$= \frac{2{,}3026}{1{,}0079}(294\cdot 0{,}8088-49\cdot 4{,}82332),$$

$$K^2 = 3{,}3001 < \chi^2_{99,0} = 15{,}1$$

(a próba szintje 99% szabadságfok, vagyis a paraméter $k-1 = 5$). Mivel a számított érték kisebb a táblázatbelinél, *a szórások egyezését a 6 oszlopváltozóra elfogadhatjuk*. [Mivel a mérések száma egy-egy változóra 50 (jóval nagyobb, mint 4), a χ^2-eloszlással való közelítés megfelelő.]

A várható értékek egyezésének vizsgálata (szórásanalízis). Ki kell számítanunk a *csoportok közti négyzetösszeget* és abból az

$$S_1^2 = \frac{\sum_{i=1}^{r} n_i(\bar{x}_i-\bar{x})^2}{r-1}$$

értéket. Itt most a tesztelendő változók száma $r=6$, mindegyik mintaelemszám $n_i=50$, az egész (6×50 elemű) minta átlaga $\bar{x}=100{,}90$, az egyes oszlopok $\bar{x}_i$ átlaga, és belőlük a csoportok közti négyzetösszeg számítása látható az alábbi táblázatban:

Oszlopok	$\bar{x}_i$	$\bar{x}$	$\bar{x}_i-\bar{x}$	$(\bar{x}_i-\bar{x})^2$
1	99,62		−1,28	1,6384
2	102,10		1,20	1,4400
3	99,88		−1,02	1,0404
4	101,03	100,90	0,13	0,0169
5	102,59		1,69	2,8561
6	100,20		−0,70	0,4900
Összesen:				7,4818

$$S_1^2 = \frac{50\cdot 7{,}4818}{5} = 74{,}818.$$

Ki kell számítanunk *a csoporton belüli négyzetösszeget* és abból az

$$S_2^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ji}-\bar{x}_i)^2}{n-r}$$

értékét. Itt $n = \sum_{i=1}^{r} n_i = 6\cdot 50 = 300$, a 6 darab $\bar{x}_i$ átlagot az előző táblázatban közöltük. A belső szummák:

$$\sum_{j=1}^{n_i}(x_{ji}-\bar{x}_i)^2 \qquad (i = 1, 2, \ldots, 6)$$

értékét úgy kapjuk, hogy minden oszlopban az egyes adatok négyzetes eltérését vesszük az oszlop elemeinek átlagától és összegezünk. Így a belső szummák tulajdonképpen a Bartlett-próbánál szereplő S_i^{*2} értékek (n_i-1)-szeresei, azaz 49-szeresei. Ezek összege:

$$\sum_{i=1}^{6}\sum_{j=1}^{50}(x_{ji}-\bar{x}_i)^2$$

az ottani $\sum_{i=1}^{6} S_i^{*2} = 38{,}63$ értéknek a 49-szerese. Így

$$S_2^2 = \frac{49 \cdot 38{,}63}{300-6} = 6{,}4383.$$

Ha a nullhipotézis igaz, tehát ha az oszlopváltozók várható értéke megegyezik, akkor a csoportok közötti és csoportokon belüli *(külső és belső) szórásnégyzet* hányadosa F-eloszlású.

$$F = \frac{74{,}818}{6{,}4383} = 11{,}62077 > F_{95} = 2{,}25,$$

az oszlopváltozók várható értékének egyenlőségét elvetjük. (Az F értéket a 9. táblázatból vettük, a számláló szabadságfoka $f_1 = r-1 = 5$, a nevező szabadságfoka $f_2 = n-r = 294$.)

Végeredményben tehát azt kaptuk, hogy *a klimatizálás után a négyzetmétersúly oszlopváltozók nem azonos normál eloszlásúak, célszerű egy tekercset továbbra is keresztirányban 6 helyen mérni.*

Az adatok és a számítások részletezése nélkül bemutatjuk a vizsgálat összesített eredményét:

I. A 6 oszlopváltozó szórásának és várható értékének az egyezése:

Eset		Szórások egyeznek-e	Várható értékek egyeznek-e
m²-súly	*a)*	nem	—
	b)	igen	nem
	c)	igen	nem
Nedvesség	*d)*	igen	igen
	e)	igen	igen

II. Az 50 sorváltozó szórásának és várható értékének egyezése:

Eset		Szórások egyeznek-e	Várható értékek egyeznek-e
m²-súly	*a)*	igen	igen
	b)	igen	igen
	c)	igen	igen
Nedvesség	*d)*	igen	nem
	e)	igen	nem

Következtetés:

– a négyzetmétersúlyt célszerű tekercsenként 6 helyen mérni, de nem kell 50 tekercset vizsgálni;

– a nedvességtartalmat elég tekercsenként 1 helyen mérni, de célszerű több tekercset vizsgálni.

Mindez azt eredményezte, hogy *a mérések száma a minőségellenőrzéskor jelentősen csökkenthető volt anélkül, hogy az ellenőrzés biztonságát csökkentettük volna.*

Megjegyzés

Az ún. teljes szórásnégyzetet, az

$$S^{*2} = \frac{1}{n-1} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ji} - \bar{x})^2$$

értéket ki sem számítottuk. A teljes szórásnégyzet kiszámítása egyrészt arra jó, hogy az

$$(r-1)S_1^2 + (n-r)S_2^2 = (n-1)S^{*2}$$

összefüggés alapján ellenőrizzük, jól számítottuk-e ki a külső és belső szórásnégyzetet. Másrészt feloszthatjuk a teljes szórásnégyzet $(n-1)$-szeresét (a teljes négyzetösszeget) a külső és belső négyzetösszegek összegére. A belső négyzetösszeg a csoporttagok négyzetes eltérését jellemzi a csoportátlagtól, minden változó szóródik a maga „átlaga" körül, tehát ez a négyzetösszeg a *véletlen hibát* jellemzi. A külső négyzetösszeg a csoportátlagoknak a „főátlagtól" (az egész minta átlagától) való négyzetes eltérését méri (természetesen a csoport „súlyával" súlyozottan), ez tehát a *nem véletlen okozta (a szisztematikus) hibára* enged következtetni, vagyis arra, hogy a mintaátlagok eltérése ellenére az egész sokaságban a várható értékek egyenlők-e vagy sem,

$$M(x_i) = m \qquad (i = 1, 2, \ldots, r)$$

fennáll-e vagy nem, amivel együtt jár, hogy

$$M(\bar{x}_i) = M(\bar{x}) = m \qquad (i = 1, 2, \ldots, r)$$

fennáll-e vagy sem.

21. Kúpgörgős csapágy belső gyűrűjének kúpszögét mérjük egy hitelesített A műszerrel, és egy hitelesítendő B műszeren. Az A műszeren végzett mérés eredménye, X normál eloszlású, a B műszeren végzett mérés eredménye, Y szintén normál eloszlású. Akkor tekinthető hitelesnek a B műszer is, ha a

$$H_0:\ M(X)=M(Y)$$

hipotézis fennáll. A mérési eredmények:

Műszer	Mérések száma	Mintaátlag, μ	Tapasztalati szórás, μ
A	$n=100$	$\bar{x}=0{,}625$	$s_n=0{,}754$
B	$m=100$	$\bar{y}=0{,}471$	$s_m=1{,}269$

a) Feltehető-e, hogy az egész sokaságban a szórások egyenlők?
b) Feltehető-e, hogy az egész sokaságban a várható értékek egyenlők (hitelesnek tekinthető-e a B műszer)?

---

a) Mivel a korrigált és a korrigálatlan szórásnégyzetek aránya egyenlő, az F próbastatisztika aktuális értékét számíthatjuk a korrigálatlan szórásnégyzetek arányával:

$$F=\frac{1{,}269^2}{0{,}754^2}=2{,}83>F_{95}=1{,}39.$$

Az F_{95} értéket a 9. táblázatból, a (99,99) paraméterpárhoz tartozó sorból és oszlopból vettük. A tapasztalati szórásnégyzetek aránya túl nagy ahhoz, hogy az egész sokaságbeli szórásnégyzeteket egyenlőknek feltételezhessük, a $D(X)=D(Y)$ feltevést elvetjük.

b) Mivel a szórásokról nem tehető fel az egyenlőség, kétmintás t-próbával nem dolgozhatunk, csak *Welch-próbával.*

A számítandó, Student-eloszlású t érték:

$$t=\frac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{S_n^2}{n}+\dfrac{S_m^2}{m}}}=\frac{0{,}625-0{,}471}{\sqrt{\dfrac{0{,}754^2}{100}+\dfrac{1{,}269^2}{100}}}=1{,}043.$$

Az f szabadságfok az alábbi egyenletből számolandó:

$$\frac{1}{f}=\frac{1}{m-1}\left(\frac{\dfrac{S_m^2}{m}}{\dfrac{S_n^2}{n}+\dfrac{S_m^2}{m}}\right)^2+\frac{1}{n-1}\left(\frac{\dfrac{S_n^2}{n}}{\dfrac{S_n^2}{n}+\dfrac{S_m^2}{m}}\right)^2=$$

$$=\frac{1}{99}\left(\frac{1{,}610361}{2{,}178877}\right)^2+\frac{1}{99}\cdot\left(\frac{0{,}568516}{2{,}178877}\right)^2=\frac{2{,}916473}{470{,}00299},$$

$$f=161{,}154\,59\approx 161,$$

$$|t|=1{,}043<t_{95}=1{,}960.$$

A t_{95} értékét a 8. táblázatból, az $n\to\infty$ sorból vettük. A minta elegendő nagyságú ahhoz, hogy t értékét közelítőleg standard normál eloszlásúnak vegyük.

22. Legyen X 1-szórású, normális eloszlású változó. A várható értékét nem ismerjük, feltételezésünk az m várható értékre

$$H_0:\ m=m_0=9{,}5,$$

a

$$H_1:\ m=m_1=10$$

ellenhipotézissel szemben.

Gyakorlati meggondolások, a különféle fajtájú hibák okozta károk figyelembevételével engedjünk meg $\alpha=0{,}05$, ill. $\beta=0{,}10$ első- és másodfajú hibát.

Döntsünk szekvenciális próbával, elfogadjuk-e, hogy a várható érték $m=9{,}5$, vagy inkább az $m=10$ ellenhipotézis mellett döntünk. A változó mért értékei sorban:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
8	12	6	16	4	14	8	12	6	2	2

.

A *sűrűségfüggvény:*

$$f(x,m)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2}},$$

a döntési tartományok határai:

$$B = \frac{\beta}{1-\alpha} = \frac{0{,}10}{0{,}95} = 0{,}105\,263\,16,$$

$$A = \frac{1-\beta}{\alpha} = \frac{0{,}90}{0{,}05} = 18,$$

ill.

$$\ln B = -2{,}251\,291\,8, \qquad \ln A = 2{,}890\,371\,7.$$

Ha $k-1$ lépésben az indifferens (közömbös) tartományban maradtunk, akkor a k-adik lépés után *a likelihood- (valószínűség-) hányados* logaritmusa:

$$\ln L = \sum_{i=1}^{k} \ln \frac{f(x_i, m_1)}{f(x_i, m_0)} = \sum_{i=1}^{k} \ln \frac{e^{-\frac{(x_i-m_1)^2}{2}}}{e^{-\frac{(x_i-m_0)^2}{2}}} =$$

$$= \sum_{i=1}^{k} \frac{1}{2}[(x_i-m_0)^2-(x_i-m_1)^2],$$

$$\ln L = \sum_{i=1}^{k} \left[\frac{1}{2}(m_0^2-m_1^2)+(m_1-m_0)x_i\right] =$$

$$= (m_1-m_0)\left[\sum_{i=1}^{k} x_i - \frac{k}{2}(m_1+m_0)\right].$$

Mindaddig nem jutunk döntésre (az indifferens tartományban maradunk), míg

$$\ln B < \ln L = (m_1-m_0)\left[\sum_{i=1}^{k} x_i - \frac{k}{2}(m_1+m_0)\right] < \ln A,$$

vagyis addig, amíg

$$\frac{1}{m_1-m_0}\ln B + \frac{k}{2}(m_1+m_0) < \sum_{i=1}^{k} x_i < \frac{1}{m_1-m_0}\ln A + \frac{k}{2}(m_1+m_0).$$

Az alsó és felső határ k-nak lineáris függvénye, azonos pozitív iránytangenssel. *Mindaddig az indifferens tartományban maradunk, amíg* $\sum_{i=1}^{k} x_i$ *értékéből*

kapott $\left(k; \sum_{i=1}^{k} x_i\right)$ *pont a két emelkedő egyenes között marad.* Mivel $m_1 > m_0$, ezért

$$\frac{1}{m_1-m_0}\ln B < \frac{1}{m_1-m_0}\ln A,$$

az alsó határ valóban „mélyebben" (lejjebb) haladó egyenest határoz meg, mint a felső.

Döntés akkor történik, ha a kapott pont kívül esik az indifferens zónán. Ha a felső határ egyenese felett van a pont, a H_0 hipotézist elvetjük, ha az alsó határ egyenese alatt, a H_0 hipotézist elfogadjuk.

A $k>0$ félsík pontjait két párhuzamos egyenessel tehát három tartományra osztjuk: egy alsó, elfogadási zónára, egy középső, indifferens zónára, és egy felső, elvetési zónára.

Numerikus adatokra térve azt kapjuk, hogy mindaddig az indifferens tartományban vagyunk, míg

$$9{,}75k - 4{,}50 < \sum_{i=1}^{k} x_i < 9{,}75k + 5{,}78$$

teljesül. Az egymás utáni adatokat tekintve:

x_i	8	12	6	16	4	14	8	12	6	2	2
Σx_i	8	20	26	42	46	60	68	80	86	88	90

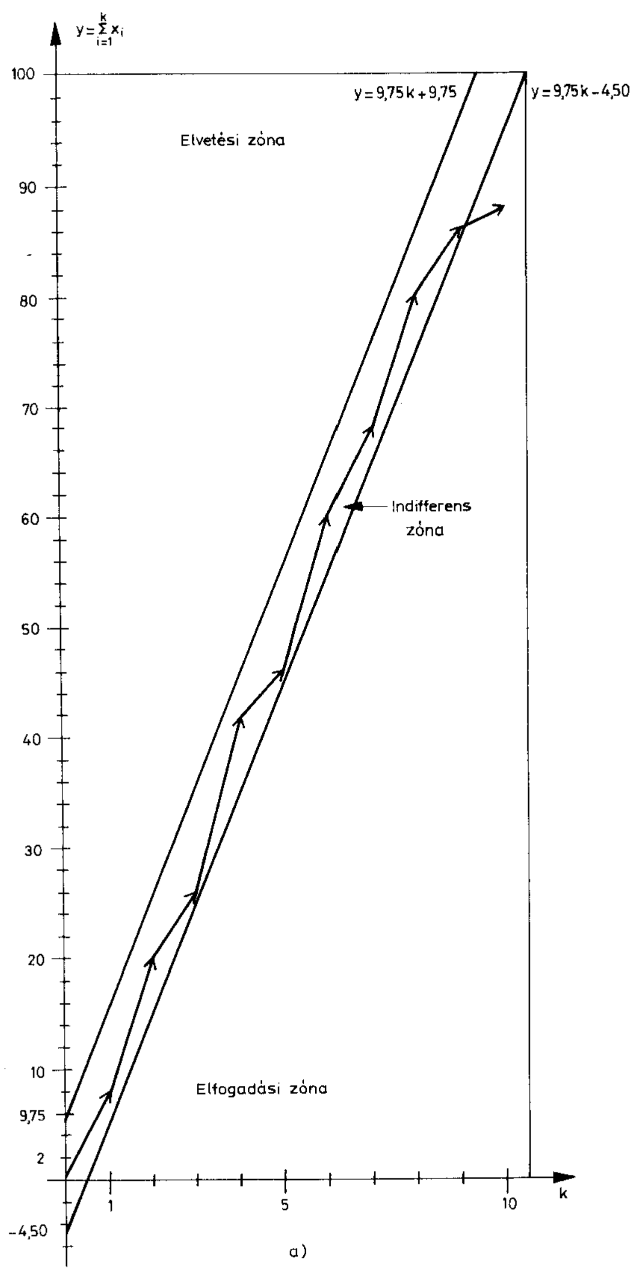
A tartományokat és az (1,8), (2,20), (3,26), (4,42), (5,46), (6,60), (7,68), (8,80), (9,86), (10,88) pontokat ábrázolva (36*a* ábra), a 10. pont lép ki először az indifferens tartományból, mégpedig az elfogadási zónába. A H_0 hipotézist,

$$m = 9{,}5$$

elfogadhatjuk. A további mérési eredményre már nincs is szükség.

Megjegyzések

a) Tíz mérés elég volt a döntéshez. A t-próbához „átlagosan" ennél több mérés kell. A 11. mérés fölösleges volt.



36*a* ábra

b) A szekvenciális próbánál nem úgy kell elképzelnünk, hogy előre adott (mondjuk) 100 mérés, és utána látunk hozzá a próba végrehajtásához. Egy-egy mérést kell elvégeznünk, és megmondani, kell-e újabb.

c) Előre megadható, hogy milyen Σx_i esetén maradunk az indifferens (döntésképtelen) zónában. Az első mérésnél például

$$-4{,}50+9{,}75 = 5{,}25 < \Sigma x_i < 5{,}78+9{,}75 = 15{,}53,$$

a második mérésnél

$$15 < \Sigma x_i < 25{,}28$$

esetén nem tudunk döntést hozni, és igy tovább. Mindez leolvasható az ábráról is.

d) Kapcsolatba hozható mindez azzal is, hogy milyen mintaátlag esetén vetjük el, fogadjuk el a nullhipotézist, vagy kérünk további mérést. Az indifferens tartományban maradunk, valahányszor az $\bar{x}$ mintaátlagra fennáll:

$$-\frac{4{,}50}{k}+9{,}75 < \bar{x} < \frac{5{,}78}{k}+9{,}75$$

(k a mérések száma). Tíz mérés esetén például

$$9{,}3 < \bar{x} < 10{,}328$$

adja meg az indifferens tartományt. Ha $k \to \infty$, az alsó és felső határ egyaránt 9,75-re szűkül, egyre nagyobb mérésszám esetén tehát egyre nagyobb valószínűségű, hogy kilépünk az indifferens zónából (döntéshez jutunk).

e) A szekvenciális próba egyenértékű egy speciális *bolyongási problémával:* a mozgó pont mindig 1-et jobbra lép, ugyanakkor adott valószínűséggel lép fölfelé adott távolságot. Ha a mozgó pont útja az alsó egyenest átmetszi, H_0-t elfogadjuk, ha a felső egyenest átmetszi, akkor H_0-t elvetjük.

23. Legyen X binomiális eloszlású ismeretlen p paraméterrel. Hogyan kell elvégezni a szekvenciális próbát a H_0: $p=p_0$ nullhipotézis, H_1: $p=p_1>p_0$ alternatív hipotézis esetén, ha előre megszabjuk az elsőfajú hiba α és a másodfajú hiba β valószínűségét?

Legyen az X változó értéke 1, ha a szóban forgó esemény bekövetkezett, legyen X értéke 0, ha az esemény nem következett be. $P(X=1) = p$, $P(X=0) = 1-p$.

Itt most

$$\lg \frac{P(X=x_i/p_1)}{P(X=x_i/p_0)} = \begin{cases} \lg \dfrac{p_1}{p_0}, & \text{ha} \quad x_i=1, \\[2ex] \lg \dfrac{1-p_1}{1-p_0}, & \text{ha} \quad x_i=0. \end{cases}$$

Legyen az első n mintaelem között K_n darab 1 és $(n-K_n)$ darab 0. Ekkor

$$\lg L = \sum_{i=1}^{n} \lg \frac{P(X=x_i/p_1)}{P(X=x_i/p_0)} = K_n \lg \frac{p_1}{p_0} + (n-K_n) \lg \frac{1-p_1}{1-p_0} =$$

$$= n \cdot \lg \frac{1-p_1}{1-p_0} + K_n \left(\lg \frac{p_1}{p_0} - \lg \frac{1-p_1}{1-p_0} \right) =$$

$$= n \cdot \lg \frac{1-p_1}{1-p_0} + K_n \lg \left(\frac{p_1}{p_0} \cdot \frac{1-p_0}{1-p_1} \right).$$

Újabb és újabb mintát kell vennünk mindaddig, amíg

$$\lg B < \lg L < \lg A \qquad \left(\text{itt } B = \frac{\beta}{1-\alpha},\ A = \frac{1-\beta}{\alpha}\right),$$

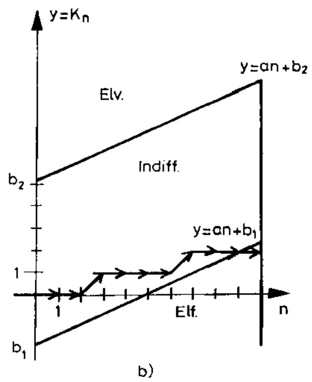
azaz mindaddig, amíg

$$\underbrace{\frac{\lg B}{\lg \dfrac{p_1(1-p_0)}{p_0(1-p_1)}}}_{b_1} + n \cdot \underbrace{\frac{\lg \dfrac{1-p_0}{1-p_1}}{\lg \dfrac{p_1(1-p_0)}{p_0(1-p_1)}}}_{a} < K_n < \underbrace{\frac{\lg A}{\lg \dfrac{p_1(1-p_0)}{p_0(1-p_1)}}}_{b_2} + n \cdot \underbrace{\frac{\lg \dfrac{1-p_0}{1-p_1}}{\lg \dfrac{p_1(1-p_0)}{p_0(1-p_1)}}}_{a}.$$

Itt a, b_1, b_2 a hipotézisekből kiszámítható állandók, az indifferens tartomány határai tehát az

$$y = an + b_1 \quad \text{és} \quad y = an + b_2$$

párhuzamos egyenesek. Mindig újabb és újabb kísérletet végezve, ábrázoljuk az $(1, K_1)$, $(2, K_2)$, ... pontokat, mindaddig, míg azok az indifferens zónába esnek. Ha kilépünk az alsó egyenes alá, akkor a nullhipotézis elfogadásával, ha kilépünk a felső egyenes fölé, a H_1 hipotézis elfogadásával ér véget a próba (36*b* ábra).



36*b* ábra

Megjegyzés

a) Könnyű belátni, hogy a $p_1 > p_0$ feltevés miatt $b_1 < b_2$, $a > 0$, ezenkívül az egyenlőtlenség-rendszer rendezése során pozitív számmal osztottunk, tehát az egyenlőtlenségek iránya nem változott.

b) Ez is egyenértékű egy bolyongási problémával, amelynek során a mozgó pont 1-et mindig jobbra lép, miközben fölfelé $1-p$ valószínűséggel 0-t lép, és 1-et lép p valószínűséggel.

24. *Adjunk próbát a binomiális eloszláshoz a H_0: $P(A)=p_0$ nullhipotézis ellenőrzésére!*

Említettük, hogy minden konfidencia-intervallumhoz hozzárendelhető egy próba. Az intervallumbecsléseknél megadtunk olyan (c_1, c_2) intervallumot, amelybe a binomiális eloszlású k gyakoriság nagy, $1-\varepsilon$ valószínűséggel esik:

$$P(c_1 \leqq k \leqq c_2/H_0) = \sum_{i=c_1}^{c_2} \binom{n}{i} p_0^i (1-p_0)^{n-i} = 1-\varepsilon.$$

Ha a gyakoriság erre az intervallumra esik, H_0-t elfogadjuk, ha a

$$k < c_1$$

vagy a

$$k > c_2$$

kritikus tartományra, akkor H_0-t visszautasítjuk. A c_1 és c_2 megkeresése a

$$P(k<c_1/H_0) = \frac{\varepsilon}{2} \quad \text{és} \quad P(k>c_2/H_0) = \frac{\varepsilon}{2}$$

egyenletekből történik, a binomiális eloszlásra vonatkozó 1. vagy 2. táblázat használatával. Az 1. táblázat tetszőleges p-re használható, mert ha pl. $P(A)=0{,}60$ (és ez nincs a táblázatban), akkor az ellentett esemény valószínűségére teszünk nullhipotézist, $P(\bar{A})=0{,}40$ már megtalálható az 1. táblázatban.

A hiba inkább ott van, hogy az 1. táblázatban $1 \leqq n \leqq 20$.

Nagy mintaelemszám esetén a binomiális eloszlást normálissal közelítjük. Tudjuk ugyanis, hogy nagy n-re

$$\frac{k-np}{\sqrt{np(1-p)}}$$

közelítőleg $N(0, 1)$-eloszlású. Így a

$$\Phi(u_\varepsilon)-\Phi(-u_\varepsilon) = 2\cdot\Phi(u_\varepsilon)-1 = 1-\varepsilon$$

összefüggésből meghatározunk olyan u_ε-t, amelyre

$$P(-u_\varepsilon \leqq \frac{k-np}{\sqrt{np(1-p)}} \leqq u_\varepsilon/H_0) \approx 1-\varepsilon.$$

Így egy próbát kaptunk k-ra,

$$X_k = \{k < np_0-u_\varepsilon\sqrt{np_0(1-p_0)} \quad \text{vagy} \quad k > np_0+u_\varepsilon\sqrt{np_0(1-p_0)}\}$$

kritikus tartománnyal. Igazolható, hogy a normál eloszlással való közelítés jó, ha

$$n \geqq \frac{9}{p_0(1-p_0)}.$$

Nagy n, kis p_0 esetén a binomiális eloszlást Poisson-eloszlással is jól közelíthetjük.

Megjegyzés

Igazolható, hogy a H_0: $p>p_0$ nullhipotézis vizsgálatára egyenletesen legerősebb a próba, ha $X_k = \{k \leqq k_0\}$ a kritikus tartomány. Itt p_0 és k_0 olyan konstans értékek, melyekre $1-\varepsilon$-szintű próba esetén a

$$P(k<k_0/p_0) = 1-\varepsilon$$

feltétel teljesül.

3.3. Illeszkedés-, homogenitás- és függetlenségvizsgálat

Azt, hogy *egy X véletlen változó adott eloszlású-e,* eddig a következő módon vizsgáltuk. Diszkrét esetben összehasonlítottuk a $P(X=x_i)$ valószínűségeket a mintában és a feltételezett elméleti eloszlásban. Folytonos eloszlásnál a tapasztalati és a feltételezett elméleti sűrűségfüggvény összehasonlításából következtettünk. Ha az elméleti és mintabeli valószínűségek, ill. sűrűségfüggvények jól illeszkednek egymáshoz, akkor feltevésünket az eloszlást illetően elfogadtuk, ellenkező esetben elvetettük.

Igen ám, de ebben az egyszerű módszerben van egy hiba: a szubjektivitás. Hiszen szubjektív az, hogy az elméleti és tapasztalati értékek milyen mérvű illeszkedése esetén fogadunk el, ill. utasítunk vissza. Egyáltalán: jó volna mérőszámot létrehozni az illeszkedés mértékére.

A tapasztalati és a feltételezett elméleti eloszlásfüggvények illeszkedésének vizsgálatára, a mintának az elméleti eloszláshoz való illeszkedésének eldöntésére szolgáló eljárás (próba) az ún. illeszkedésvizsgálat. Tiszta illeszkedésvizsgálatról beszélünk, ha paramétert nem kell becsülni, feltevésünk egyértelműen egyetlen eloszlásfüggvényre vonatkozik. *Becsléses az illeszkedésvizsgálat,* ha hipotézisünk csupán az eloszlás jellegét, az eloszlásfüggvény típusát adja meg (normál, exponenciális stb.), az eloszlásfüggvény paramétereit a mintából kell becsülni. Ez utóbbi a gyakoribb.

A most következő statisztikai problémákat három típusba sorolhatjuk:

a) egy minta tekinthető-e adott eloszlásból származónak *(illeszkedésvizsgálat),*

b) két minta homogén-e, tekinthető-e azonos eloszlásból származónak *(homogenitásvizsgálat),*

c) két valószínűségi változó a minta alapján függetlennek tekinthető-e *(függetlenségvizsgálat).*

1. Grafikus normalitásvizsgálat Gauss-papíron

Említettük, hogy a változó normális voltát ellenőrizhetjük grafikusan, hisztogram megszerkesztésével. Most a normalitás vizsgálatára egy másik grafikus módszert mutatunk, az ún. Gauss-papíron való szemléltetést. Ezt különösen kis elemszámú minták esetén használják, vagy ha csak tájékozódó jellegű a vizsgálat, vagy ha ellenőrizni kívánják, nem változott-e meg egy változó korábban kimutatott normalitása.

„Egyenesítsük ki" először a standard normális eloszlás $\Phi(x)$ eloszlásfüggvényét!

Ismeretes, hogy a Φ függvény $-3, -2, -1, 0, 1, 2, 3$ helyen felvett értékei a következők:

$$\Phi(-3)=0{,}00135, \quad \Phi(3)=0{,}99865,$$
$$\Phi(-2)=0{,}0228, \quad \Phi(2)=0{,}9772,$$
$$\Phi(-1)=0{,}1587, \quad \Phi(1)=0{,}8413,$$
$$\Phi(0)=0{,}5.$$

(A változó értékei a 0 várható értéktől csaknem teljes egészében -3-szoros és $+3$-szoros szórásra, -3 és $+3$ közé esnek.)

Ábrázoljuk most a következő pontokat:

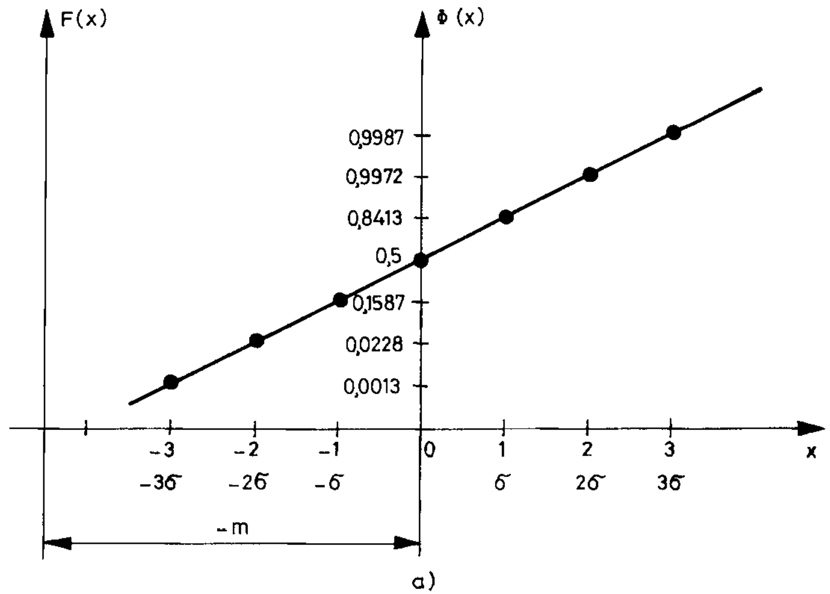
$x=$ 0-nál mérjük fel ordinátaként a $\Phi(0)$ $=0{,}5$-et,
$x=$ 1-nél mérjük fel ordinátaként a $\Phi(1)$ $=0{,}8413$-et,
$x=-1$-nél mérjük fel ordinátaként a $\Phi(-1)=0{,}1587$-et,
$x=$ 2-nél mérjük fel ordinátaként a $\Phi(2)$ $=0{,}9772$-et,
$x=-2$-nél mérjük fel ordinátaként a $\Phi(-2)=0{,}0228$-et

stb., de úgy, hogy $\Phi(1)$ 1 távolságegységgel feljebb, $\Phi(-1)$ 1 távolságegységgel lejjebb, $\Phi(2)$ 2 távolságegységgel feljebb, $\Phi(-2)$ 2 távolságegységgel lejjebb kerüljön, mint a $\Phi(0)=0{,}5$ stb.

Az ordinátatengely skálázása ezáltal egyenlőtlen lett, viszont így a $\Phi(x)$ függvény képe egy egyenes, ahol nyilván a többi érték is ezen az egyenesen fekvő pontot határoz meg (37*a* ábra).

Ha az x tengelyt most átskálázzuk (-3 helyébe -3σ, -2 helyébe -2σ kerül stb., l. az ábrát), akkor az előző pontok egy 0 várható értékű, σ szórású normális eloszlású változó eloszlásfüggvényének a pontjai.

Végül toljuk el az ordinátatengelyt eredeti helyzetével párhuzamosan az x tengely mentén $-m$ egységgel. Az így kapott koordinátarendszerben az m várható értékű, σ szórású normális eloszlású változó $F(x)$ eloszlásfüggvényének a képe egy egyenes.



37*a* ábra

Ezek a lépések annak felelnek meg, hogy

$$F(x) = \Phi\left(\frac{x-m}{\sigma}\right),$$

vagyis az $F(x)$ függvény $\Phi(x)$-ből 2 lineáris transzformációval kapható meg:

a) $\Phi(x)$-ből $\Phi\left(\frac{x}{\sigma}\right)$-t az x tengely menti egységek σ-ra változtatásával.

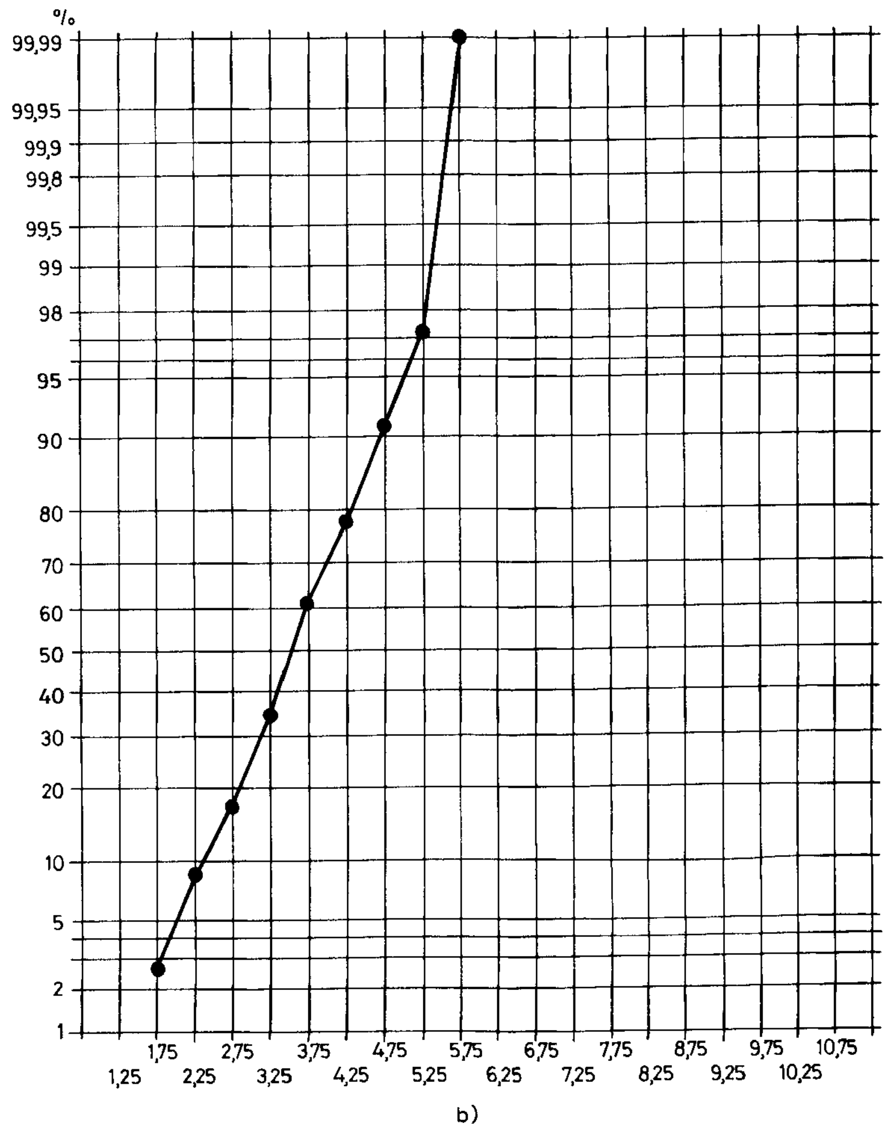
b) $\Phi\left(\frac{x}{\sigma}\right)$-ból $\Phi\left(\frac{x-m}{\sigma}\right)$-t az ordinátatengely párhuzamosan, $-m$ egységgel történő eltolásával kaphatjuk meg.

Ha a minta alapján kapott tapasztalati eloszlásfüggvényt az előbbi koordináta-rendszerben (ún. Gauss-papíron) ábrázoljuk, normális eloszlás esetén a pontok közelítőleg egy egyenesre esnek (azaz egy egyenes körül csoportosulnak), annak megfelelően, hogy a tapasztalati eloszlásfüggvény nagy valószínűséggel jól közelíti az elméleti eloszlásfüggvényt, ha n elég nagy.

Nézzük most az előzőekben már szerepelt cipőgyártási példánkban a talpfelerősítési szilárdság értékeit! Tekintsük úgy, hogy egy-egy intervallumban a szilárdság az osztályhatárok számtani közepét (az osztályközepet) veszi fel.

Szilárdsági osztály	Osztályközép	Valószínűség	A valószínűségek összege (az eloszlásfüggvény értéke)
1,5–2,0	1,75	$\frac{1}{36} = 0{,}028$	0,028
2,0–2,5	2,25	0,055	0,028 + 0,055 = 0,083
2,5–3,0	2,75	0,083	0,166
3,0–3,5	3,25	0,167	0,333
3,5–4,0	3,75	0,278	0,611
4,0–4,5	4,25	0,167	0,778
4,5–5,0	4,75	0,139	0,917
5,0–5,5	5,25	0,055	0,972
5,5–6,0	5,75	0,028	1,000
Összesen:		1,000	

A 37*b* ábrán feltüntetett Gauss-papíron ábrázolva az (1,75; 0,028), (2,25; 0,083) stb. pontokat, látható, hogy azok közelítőleg egy egyenesre esnek, tehát a szilárdság a minta alapján normális eloszlásúnak tekinthető. (A hálózat ordinátatengelye az előzőleg mutatottnál sűrűbb beosztású, a valószínűségértékek %-ban vannak bejelölve.)



37*b* ábra

2. Egy egyszerű normalitásvizsgálat

Ha T standard normál eloszlású, akkor sűrűségfüggvénye:

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} < \frac{1}{\sqrt{2\pi}}, \quad \text{ha} \quad t \neq 0,$$

$$\ln \sqrt{2\pi}\,\varphi(t) = -\frac{t^2}{2},$$

$$\ln(-2\ln\sqrt{2\pi}\,\varphi(t)) = 2\ln t, \quad \text{ha} \quad t > 0,$$

az $y = \ln(-2\ln\sqrt{2\pi}\,\varphi(t))$, $\ln t = z$ helyettesítést végrehajtva, *az $(y; z)$ koordináta-rendszerben a most kapott függvény képe egyenes.*

Ha $t > 0$, akkor $\sqrt{2\pi}\,\varphi(t) < 1$, ezért

$$-2\ln\sqrt{2\pi}\,\varphi(t) > 0,$$

y-nak tehát mindig van ez esetben értelme, $\ln t$ is létezik. Ha logaritmikus skálájú t tengelyt veszünk, akkor a z értékre a gyakorlatban nem is kell áttérnünk. Ahhoz, hogy a papírboltokban vásárolható logaritmikus beosztású tengelyt használhassunk, 10-es alapú logaritmust kell alkalmazni. Ekkor

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} 10^{-\frac{t^2}{2}\lg e},$$

$$\lg\sqrt{2\pi}\,\varphi(t) = -\frac{t^2}{2}\lg e,$$

$$\lg u = \lg\left(-\frac{2}{\lg e}\lg\sqrt{2\pi}\,\varphi(t)\right) = 2\lg t,$$

ha $t > 0$. *10-es logaritmikus beosztású t tengelyt, 10-es logaritmikus beosztású u tengelyt alkalmazva, a $(t; u)$ koordináta-rendszerben a most kapott függvény képe egyenes lesz.*

Az $N(m, \sigma)$-eloszlású, $f(x)$ sűrűségfüggvényű X változót először standardizáljuk a

$$T = \frac{X - m}{\sigma}$$

transzformációval. Így sűrűségfüggvénye a $\varphi(t)$ függvény. Ez kifejezhető $f(x)$ segítségével.

A már tárgyalt

$$f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sigma}\varphi(t)$$

egyenletből:

$$\varphi(t) = \varphi\left(\frac{x-m}{\sigma}\right) = \sigma f(x).$$

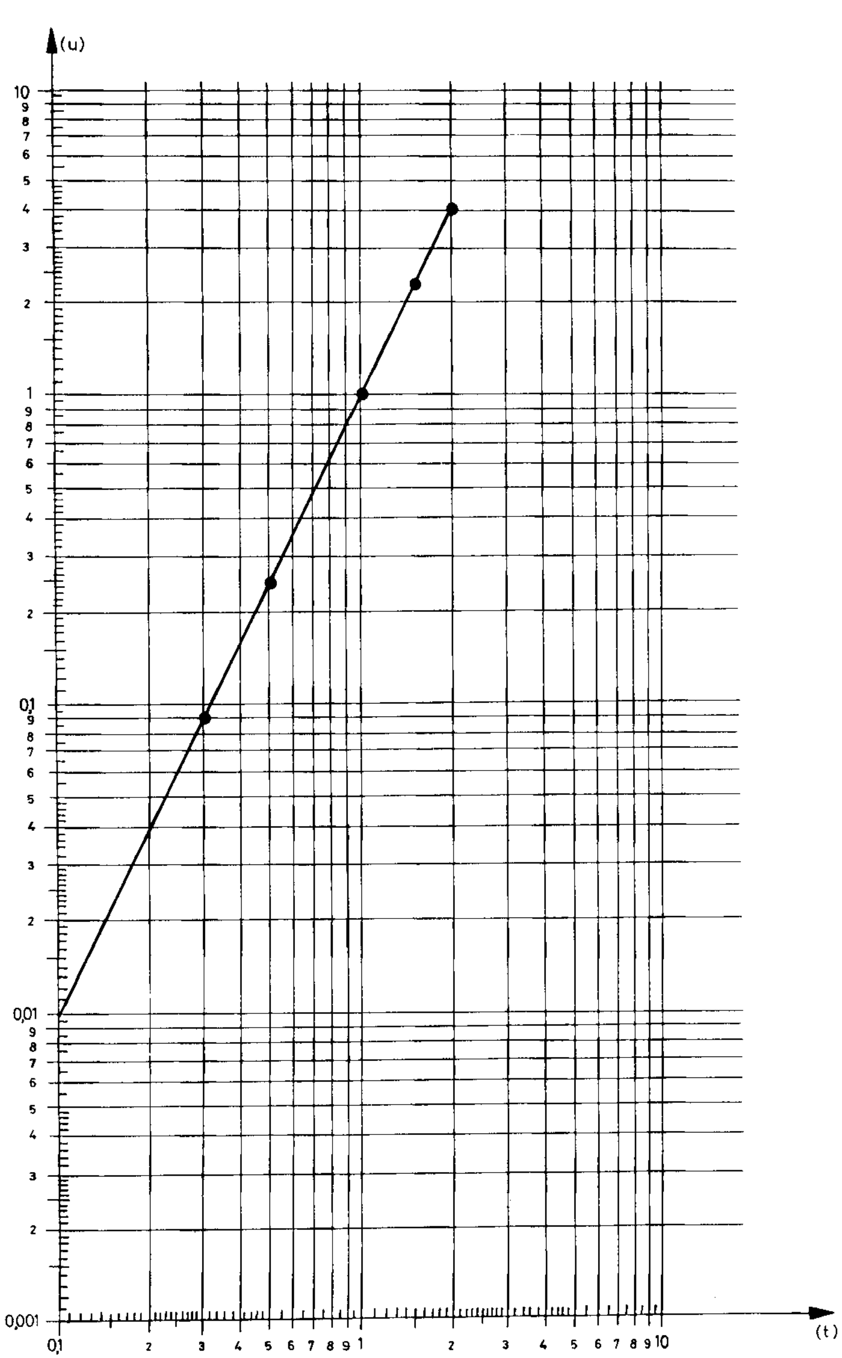
Ha a mintából m értékét $\bar{x}$-gal, σ értékét s^*-gal, $\sigma f(x)$ értékét $s^* f_n(x)$-szel becsüljük, akkor $\varphi(t)$ értékeit megkapjuk a mintabeli becslés alapján. Az így kapott becslésekkel az előbbi módszerrel ábrázolt pontok X normális eloszlása esetén közelítőleg egy egyenesen sorakoznak.

Példaként ábrázoljuk a mindkét tengelyen logaritmikus $(t; u)$ koordináta-rendszerben a

$$\lg u = \lg\left(-\frac{2}{\lg e}\lg\sqrt{2\pi}\,\varphi(t)\right) = 2\lg t$$

függvényt. A $\varphi(t)$ értékeket a normál eloszlás táblázatából vettük, a szükséges függvényértékeket számítógéppel számítottuk.

t	$\varphi(t)$	$u = -\frac{2}{\lg e}\lg\sqrt{2\pi}\,\varphi(t)$
0,10	0,3970	$9{,}760\,934\,8 \cdot 10^{-3}$
0,30	0,3814	$8{,}993\,611\,2 \cdot 10^{-2}$
0,50	0,3521	$2{,}498\,030\,9 \cdot 10^{-1}$
1,00	0,2420	$9{,}997\,581\,1 \cdot 10^{-1}$
1,50	0,1295	2,250 271 9
2,00	0,0540	3,999 655 6



38. ábra

A 38. ábrán a $(t;\ u)$ koordináta-rendszerben ábrázolva a kapott pontokat, látszik, hogy azok egy egyenesen sorakoznak.

3. Egy egyszerű vizsgálat az eloszlás exponenciális voltára

Ha a minta exponenciális eloszlású sokaságból származott, akkor az elméleti sűrűségfüggvény:

$$f(x)=\lambda e^{-\lambda x}, \quad \text{ha} \quad x>0, \quad \text{és} \quad f(x)=0 \text{ egyébként.}$$

Logaritmáljunk az $x>0$ esetben!

$$\ln f(x) = -\lambda x+\ln \lambda.$$

Az $y=\ln f(x)$ tengelyen logaritmikus beosztást alkalmazva, az $y = -\lambda x+\ln \lambda$ egyeneshez jutottunk. λ értékét $\frac{1}{\bar{x}}$-gal becsülhetjük.

Közelítsük az elméleti sűrűségfüggvényt a tapasztalataival, minden elemet helyettesítsünk annak az osztálynak a közepével, amelybe esett! Ha az eredeti eloszlás exponenciális, az így kapott pontok közelítőleg egy egyenesre esnek. Ha a vásárolható logaritmikus beosztású papírt akarjuk használni, 10-es logaritmust kell alkalmazni (mint előbb).

Megjegyzés

A két legutóbb ajánlott módszer közelítő, gyors tájékozódásra jó, hiszen mindkettőben különböző közelítéseket használtunk.

4. A χ^2-próba és alkalmazásai

Ha az $A_1, A_2, \ldots, A_r$ események kizárják egymást és teljes eseményrendszert alkotnak, akkor a

$$H_0\text{: } P(A_i)=p_i \quad \left(i = 1, 2, \ldots, r \quad \text{és} \quad \sum_{i=1}^{r} p_i=1\right),$$

nullhipotézis vizsgálatára az alábbi próba adható meg.

Tegyük fel, hogy n megfigyelésből az A_i esemény k_i-szer következett be ($i = 1, 2, \ldots, r$), világos, hogy $\sum_{i=1}^{r} k_i = n$. A k_i gyakoriságok binomiális eloszlású változók np_i várható értékkel, a $(k_1, k_2, \ldots, k_r)$ vektorváltozó polinomiális eloszlású, tehát H_0 fennállása esetén

$$P(k_1 = n_1, \ldots, k_r = n_r/H_0) = \frac{n!}{\prod_{i=1}^{r} n_i!} \prod_{i=1}^{r} p_i^{n_i},$$

ha $\sum_{i=1}^{r} n_i = n$, más esetben a valószínűség 0.

Adjuk meg a következő próbastatisztikát:

$$\chi^2 = \sum_{i=1}^{r} \frac{(k_i - np_i)^2}{np_i}.$$

A statisztika minden egyes tagja: a k_i gyakoriság négyzetes eltérése az np_i várható értéktől osztva a várható értékkel. *A χ^2 statisztika tehát azt jellemzi, hogy milyen arányú a mintabeli gyakoriság és a feltételezett elméleti várható érték eltérése.*

Bizonyítható, hogy *ha $n \to \infty$, akkor a próbastatisztika eloszlása tart a) az $r-1$ paraméterű χ^2-eloszláshoz (tiszta illeszkedésvizsgálat esetén), b) az $r-1-s$ paraméterű χ^2-eloszláshoz (becsléses illeszkedésvizsgálatnál, ha a becsült paraméterek száma s). A próbastatisztika tehát közelítőleg χ^2-eloszlású.*

A közelítés megfelelő, ha mindegyik $np_i \geqq 10$.

Az $1-\varepsilon$-szintű próbához a kritikus tartomány H_0-t feltételezve:

$$X_k = \{\chi^2 > \chi^2_{r-1}(\varepsilon)\} \quad \text{vagy} \quad X_k = \{\chi^2 > \chi_{r-1-s}(\varepsilon)\}.$$

A kritikus értékeket a 7. táblázat tartalmazza. A próba *egyoldali* próba, mivel az elfogadási tartományt úgy adjuk meg, hogy

$$P(\chi^2 \leqq \chi^2_{r-1}(\varepsilon)/H_0) = 1-\varepsilon \quad \text{(tiszta illeszkedés)},$$

vagy

$$P(\chi^2 \leqq \chi^2_{r-1-s}(\varepsilon)/H_0) = 1-\varepsilon \quad \text{(becsléses illeszkedés)}$$

legyen. *A χ^2-próbához elég nagy mintaelemszámra van szükség.* Ha például a legkisebb $p_i = 0{,}05$, akkor $np_i \geqq 10$ egyenletből:

$$n \geqq \frac{10}{0{,}05} = 200,$$

tehát a jó közelítéshez legalább 200 mérésre van szükség.

Az Olvasónak az eddigiekből úgy tűnhet, hogy a χ^2-próba csak diszkrét eloszlású változóra jó (hiszen r esemény bekövetkezési valószínűségeire tettünk hipotézist). *A χ^2-próba azonban folytonos eloszlású változó eloszlására tett feltevés vizsgálatára is alkalmas:* ha X folytonos eloszlású változóra van mintánk, akkor soroljuk azt r osztályba, minden változóértéket helyettesítsünk a megfelelő osztály közepével. Az A_i esemény az, hogy X egyenlő az i-edik osztályközéppel, a p_i valószínűség annak a valószínűsége, hogy X az i-edik osztályba esett.

Itt is szükséges, hogy a legkisebb np_i se legyen kisebb, mint 10. Ha ez nem teljesül, akkor ezen úgy próbálunk segíteni, hogy az eredetinél nagyobb terjedelmű osztályokba sorolva a p_i valószínűségeket növeljük. Ha így sem sikerül elérni, hogy minden $np_i \geqq 10$ legyen, más próbát alkalmazunk.

Megjegyzés

Azt találtuk, hogy *ha az illeszkedésvizsgálat becsléses, a próbastatisztika szabadságfoka a becsült paraméterek számával csökken. Ez más vizsgálatokra is érvényes.* Például már a kétmintás t-próbánál is az $n_1 + n_2$ szabadságfok 2-vel csökkent (mert az m és σ értékét a mintából becsültük). Ez érthető: a szereplő változók közt a mintabeli becslésből annyi egyenletünk van, ahány paramétert becslünk (ennyivel kevesebb a „szabad" változók száma).

a) Az illeszkedésvizsgálatot χ^2-próbával tulajdonképpen vázoltuk, diszkrét és folytonos változóra.

Megjegyezzük még a χ^2-próba egy hibáját. Ha az eloszlás „szélein" levő intervallumban a kelleténél több X_i érték fordul elő, akkor azt az egyszerű, hisztogramos eljárásnál nem vesszük figyelembe, a véletlennek tulajdonítjuk. A χ^2-próbánál azonban ez a $(k_i - np_i)^2$ túl nagy lesz a hozzátartozó np_i-hez képest, a χ^2-próba elutasításhoz vezet.

A statisztikában szokásos eljárás az „outlyer”, túlságosan kirívó adatok elhagyása („az adatok tisztítása”; Cleaning data). Például versenyeken gyakran elhagyják a legalacsonyabb és legmagasabb pontszámot. Persze vigyázni kell, nehogy az eljárással meghamisítsuk az eloszlást!

b) A homogenitásvizsgálat azt tárgyalja, hogy két valószínűségi változó azonos eloszlásúnak (két sokaság „homogénnek”, egynemű-nek) tekinthető-e. Ha a két véletlen változót X-szel, ill. Y-nal jelöljük, akkor a nullhipotézis:

$$H_0\colon\ P(X<x) = P(Y<x), \quad \text{vagyis} \quad F(x) \equiv G(x).$$

A χ^2-próbás homogenitásvizsgálatban az azonosnak feltételezett eloszlásfüggvény nem is szerepel. Osszuk fel a számegyenest a

$$-\infty = z_0 < z_1 < z_2 < \ldots < z_r = \infty$$

osztópontokkal r számú osztályra! Essen a (z_{i-1}, z_i) intervallumba az X változóra vett mintából f_i, az Y változóra vett mintából k_i $(i = 1, 2, \ldots, r)$. Legyen az X változóra vett minta elemszáma $\sum_{i=1}^{r} f_i = n$, az Y változóhoz tartozó minta elemszáma $\sum_{i=1}^{r} k_i = m$. Kimutatható, hogy a

$$\chi^2 = mn \sum_{i=1}^{r} \frac{\left(\dfrac{f_i}{n} - \dfrac{k_i}{m}\right)^2}{f_i + k_i}$$

próbastatisztika $n \to \infty$ és $m \to \infty$ esetén $r-1$ paraméterű, χ^2-eloszlású. Ha tehát m és n elég nagy, akkor χ^2-próba alkalmazható a homogenitás vizsgálatára.

c) Függetlenségvizsgálat χ^2-próbával. A függetlenségvizsgálat célja az, hogy eldöntsük: két valószínűségi változó függetlennek tekinthető-e, tehát igaz-e, hogy

$$H_0\colon\ P(X<x,\ Y<y) = P(X<x)P(Y<y),$$

vagy az eloszlásfüggvényekkel felírva fennáll-e, hogy

$$H_0\colon\ H(x, y) = F(x)G(y).$$

Legyen az (X, Y) párra vonatkozó n-elemű minta $(X_1, Y_1), (X_2, Y_2), \ldots$ $\ldots, (X_n, Y_n)$. Osszuk fel az X tengelyt a $-\infty = x_0 < x_1 < x_2 < \ldots$ $\ldots < x_r = \infty$ osztópontokkal r részintervallumra, az Y tengelyt $-\infty = y_0 < y_1 < y_2 < \ldots < y_s = \infty$ osztópontokkal s számú osztályra. Így a síkot rs „téglalapra” bontottuk.

$$\text{Az} \quad A_i = \{x_{i-1} \leqq X < x_i\} \quad \text{és a} \quad B_j = \{y_{j-1} \leqq Y < y_j\}$$

események:

– külön-külön való bekövetkezésének valószínűsége

$$p_i = F(x_i) - F(x_{i-1}),$$
$$q_j = G(y_j) - G(y_{j-1}),$$

– együttes bekövetkezéseinek (azaz egy kiszemelt téglalapba való eséseknek, az A_iB_j esemény bekövetkezéseinek) a száma k_{ij}.

A függetlenség eldöntésére szolgáló próbastatisztika (ha az $F(x)$ és $G(y)$ eloszlásfüggvények ismeretesek, *tiszta függetlenségvizsgálat*):

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(k_{ij} - np_iq_j)^2}{np_iq_j}.$$

Ez a próbastatisztika a nullhipotézis (vagyis a függetlenség) fennállása esetén $rs-1$ paraméterű, aszimptotikusan χ^2-eloszlású (a χ^2-próba a szokott módon elvégezhető).

A gyakorlatban sűrűbben fordul elő, hogy a p_i és q_j valószínűségeket nem ismerjük, hanem a mintából becsüljük *(becsléses függetlenségvizsgálat)*. Szokás ilyenkor az alábbi *kontingenciatáblázatot* készíteni (contingency = véletlenség, eshetőség). Miként a kétdimenziós (kétváltozós) eloszlásoknál tettük, a táblázat peremén a *vetületeloszlásoknak* megfelelő gyakoriságok találhatók:

$$f_{i.} \quad \text{az } A_i \text{ esemény gyakorisága,} \quad f_{i.} = \sum_{j=1}^{s} k_{ij},$$

$$f_{.j} \quad \text{a } B_j \text{ esemény gyakorisága,} \quad f_{.j} = \sum_{i=1}^{r} k_{ij},$$

$$P(A_i) = p_i \approx \frac{f_{i.}}{n}, \qquad P(B_j) = q_j \approx \frac{f_{.j}}{n}.$$

X	Y						Összesen
	B_1	B_2	...	B_j	...	B_s	
A_1	k_{11}	k_{12}	...	k_{1j}	...	k_{1s}	$f_{1.}$
A_2	k_{21}	k_{22}	...	k_{2j}	...	k_{2s}	$f_{2.}$
A_i	k_{i1}	k_{i2}	...	k_{ij}	...	k_{is}	$f_{i.}$
A_r	k_{r1}	k_{r2}	...	k_{rj}	...	k_{rs}	$f_{r.}$
Összesen:	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.s}$	n

Kimutatható, hogy a

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(k_{ij} - \frac{f_{i.}f_{.j}}{n}\right)^2}{f_{i.}f_{.j}}$$

próbastatisztika aszimptotikusan $(r-1)(s-1)$ paraméterű, χ^2-eloszlású, a szokásos χ^2-próba vele elvégezhető. A paraméterek száma most (a becsléses függetlenségvizsgálatnál):

$$rs-1-(r-1+s-1) = (r-1)(s-1),$$

hiszen a becsült paraméterek száma $(r-1)+(s-1)$, mert a valószínűségek összege mind az X, mind az Y változóra 1, és így például az első $(r-1)$ számú becsült p_i-ből az utolsó, p_r becslése már adódik; hasonlóan az első $(s-1)$ számú becsült q_j-ből az utolsó, q_s becslése már megkapható.

Átalakítással számításra alkalmasabb, egyszerűbb képletet nyerhetünk:

$$\chi^2 = n\left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{k_{ij}^2}{f_{i.}f_{j}} - 1\right).$$

Különösen egyszerű a statisztika, ha mindkét változóra $r=s=2$ mérésünk van. *A 2×2-es kontingenciatáblázat:*

X	Y		Összesen
	B_1	B_2	
A_1	k_{11}	k_{12}	$k_{11}+k_{12}=f_{1.}$
A_2	k_{21}	k_{22}	$k_{21}+k_{22}=f_{2.}$
Összesen:	$k_{11}+k_{21}=f_{.1}$	$k_{12}+k_{22}=f_{.2}$	$n=k_{11}+k_{12}+k_{21}+k_{22}$

A próbastatisztika:

$$\chi^2 = n\frac{(k_{11}k_{22}-k_{12}k_{21})^2}{f_{1.}f_{2.}f_{.1}f_{.2}}$$

aszimptotikusan χ^2-eloszlású, $(r-1)(s-1) = 1$ paraméterrel.

Mivel nem csupán két mennyiségi, hanem két minőségi ismérv függetlenségéről is beszélhetünk, *a problémát az $A_1, A_2, \ldots, A_r$ és $B_1, B_2, \ldots, B_s$ teljes eseményrendszerek függetlenségének az ellenőrzésére is megfogalmazzuk.* Nullhipotézis: a két eseményrendszer független, vagyis:

$$P(A_iB_j) = P(A_i)P(B_j) \qquad (i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, s).$$

Ha a $P(A_i)=p_i$ és a $P(B_j)=q_j$ elméleti valószínűségek ismertek, akkor az előzőekben ismertetett *tiszta függetlenségvizsgálatot* végezzük el.

Gyakoribb azonban, hogy nem az elméleti valószínűségek, hanem ezeknek mintabeli becslései ismeretesek, vagyis a

$$p_i \approx \frac{f_{i.}}{n} = \frac{\sum_{j=1}^{s} k_{ij}}{n}, \qquad q_j \approx \frac{f_{.j}}{n} = \frac{\sum_{i=1}^{r} k_{ij}}{n}$$

mennyiségek (l. az előző kontingenciatáblázatot). Ekkor az előzőekben bemutatott *becsléses illeszkedésvizsgálati χ^2-próbát* végezzük el.

5. Normalitásvizsgálat a mintaelemek transzformációja alapján *(Sarkadi-próba)*

Egy mintáról eldönthető, hogy normális eloszlásból származónak tekinthető-e anélkül, hogy a várható értéket vagy a szórást ismernénk. Sarkadi olyan transzformációt javasol, amely az eredeti mintaelemeket mint változókat független, $N(0, 1)$-eloszlású változókká transzformálja akkor és csak akkor, ha az eredeti változók normál eloszlásúak voltak. Így visszavezettük a feladatot olyan statisztikai minta vizsgálatára, amelynél a hipotézis az $N(0, 1)$-eloszlás (normalitásvizsgálat ismert paraméterekkel). Ennél az ún. tiszta illeszkedésvizsgálatnál több próba, például a bemutatandó Kolmogorov-próba is jól alkalmazható. A feladat több kismintára (különböző várható értékekre és szórásokra) is elvégezhető, a transzformáció után kapott $N(0, 1)$-eloszlású kismintákat egyesíthetjük egyetlen $N(0, 1)$-eloszlású nagymintában.

Legyenek a mintaelemek (mintavétel szerinti sorrendben) $X_1, X_2, \ldots, X_n$, végezzük el a következő transzformációt:

$$Y_i = \frac{X_i - \bar{X}'}{S}\,\psi_{n-2}\left(\frac{|X_{n-1} - X_n|}{S}\right), \qquad (i = 1, 2, \ldots, n-2),$$

itt az egyes változók értelme ($\bar{X}$-sal jelölve a $\frac{\sum X_i}{n}$ mintaátlagot):

$$\bar{X}' = \frac{\bar{X} + \sqrt{\frac{n}{2}}(X_{n-1} + X_n)}{n + \sqrt{2n}},$$

$$S^2 = \frac{2}{n-2}\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\bar{X}^2 - \frac{1}{2}(X_{n-1} - X_n)^2\right].$$

Az $\psi_{n-2}(x)$ függvény értékét az $n-2$ paraméterű Student-, ill. χ^2-eloszlású változó eloszlásfüggvényéből határozzuk meg, a következőképpen. Tekintsük az f paraméterű, Student-eloszlású változó $P(t/f)$ eloszlásfüggvényét:

$$P(t/f) = \frac{1}{\sqrt{f\pi}}\,\frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)}\int_{-\infty}^{t}\frac{du}{\left(1 + \frac{u^2}{f}\right)^{\frac{f+1}{2}}},$$

és tekintsük az f paraméterű χ^2-eloszlású véletlen változó $1 - Q(t/f)$ eloszlásfüggvényét, tehát a

$$Q(t/f) = \frac{1}{2^{\frac{f}{2}}\Gamma\left(\frac{f}{2}\right)}\int_{t}^{\infty} u^{\frac{f-1}{2}} e^{-\frac{u}{2}}\,du$$

függvényt!

Ezekből határozható meg a keresett $\psi_{n-2}(t)$, visszakereséssel az alábbi összefüggésből:

$$Q([\psi_{n-2}(t)]^2/n-2) = 2P(t/n-2) - 1 \qquad (t \geq 0).$$

A könyvben található χ^2 és Student-eloszlás táblázatok ehhez a próbához általában nem elégségesek. Vannak olyan statisztikai munkák, amelyekben a kellő számú táblázat megtalálható, például *Sarkadi, K.*: On testing for normality. MTA Mat. Int. Közl. V. A. 3. Budapest, 1960.

A leírt transzformációval az eredetinél 2-vel kevesebb $Y_1, Y_2, \ldots, Y_{n-2}$ független változókhoz jutunk, amelyek akkor és csak akkor $N(0, 1)$-eloszlású változók, ha az eredeti minta normális eloszlásból származott.

6. Hipotézisvizsgálat az eloszlás exponenciális voltára *(Störmer-próba)*

Tekinthető-e az $X_1, X_2, \ldots, X_n$ független mintaelemekből álló minta exponenciális eloszlásból származónak? Az exponenciális eloszlás paramétere nem ismert. Az

$$X_1^* \leq X_2^* \leq \ldots \leq X_n^*$$

rendezett mintaelemekhez vegyük hozzá még a $X_0^*=0$ elemet, és végezzük el a következő transzformációt:

$$Y_k^* = \frac{\sum_{i=0}^{k-1} (n-i)(X_{i+1}^* - X_i^*)}{\sum_{j=1}^{n} X_j} \qquad (k = 1, 2, \ldots, n-1).$$

Ha az eredeti eloszlás exponenciális, akkor az így kapott (az eredetinél eggyel kevesebb számú) Y_k^* változók úgy tekinthetők, mint a (0, 1) intervallumban egyenletes eloszlású változóból vett $n-1$-elemű minta. Erre a tiszta illeszkedésvizsgálatra például a Kolmogorov-próba alkalmazható, a kombinált módszer minden folytonos ellenhipotézisre aszimptotikusan konzisztens.

7. Tiszta vagy becsléses illeszkedésvizsgálat nagy mintaelemszám esetében, *a Kolmogorov-féle (egymintás) próba*

Az X folytonos eloszlású változóra tett nullhipotézis:

$$H_0\colon P(X<x) = F(x).$$

A kétoldali ellenhipotézis:

$$H_1\colon P(X<x) \not\equiv F(x).$$

Ha az $X_1, X_2, \ldots, X_n$ minta empirikus eloszlásfüggvénye $F_n(x)$, akkor a próbastatisztika:

$$D_n = \max_{-\infty<x<\infty} |F_n(x) - F(x)|.$$

Ha H_0 igaz, akkor

$$\lim_{n\to\infty} P(\sqrt{n}\, D_n < z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2z^2} = K(z) \qquad (z>0).$$

A Kolmogorov-féle $K(z)$ függvény segítségével:

$$P\left(D_n < \frac{z}{\sqrt{n}}\right) \approx K\left(\frac{z}{\sqrt{n}}\right),$$

így megszerkesztették a 11. táblázatot, amelynek használata egyszerű: *ha* $D_n = \max_{x\text{-re}} |F_n(x) - F(x)|$ *nagyobb, mint a táblázatban 95%-os* (ill. 99%-os) *biztonsági szintre megadott kritikus érték, akkor elutasítjuk azt a hipotézist, hogy a valószínűségi változó eloszlásfüggvénye $F(x)$; ellenkező esetben elfogadjuk.*

Például ha $n=25$, $D_n=0{,}8$ (a feltételezett elméleti és tapasztalati eloszlásfüggvény maximális eltérése 0,8), akkor a táblázatból az alábbi részletet kell figyelni:

n	0,95	0,99
25	0,2640	0,3166

A számított 0,8 eltérés sokkal nagyobb 0,3166-nál is, 0,2640-nél is, visszautasítjuk azt, hogy a sokaság eloszlásfüggvénye valóban a feltételezett $F(x)$ volna: $P(X<x) \not\equiv F(x)$. *(A táblázat értékei kétoldali ellenhipotézisre vonatkoznak.)*

A Kolmogorov-próba az $F_n(x)$ és $F(x)$ közötti lényeges eltérésre erősebben reagál, mint a χ^2-próba, ilyen esetekben gyakrabban vezet a nullhipotézis elvetésére. Másrészről azonban előfordulhat, hogy a tapasztalati eloszlás lényegesen és szisztematikusan különbözik a felvett elméleti eloszlástól, de a D_n eltérés mégsem elég nagy ahhoz, hogy a nullhipotézis elvetésére vezetne.

Megjegyzés

Beszéltünk arról, hogy Szmirnov tétele szerint a

$$D_n^+ = \sup_{-\infty<x<\infty} (F_n(x) - F(x))$$

statisztikával:

$$\lim_{n\to\infty} P(\sqrt{n}\, D^+ < z) = 1 - e^{-2z^2} = S(z), \qquad (z \geqq 0).$$

A $\sqrt{n}\,D^+$ eloszlását elég nagy n-re közelítőleg a Szmirnov-féle $S(z)$ függvény határozza meg:

$$P(\sqrt{n}\,D^+ < z) \approx 1 - e^{-2z^2}.$$

Ha a nullhipotézis

$$H_0\colon P(X<x) = F(x),$$

és az egyoldali ellenhipotézis

$$H_1\colon P(X<x) < F(x),$$

akkor a H_0 elfogadására vagy elutasítására elegendő a $\sqrt{n}\,D^+$ statisztika alapján $1-\varepsilon$ szinten olyan z_0 kritikus értéket meghatározni, hogy

$$e^{-2z_0^2} = \varepsilon$$

legyen, amiből

$$z_0 = \sqrt{-\frac{1}{2}\ln\varepsilon}.$$

A kritikus tartomány a számegyenes z_0-nál nagyobb értékeket tartalmazó szakasza.

A $\sqrt{n}\,D^+$ statisztika sűrűségfüggvénye:

$$\frac{dS}{dz} = 4ze^{-2z^2},$$

így várható értéke:

$$M(\sqrt{n}\,D^+) = -\int_0^\infty z(-4ze^{-2z^2})dz = [-ze^{-2z^2}]_0^\infty + \int_0^\infty e^{-2z^2}dz =$$

$$\begin{array}{c|c} u & v' \\ \hline u'=1 & v=e^{-2z^2} \end{array}$$

$$= \frac{\sqrt{2\pi}}{4} \approx 0{,}625.$$

8. A Kolmogorov–Szmirnov-féle kétmintás próba

Ez homogenitásvizsgálatra használható, segítségével arról döntünk, hogy két folytonos eloszlású X és Y véletlen változó eloszlása azonos-e. Az eloszlásfüggvények:

$$P(X<x) = F(x) \quad \text{és} \quad P(Y<x) = G(x),$$

így a nullhipotézis most:

$$H_0\colon F(x) \equiv G(x).$$

Az X-re vett

$$X_1, X_2, \ldots, X_n$$

minta és az Y-ra vett

$$Y_1, Y_2, \ldots, Y_m$$

minta rendezésével készítsük el az $F_n(x)$, ill. a $G_m(x)$ tapasztalati eloszlásfüggvényeket! Szmirnov kimutatta, hogy

$$\lim_{n,m\to\infty} P\left\{\sqrt{\frac{nm}{n+m}}\max_{x\text{-re}}[F_n(x)-G_m(x)] < z/H_0\right\} = 1 - e^{-2z^2-4cz},$$

ahol $c = \dfrac{m-n}{2\sqrt{m+n}}$, továbbá

$$\lim_{n,m\to\infty} P\left(\sqrt{\frac{nm}{n+m}}\max_{x\text{-re}}|F_n(x)-G_m(x)| < z/H_0\right) =$$

$$= K(z) = \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2z^2},$$

ahol $z>0$.

– A $H_1\colon F(x) > G(x)$ *egyoldali ellenhipotézis* esetén a próbastatisztika:

$$D^+_{n,m} = \max_{(x)}[F_n(x)-G_m(x)],$$

a kritikus tartomány $1-\varepsilon$ szinten a $\{D_{n,m} > D_\varepsilon\}$ tartomány, ahol D_ε értékét a

$$P(D^+_{n,m} < D_\varepsilon / H_0) = 1-\varepsilon$$

relációból kaphatjuk meg.

– A H'_1: $F(x) \not\equiv G(x)$ *kétoldali ellenhipotézissel* szemben a próbastatisztika:

$$D_{n,m} = \max_{(x)} |F_n(x) - G_m(x)|,$$

a kritikus tartomány $1-\varepsilon$ szinten a $\{D_{n,m} > D'_\varepsilon\}$ tartomány, ahol D'_ε értékét a

$$P(D_{n,m} < D'_\varepsilon / H_0) = 1-\varepsilon$$

összefüggésből határozhatjuk meg. Nagy n és m esetén használható a $K(z)$ függvény táblázata (l. a 14. táblázatot).

Az alábbi táblázatrészlet azt mutatja, hogy nagy, kb. 95% valószínűséggel:

$$\sqrt{\frac{nm}{n+m}} \max_{x\text{-re}} |F_n(x) - G_m(x)| < 1{,}36.$$

z	$K(z)$
1,35	0,9477
1,36	0,9505

Ha a próbastatisztika számított értéke 1,36-nál nagyobb, a nullhipotézist elvetjük; ellenkező esetben elfogadjuk.

Egyenlő mintadarabszámok ($m=n$) esetén használható a 12. táblázat 95% és 99% szinteken, amely az $n\,D'_\varepsilon$ értékeket adja meg, ezeket az $nD_{n,n}$ értékekkel kell összehasonlítani.

Ha $n=m$ mintadarabszám esetén $nD^+_n = n \max |F_n(x) - G(x)|$ meghaladja a táblázatban található kritikus értéket, akkor elutasítjuk azt a feltevést, hogy a két véletlen változó eloszlása megegyezik; ellenkező esetben elfogadjuk. A táblázat értékei kétoldali ellenhipotézisre vonatkoznak.

Például $n=m=7$, 95% biztonsági szint esetén a 12. táblázatbeli kritikus érték $nD'_\varepsilon = 6$:

n	0,95	0,99
7	6	6

Ha most az $F_7(x)$ és $G_7(x)$ függvények maximális abszolút eltérése $\frac{4}{7}$, akkor

$$nD_{n,n} = 7D_{7,7} = 7 \cdot \frac{4}{7} = 4,$$

ez kisebb a kritikus értéknél, a H_0 hipotézist elfogadjuk. Feltehető, hogy a két minta azonos eloszlásból származik (homogén).

Megjegyzés

Egyoldali ellenhipotézis sokszor merül fel a gyakorlatban a következő esetben. Egy alkatrész X élettartamának eloszlását $F(x)$ eloszlásfüggvény adja meg. Valamilyen eljárással az élettartam növelését akarjuk elérni. Ha az új eljárással készült alkatrészek Y élettartamának eloszlásfüggvénye $G(x)$, akkor az élettartam növekedését az jelenti, hogy

$$P(Y>x) > P(X>x),$$

$$1-G(x) > 1-F(x),$$

$$F(x) > G(x).$$

9. Wilcoxon-próba

A próbát két eloszlás egyezésének vizsgálatára használják. Kidolgozták arra az esetre is, amikor két várható érték egyezése a nullhipotézis (H_0: $M(X)=M(Y)$), vagy amikor azt feltételezzük, hogy X átlagosan ugyanannyiszor kisebb Y-nál, mint ahányszor nagyobb

$$\left(H_0\colon P(X<Y) = P(X>Y) = \frac{1}{2}\right).$$

Tekintsük most az általános esetet, amikor a nullhipotézis az X változó $F(x)$ és az Y változó $G(x)$ eloszlásfüggvényeinek az azonossága, vagyis

$$H_0\colon F(x) \equiv G(x).$$

Rendelkezésünkre áll az

$$X_1, X_2, \ldots, X_n,$$

ill. az

$$Y_1, Y_2, \ldots, Y_m$$

minta.

Egyesítsük a két minta elemeit, és a kapott $m+n$-elemű minta elemeit rendezzük növekvő sorrendbe! Legyen az egyesített mintában az $X_1, X_2, \ldots, X_n$ elemek sorszáma (rangja) $r_1, r_2, \ldots, r_n$, ugyanakkor az $Y_1, Y_2, \ldots, Y_m$ elemek rangja $s_1, s_2, \ldots, s_m$.

Adjuk össze az X változóra vett minta elemeinek rangjait az egyesített, rendezett mintában. $\sum_{i=1}^{n} r_i$ legkisebb értékét akkor kapjuk, ha az első minta elemei (az X_i értékek) a közös (egyesített) minta elemeinek növekvő sorrendjében az első n helyet foglalják el,

$$\min \sum_{i=1}^{n} r_i = 1+2+\ldots+n = \frac{n(n+1)}{2}.$$

Itt $\sum r_i$ legnagyobb értékét nyilván akkor kapjuk, ha az első minta elemei a közös sorrend utolsó n helyét foglalják el:

$$\max \sum_{i=1}^{n} r_i = (m+1)+(m+2)+\ldots+(m+n) = \frac{(2m+n+1)n}{2}.$$

Wilcoxon a következő próbastatisztikát vizsgálta:

$$W_{n,m} = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2}.$$

A W statisztika pontos eloszlására zárt formula nem ismeretes. Mindenesetre meg tudunk állapítani néhány fontos dolgot:

– a W legkisebb értéke 0 $\left(\text{mivel } \sum r_i \text{ legkisebb értéke } \frac{n(n+1)}{2}\right)$;

– a W legnagyobb értéke:

$$\max \sum r_i - \frac{n(n+1)}{2} = \frac{(2m+n+1)n}{2} - \frac{n(n+1)}{2} =$$
$$= \frac{n \cdot 2m}{2} = mn;$$

– a nullhipotézis fennállása esetén W várható értéke:

$$M(W_{n,m}) = \frac{mn}{2},$$

szórásnégyzete:

$$D^2(W_{n,m}) = \frac{mn(m+n+1)}{12};$$

– a W véletlen változó eloszlása a várható értékre $\left(\frac{mn}{2}\text{-re}\right)$ szimmetrikus,

– W eloszlására érvényes a következő rekurziós formula:

$$P(W_{n,m}=k) = \frac{m}{m+n} P(W_{n-1,m}=k) + \frac{n}{m+n} P(W_{n,m-1}=k);$$

– ha $m \to \infty$, $n \to \infty$, akkor a standardizált W érték:

$$\frac{W_{n,m} - \dfrac{mn}{2}}{\sqrt{\dfrac{mn(m+n+1)}{12}}}$$

határeloszlása az $N(0, 1)$-eloszlás.

Ez utóbbi alapján:

– nagy n-re az u-próba jó közelítésben használható (nem is kell standardizálni, mert W jó közelítésben normál eloszlású, $M(W)$ várható értékkel, $D(W)$ szórással);

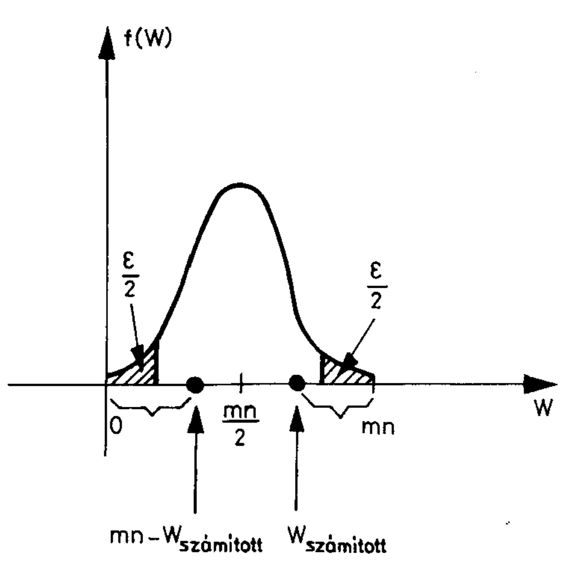
– kis m-ekre, kis n-ekre (5-től 20-ig) az előbbi rekurzív formula segítségével elkészítettük a W-statisztika kritikus értékeit 95%-os, ill. 99%-os kétoldali szintre (13. táblázat). Az elfogadási tartomány:

$$W_{\frac{\varepsilon}{2}} \leqq W \leqq W_{1-\frac{\varepsilon}{2}},$$

a táblázatban csak az alsó, $W < \dfrac{mn}{2}$ tartományhoz tartozó $W_{\frac{\varepsilon}{2}}$ kritikus érték található. Ha azt találjuk, hogy $W_{\text{számított}} < \dfrac{mn}{2}$, akkor magát $W_{\text{számított}}$ értéket, ellenkező esetben $(mn - W_{\text{számított}})$ értéket hasonlítjuk össze a táblázatban található $W_{\frac{\varepsilon}{2}}$ kritikus értékkel (39. ábra).

Ha a próbastatisztika számított értéke a kritikus érték alatt marad, akkor a két eloszlásfüggvény egyezésére vonatkozó nullhipotézist elvetjük; ellenkező esetben elfogadjuk. Az ábrán az elfogadás esetét szemléltettük.

Például H_0: $F(x) \equiv G(x)$, H_1: $F(x) \not\equiv G(x)$ (kétoldali ellenhipotézis), $n = 14$, $m = 20$, $\sum_{i=1}^{n} r_i = 110$,



39. ábra

$$W_{n,m} = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2} = 110 - 105 = 5,$$

mivel ez a felezőponttól balra van $\left(W_{\text{számított}} = 5 < \dfrac{nm}{2} = 140\right)$, ezért magát $W_{\text{számított}}$ értéket hasonlítjuk össze a táblázatbeli, 95%-os szinthez tartozó $W_{\frac{\varepsilon}{2}}$ értékkel:

	m
$n = 14$	20
	83

Mivel $W_{\text{számított}} = 5 < W_{\text{krit}} = 83$, a két eloszlásfüggvény azonosságára vonatkozó nullhipotézist 95%-os biztonsági szinten elvetjük.

Megjegyzés

A nagyság szerinti sorrendbe állításnál egyenlő elemek is előfordulhatnak, mégpedig:

a) ugyanazon mintán belül van mondjuk két azonos X_i érték;

b) a két azonos érték közül egyik az X, a másik az Y változóra vett mintában van.

A *b)* esetben ún. *átlagos rangot* rendelünk az egyenlőkhöz, az átlagos rang a közös mintában a megegyező elemek rangszámának a számtani közepe. Ha például a közös minta 10-edik és 11-edik eleme azonos, $X^*_{10}=Y^*_{11}$, akkor mindkettő 10,5 rangot kap. Ha például az egyesített, rendezett mintában 3 elem egyenlő, $X^*_{10}=X^*_{11}=Y^*_{12}$, akkor a közös rangszám 10, 11, 12 számtani közepe, vagyis 11 (mindegyik 11-es rangszámot kap).

Az *a)* esetben nincs szükség átlagos rang megállapításra, az egyenlő elemeket tetszés szerinti sorrendben, folytatólag rangszámozzuk. Pl. $X^*_{10}=X^*_{11}$, egyik kapja a 10-es, másik a 11-es sorszámot.

c) A próbát a

$$H_0\colon\ P(X<Y)=P(X>Y)=\frac{1}{2}$$

nullhipotézissel szemben tett

$$H_1^+\colon\ P(X<Y)>\frac{1}{2}$$

egyoldali ellenhipotézis vizsgálatára is alkalmazzák *(egyoldali próba)*, vagy pedig a

$$H_1\colon\ P(X<Y)\neq P(X>Y)$$

ellenhipotézis vizsgálatára *(kétoldali próba)*.

Gyakorló feladatok

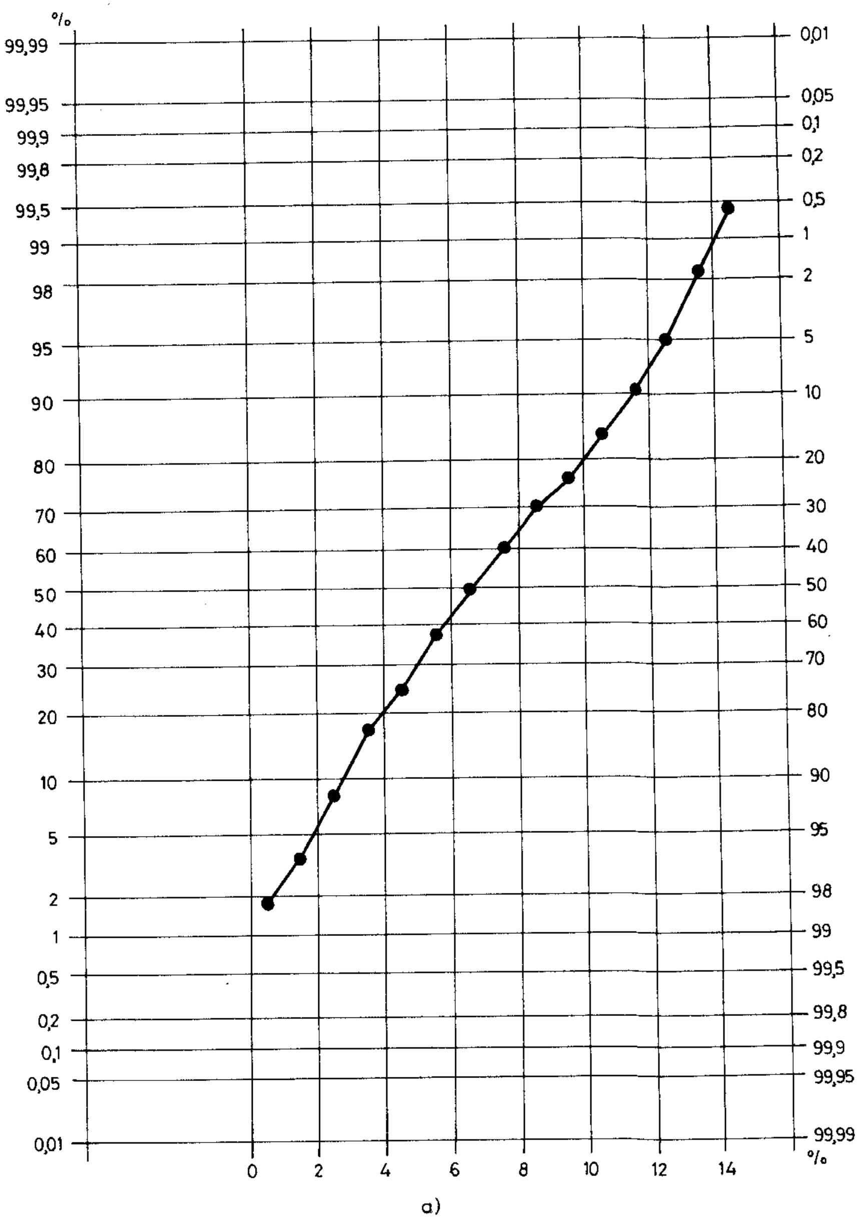
1. Kúpgörgős csapágynál az ún. nagyhomlok maximumának a névleges mérettől való eltérésére $n=170$ mérést végeztek. Az értékeket $(-0{,}5;\,0{,}5)$, $(0{,}5;\,1{,}5)$, ..., $(14{,}5;\,15{,}5)$ osztályba sorolták.

Osztályközép	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Gyakoriság	3	3	8	15	12	22	20	19	17	10	14	10	8	6	2	1

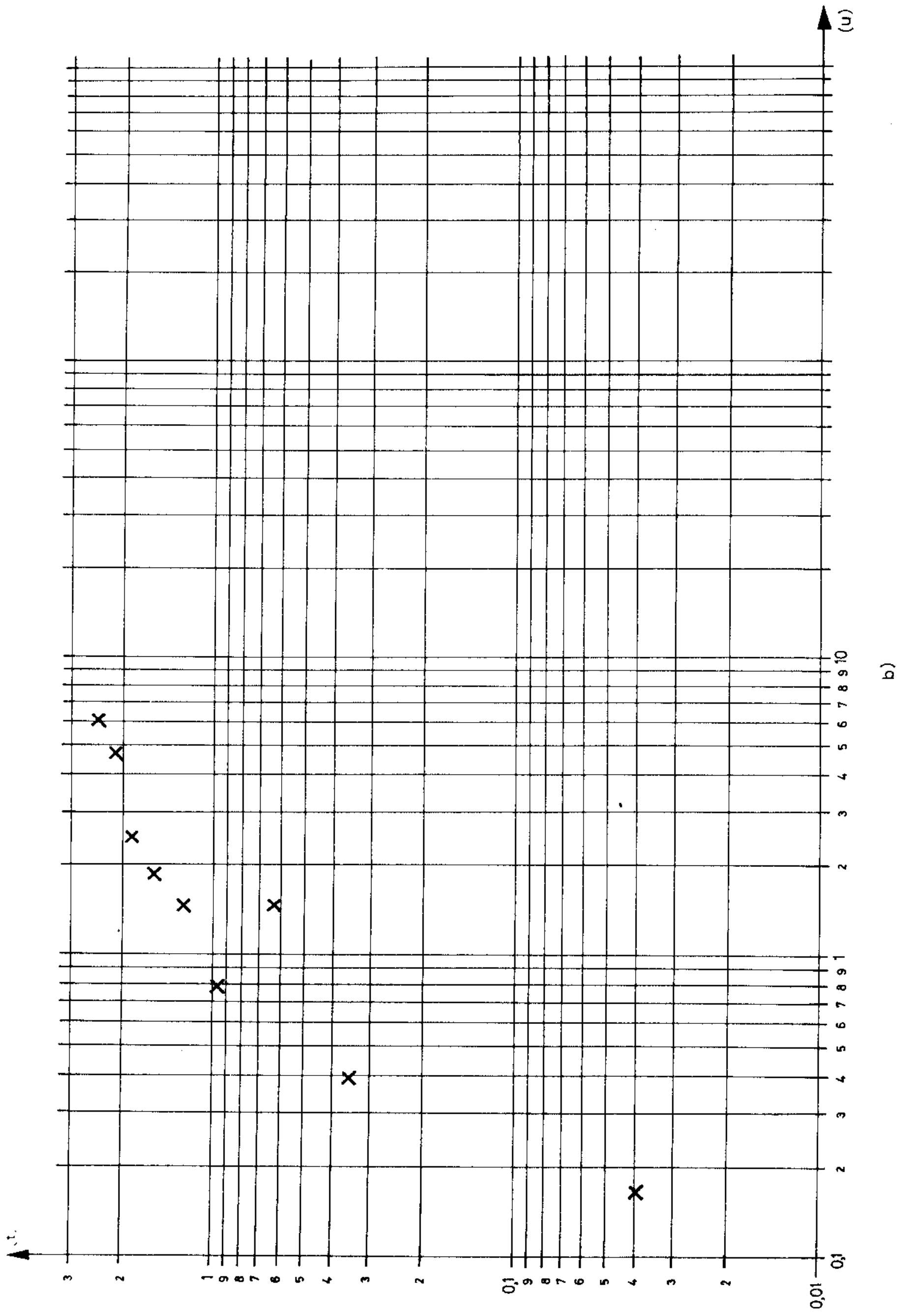
Feltehető-e, hogy az előírt mérettől való eltérések normális eloszlásúk? Döntse el *a)* Gauss-papíron való ábrázolással; *b)* a 2. pontban ajánlott egyszerű grafikus vizsgálattal; *c)* χ^2-próbával!

a) A gyakoriságokból kiszámítva a relatív gyakoriságokat, ezeket összegezve, $F(x)$ becslését kapjuk. Az $(x;\,F(x))$ pontokat Gauss-papíron ábrázoljuk (40*a* ábra). A felvitt pontok nagyjából egy egyenesre esnek, a normalitás feltehető.

Osztályközép	Gyakoriság	Relatív gyakoriság	Összegezett relatív gyakoriság
0	3	0,0176	0,0176
1	3	0,0176	0,0352
2	8	0,0471	0,0823
3	15	0,0882	0,1705
4	12	0,0706	0,2411
5	22	0,1294	0,3705
6	20	0,1177	0,4882
7	19	0,1118	0,6000
8	17	0,1000	0,7000
9	10	0,0588	0,7588
10	14	0,0824	0,8412
11	10	0,0587	0,8999
12	8	0,0471	0,9470
13	6	0,0353	0,9823
14	2	0,0118	0,9941
15	1	0,0059	1,000
Összesen:	170		



40*a* ábra



40*b* ábra

b) A mérési eredményekből becsüljük:

- a várható értéket: $m \approx \bar{x} = 6{,}8706$,
- a szórást: $\sigma \approx s^* = 3{,}29$.

Ezek után minden értéket az x_i osztályközepekkel helyettesitve:

- standardizáljuk a változót $\left(t_i = \dfrac{x_i - m}{\sigma}\right)$;
- az $f(x)$ sűrűségfüggvényt becsüljük a tapasztalati sűrűségfüggvénnyel $\left(f(x_i) \approx \dfrac{f_i}{nk}\right.$, ahol f_i a gyakoriság, n az összes mérés száma, k az osztályok terjedelme);
- kiszámítjuk az u_i értékeket:

$$u_i = -\frac{2}{\lg e} \lg \sqrt{2\pi}\sigma f(x_i).$$

Nézzük a $t_i > 0$ abszcisszájú pontokat. A (t_i, u_i) pontokat egy (mindkét tengelyen logaritmikus beosztású) papíron ábrázolva; azok nagyjából egy egyenesbe esnek, a normalitás feltehető (40*b* ábra).

x_i	$x_i - m$	$t_i = \frac{x_i - m}{\sigma}$	$f(x_i) \approx \frac{f_i}{nk}$	$u_i = -\frac{2}{\lg e} \lg \sqrt{2\pi}\,\sigma f(x_i)$
0	−6,8706			
1	−5,8706			
2	−4,8706			
3	−3,8706			
4	−2,8706			
5	−1,8706			
6	−0,8706			
7	0,1294	$3{,}933 \cdot 10^{-2}$	19/340	$1{,}631 \cdot 10^{-1}$
8	1,1294	$3{,}433 \cdot 10^{-1}$	17/340	$3{,}855 \cdot 10^{-1}$
9	2,1294	$6{,}472 \cdot 10^{-1}$	10/340	1,447
10	3,1294	$9{,}512 \cdot 10^{-1}$	14/340	$7{,}738 \cdot 10^{-1}$
11	4,1294	1,255	10/340	1,447
12	5,1294	1,559	8/340	1,893
13	6,1294	1,863	6/340	2,468
14	7,1294	2,167	2/340	4,666
15	8,1294	2,471	1/340	6,052

A számítások végzéséhez BASIC programot mellékelünk. A bemenő adatok:

- az adatok száma: n,
- az osztályok száma: r,
- a legkisebb osztályközép: x_1,
- az osztályok terjedelme: k,
- az osztályokba esés gyakorisága: f_i,
- a σ szórás becslése,
- az m várható értékbecslése.

Számítandók:

- a t_i értékek $\left(x_i = x_1 + (i-1)k,\ t_i = \dfrac{x_i - m}{\sigma}\right)$,
- az u_i értékek $\left(f(x_i) \approx \dfrac{f_i}{nk},\ u_i = -\dfrac{2}{\lg e} \lg \sqrt{2\pi}\sigma f(x_i)\right)$.

A folyamatábra a 41. ábrán látható.

Mivel a COMMODORE 64-hez tartozó LOG(Z) természetes alapú logaritmust jelent,

$$\mathrm{LOG}(Z) = \ln Z,$$

ezért

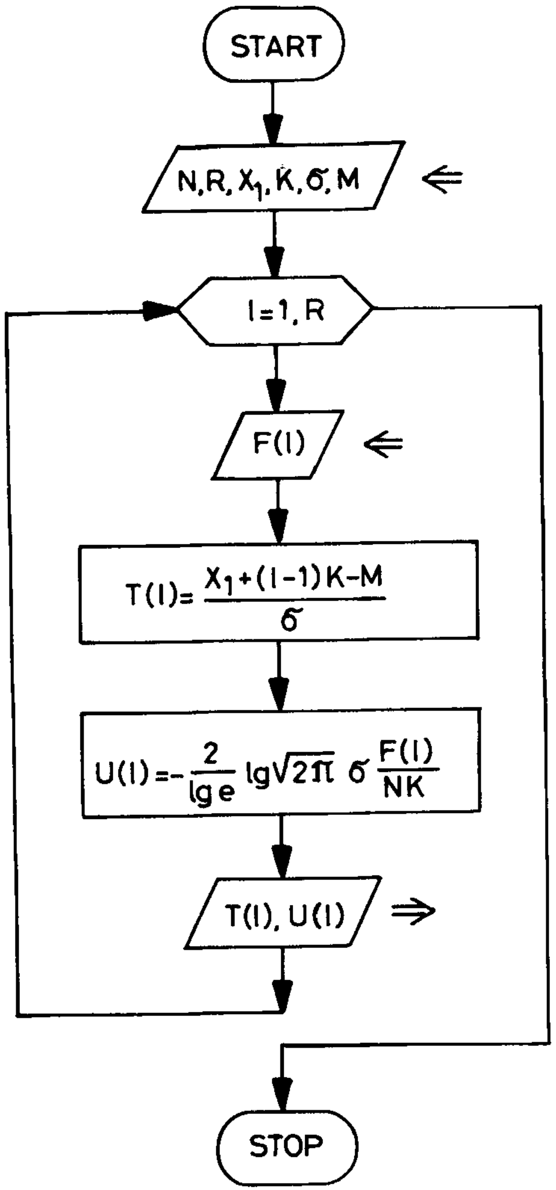
$$-\frac{2}{\lg e} = -\underbrace{\frac{2 \ln 10}{\ln e}}_{1} = -2 \ln 10,$$

$$u_i = -\frac{2}{\lg e} \lg \sqrt{2\pi}\sigma f(x_i) = -2 \ln 10 \cdot \frac{1}{\ln 10} \ln (\sqrt{2\pi}\sigma f(x_i)) =$$

$$= -2 \ln (\sqrt{2\pi}\sigma f(x_i)).$$

A program:

```
 5 REM GRAFIKUS NORMALITAS VIZSGALAT
10 INPUT "N, R, X1, K, SZ, M="; N, R, X1, K, SZ, M
15 DIM F(R)
20 FOR I=1 TO R
30 PRINT "F"I"=": INPUT F(I)
40 T(I)=(X1+(I-1)*K-M)/SZ
50 U(I)=-2*LOG(SQR(2*π)*SZ*F(I)/(N*K))
60 PRINT "T";I;"=";T(I)
70 PRINT "U";I;"="U(I)
80 NEXT I
90 END
```



41. ábra

Megjegyzés

Eljárásunk tulajdonképpen az, hogy az X változót standardizálva, a minta sűrűségfüggvényét hasonlítjuk össze az „etalonnal", a $\varphi(x)$ függvénnyel. Csak most olyan papíron tesszük ezt, amelyen mindkét függvény képe egyenes.

Az elveszített $t_i<0$ abszcisszájú pontokat is felhasználhatnánk. Nézzük ugyanis a bizonyítás következő lépését:

$$\lg\sqrt{2\pi}\,\varphi(t) = -\frac{t^2}{2}\lg e,$$

ebből

$$-\frac{2}{\lg e}\lg\sqrt{2\pi}\,\varphi(t) = t^2,$$

$$\lg\left(-\frac{2}{\lg e}\lg\sqrt{2\pi}\,\varphi(t)\right) = \lg t^2,$$

ha $t<0$, akkor $\lg t^2 = 2\lg(-t)$, a

$$\lg u = \lg\left(-\frac{2}{\lg e}\lg\sqrt{2\pi}\,\varphi(t)\right) = 2\lg(-t),$$

a most kapott egyenes egybeesik a pozitív t értékekre kapottal. Egyszerűen arról van szó, hogy ha a sűrűségfüggvény tükrös a $t=0$ egyenesre, elég csak az egyik felét használni.

Ha viszont $\varphi(t)$-t a mintából becsüljük a

$$\varphi(t) = \sigma f(x_i)$$

képlet felhasználásával, a kapott $(-t_i;\,u_i)$ pontokat is felhasználhatjuk, még több pontnak kell egyenesen sorakoznia.

A kapott grafikus eljárás mutatja, hogy az első $(3{,}933\cdot 10^{-2};\ 1{,}631\cdot 10^{-1})$ pont az, amelyik nagyon „kilóg a sorból": egyszerűen azért, mert normalitás esetén a középső intervallumban várjuk a legtöbb adatot. Az ide eső 19 adat kevés, ha több adat esik ebbe az osztályba, csökken a pont u_i ordinátája, a pont kevésbé tér el az egyenestől.

Mind a hisztogramos, mind a két, mutatott grafikus eljárás elsődleges, gyors tájékozódásra való, a döntés kissé szubjektív.

c) Bemutatjuk most az objektívebb χ^2-próbát. Az illeszkedésvizsgálat becsléses, a várható érték becslése $m\approx\bar{x}=6{,}8706$, a szórás becslése $\sigma\approx s^*=3{,}29$.

A nullhipotézis:

$$H_0\colon P(X<x) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{x} e^{-\frac{(z-m)^2}{2\sigma^2}}\,dz = F(x),$$

az ellenhipotézis:

$$H_1\colon P(X<x) \neq F(x)$$

(vagyis a nullhipotézis szerint az X eloszlása normális, az ellenhipotézis szerint nem az).

A χ^2-próba alkalmazásának előfeltétele (a

$$\chi^2 = \sum_{i=1}^{r} \frac{(f_i - np_i)^2}{np_i}$$

statisztika aszimptotikus χ^2-eloszlása miatt) az, hogy a legkisebb np_i (a legkisebb „elméleti" gyakoriság) is 10-nél nagyobb legyen. Az np_i elméleti gyakoriságokat a tapasztalati f_i gyakoriságok jól közelítik, ránézésre látszik, hogy *túl sok gyakoriság kisebb, mint 10.*

Az adatokat átsoroljuk $r=8$ osztályba; az osztályhatárok legyenek:

$$x_0 = -\infty,\; x_1 = 2{,}5,\; x_2 = 4{,}5,\; x_3 = 5{,}5,\; x_4 = 6{,}5,\; x_5 = 8{,}5,\; x_6 = 10{,}5,\; x_7 = 12{,}5,\; x_8 = \infty.$$

Így egyetlen intervallumban lesz csak 10-nél kevesebb adat (az utolsóban $f_i = 9$, $np_i = 7{,}4$, amint majd kiszámítjuk). Ettől a csekély kivételtől eltekinthetünk.

Így az osztályok száma $r=8$ lesz, a becsült paraméterek száma $s=2$, a χ^2-statisztika paramétere $r-1-s = 8-1-2 = 5$ lesz.

Az i-edik intervallumba esik a változó a nullhipotézis teljesülése esetén

$$p_i = P(x_{i-1} \leq X < x_i) = F(x_i) - F(x_{i-1}) = \Phi\left(\frac{x_i - m}{\sigma}\right) - \Phi\left(\frac{x_{i-1} - m}{\sigma}\right)$$

valószínűséggel. Ezek az „elméleti" valószínűségek, az np_i értékek az „elméleti" gyakoriságok, vesszük mindegyik np_i négyzetes eltérését az f_i „tapasztalati" gyakoriságtól, arányítjuk az elméleti np_i gyakorisághoz. *Az így kapott*

$$\frac{(f_i - np_i)^2}{np_i}$$

értékek összege jellemezni fogja az elméleti és tapasztalati gyakoriságok eltérését.

Ha a nullhipotézis igaz, akkor a

$$\chi^2 = \sum_{i=1}^{r} \frac{(f_i - np_i)^2}{np_i}$$

összegnek nem szabad túl nagynak lennie. Mekkora lehet maximálisan ahhoz, hogy elfogadjuk a nullhipotézist? Maximálisan a táblázatból kiolvasott χ^2 érték lehet ($r-1-s = 5$ szabadságfokhoz); ha annál nagyobb, a nullhipotézist (tehát azt, hogy a p_i valószínűségeket a normál eloszlás segítségével számíthatjuk, az előbbi módon) elvetjük.

Az első intervallumra esés valószínűsége:

$$p_1 = \Phi\left(\frac{2{,}5 - 6{,}8706}{3{,}29}\right) - \Phi(-\infty) = 1 - \Phi(1{,}328) =$$
$$= 1 - 0{,}9079 = 0{,}0921,$$
$$np_1 = 170 \cdot 0{,}0921 = 15{,}7;\; \frac{(f_1 - np_1)^2}{np_1} = \frac{(15 - 15{,}7)^2}{15{,}7} = 0{,}185,$$

és így tovább, kiszámítjuk ugyanezeket a második, harmadik, ..., nyolcadik intervallumra. (Igyekszünk felfelé kerekíteni, hogy a szóban forgó összeg minél nagyobb legyen.)

Osztályok, x_{i-1}; x_i	Gyakoriság, f_i	$\frac{x_i - \bar{x}}{s^*}$	p_i	np_i	$(f_i - np_i)^2$	$\frac{(f_i - np_i)^2}{np_i}$
∞ – 2,5	14	−1,328	0,0921	15,7	2,89	0,185
2,51– 4,5	27	−0,720	0,1437	24,4	6,76	0,277
4,51– 5,5	22	−0,416	0,1029	17,5	20,25	1,157
5,51– 6,5	20	−0,112	0,1167	19,8	0,04	0,002
6,51– 8,5	36	0,495	0,2343	39,8	14,44	0,363
8,51–10,5	24	1,103	0,1753	29,8	33,64	1,129
10,51–12,5	18	1,711	0,0915	15,6	5,76	0,369
12,51– ∞	9	∞	0,0435	7,4	2,56	0,346
Összesen:	170		1,000	170,0		$\chi^2 = 3{,}828$

Minthogy a statisztika számított értéke a χ^2-táblázatban az 5 paraméterértékhez és 95%-os szinthez tartozó 11,1 értékénél jelentősen kisebb, a nullhipotézist elfogadjuk, az előírástól való eltérések tekinthetők normál eloszlásúnak, a minta ennek nem mond ellent.

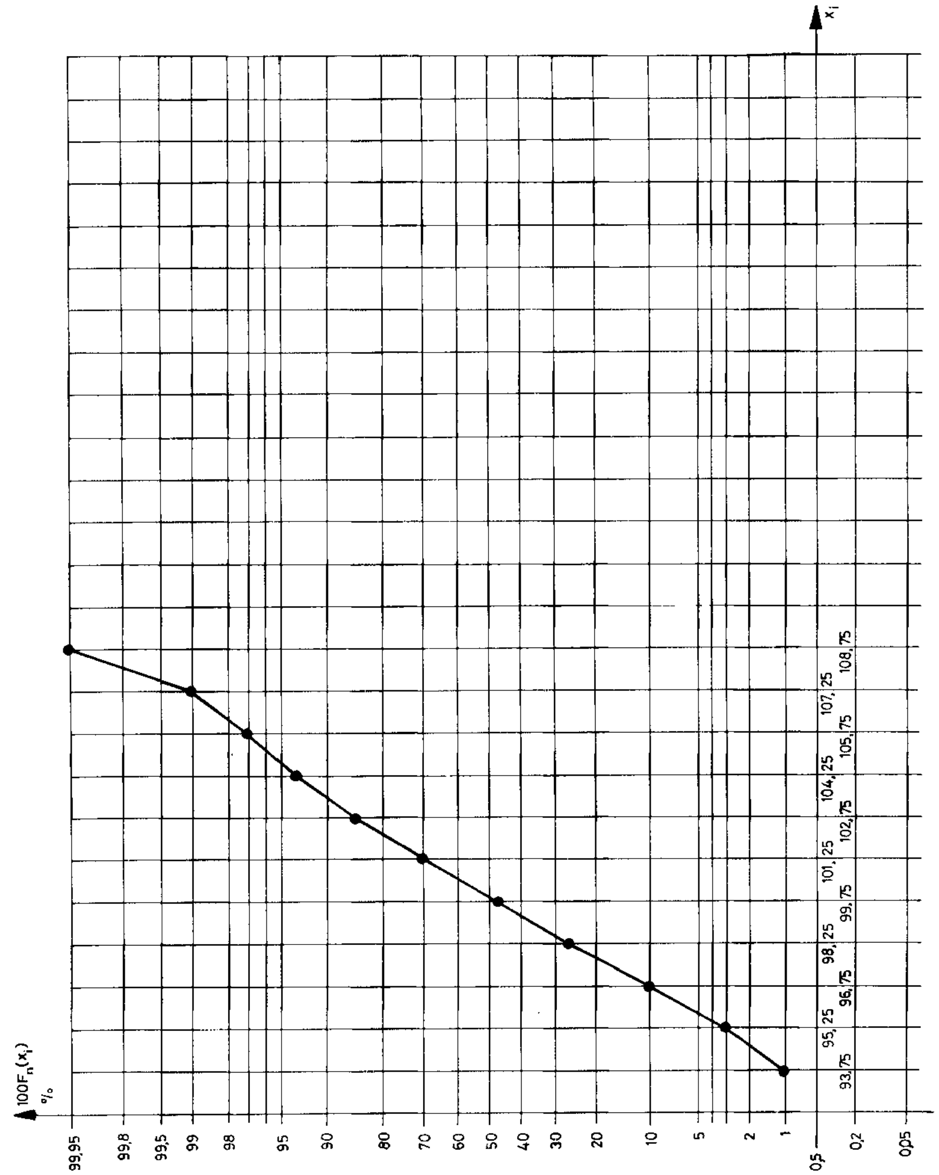
A 7. táblázat megfelelő részlete:

Szabadságfok	Valószínűség, % 95,0
5	11,1

2. Vizsgáljuk meg, hogy tekinthető-e normál eloszlásból származónak a szórásanalízisnél bemutatott papírtekercsek négyzetmétersúlya!

Az ottani mérési adatokat osztályokba soroljuk. Az $n = 6 \cdot 50 = 300$ elemű minta átlaga $\bar{x} = 100{,}9$, korrigált szórása $s^* = 2{,}75$.

A Gauss-papíron történő ábrázoláshoz a gyakoriságokból kiszámítjuk az intervallumra esés valószínűségét (az $\frac{f_i}{n}$ értékeket), ezek összegezésével az $F_n(x)$ eloszlásfüggvény értékét. Az osztályközepeket tekintve x_i értékeknél, a Gauss-papíron fölmért $F_n(x_i)$ értékek nagyjából egy egyenes körül sorakozó pontokat adnak (42. ábra), a normalitás feltételezhető. A számításokat a klimatizálás után mért osztálybasorolásos gyakorisági táblázat tartalmazza.



42. ábra

Osztályhatárok, század N/m^2	Osztályközép század N/m^2, x_i	Gyakoriság, f_i	Kumulált gyakoriság	Kumulált relatív gyakoriság, $F_n(x_i)$
93,0– 94,5	93,75	4	4	0,0133
94,5– 96,0	95,25	6	10	0,0333
96,0– 97,5	96,75	21	31	0,1033
97,5– 99,0	98,25	46	77	0,2566
99,0–100,5	99,75	60	137	0,4566
100,5–102,0	101,25	69	206	0,6866
102,0–103,5	102,75	47	253	0,8432
103,5–105,0	104,25	27	280	0,9332
105,0–106,5	105,75	10	290	0,9665
106,5–108,0	107,25	8	298	0,9933
108,0–109,5	108,75	2	300	1,0000

A χ^2-próbához észre kell vennünk, hogy az első, második, tizedik és tizenegyedik osztályban túl kevés az adat (4, 6, 8, 2 mindegyike kisebb, mint 10). Igaz, hogy ezek a tapasztalati gyakoriságok, de ha az elméleti np_i értékeket számítanánk ki ezekre az osztályokra vonatkoztatva, azok is kisebbek lennének 10-nél. Az első–második osztályokat vonjuk össze, az adatok száma a $(-\infty; 96{,}0)$ intervallumban $4+6 = 10$, ugyanígy a tizedik és tizenegyedik osztály egyesítésével a $(106{,}5; \infty)$ osztályban $8+2 = 10$ lesz.

Így $r=9$ osztályt nyertünk, a becsült paraméterek száma $s=2$, a χ^2-próba szabadságfokainak száma:

$$r-1-s = 9-1-2 = 6.$$

Számítások az új első intervallumon:

$$p_i = P(A_i) = F(x_i)-F(x_{i-1}) = \Phi\left(\frac{x_i-m}{\sigma}\right)-\Phi\left(\frac{x_{i-1}-m}{\sigma}\right),$$

$$p_1 = \Phi\left(\frac{96{,}5-100{,}9}{2{,}75}\right)-\Phi(-\infty) = \Phi(-1{,}7818)-0 = 0{,}0372,$$

$$np_1 = 300\cdot 0{,}0372 = 11{,}16,$$

$$(f_1-np_1)^2 = (10-11{,}16)^2 = 1{,}3456,$$

$$\frac{(f_1-np_1)^2}{np_1} = \frac{1{,}3456}{11{,}16} = 0{,}1206.$$

Számítások az új utolsó intervallumon:

$$p_9 = \Phi(\infty)-\Phi(2{,}0363) = 0{,}0209,$$

$$np_9 = 300\cdot 0{,}0209 = 6{,}27,$$

$$(f_9-np_9)^2 = (10-6{,}27)^2 = 13{,}9129,$$

$$\frac{(f_9-np_9)^2}{np_9} = \frac{13{,}9129}{6{,}27} = 2{,}2190.$$

Az utolsó oszlopban kapott értékek összege,

$$\sum_{i=1}^{r}\frac{(f_i-np_i)^2}{np_i} = 5{,}6831$$

megadja a számított χ^2 értéket. Ezt összehasonlítva a 95%-os biztonsági szinthez, 6 szabadságfokhoz tartozó táblázatbeli χ^2 értékkel (12,6-del), *a négyzetmétersúly normalitása elfogadható,* hiszen

$$5{,}6831<12{,}6.$$

A többi számítást az alábbi táblázat tartalmazza.

Intervallum sorszáma	Osztályhatárok, $x_{i-1}-x_i$, század N/m^2	Gyakoriság, f_i	$\frac{x_i-\bar{x}}{s^*}$	p_i	$300\,p_i$	$(f_i-300\,p_i)^2$	$\frac{(f_i-300\,p_i)^2}{300\,p_i}$
1.	$-\infty$ – 96,0	10	−1,7818	0,0372	11,16	1,3456	0,1206
2.	96,0– 97,5	21	−1,2363	0,0708	21,24	0,0576	0,0027
3.	97,5– 99,0	46	−0,6909	0,1366	40,98	25,2004	0,6149
4.	99,0–100,5	60	−0,1454	0,1975	59,25	0,5625	0,0094
5.	100,5–102,0	69	0,4000	0,2131	63,93	25,7049	0,4020
6.	102,0–103,5	47	0,9454	0,1723	51,69	21,9961	0,4255
7.	103,5–105,0	27	1,4909	0,1053	31,59	21,0681	0,6669
8.	105,0–106,5	10	2,0363	0,0472	14,16	17,3056	1,2221
9.	106,5– ∞	10		0,0209	6,27	13,9129	2,2190
							χ^2 = 5,6831

3. Próbáljunk megadni az illeszkedés-, homogenitás- és függetlenségvizsgálatnál szereplő χ^2 értékek számítására számítástechnikai szempontból egyszerűbb képleteket!

---

a) Az illeszkedésvizsgálatnál:

$$\chi^2 = \sum_{i=1}^{r} \frac{(k_i - np_i)^2}{np_i} = \frac{1}{n}\Sigma\frac{k_i^2}{p_i} - 2\Sigma k_i + n\Sigma p_i =$$

$$= \frac{1}{n}\Sigma\frac{k_i^2}{p_i} - 2n + n = \frac{1}{n}\Sigma\frac{k_i^2}{p_i} - n.$$

b) A homogenitásvizsgálatnál

$$\chi^2 = mn \sum_{i=1}^{r} \frac{\left(\frac{f_i}{n} - \frac{k_i}{m}\right)^2}{f_i + k_i} = \frac{m}{n}\Sigma\frac{f_i^2}{f_i + k_i} - 2\Sigma\frac{f_i k_i}{f_i + k_i} + \frac{n}{m}\Sigma\frac{k_i^2}{f_i + k_i}.$$

c) Tiszta függetlenségvizsgálatnál

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(k_{ij} - np_i q_j)^2}{np_i q_j} = \sum_i\sum_j \frac{k_{ij}^2 - 2nk_{ij}p_i q_j + n^2 p_i^2 q_j^2}{np_i q_j} =$$

$$= \sum_i\sum_j \frac{k_{ij}^2}{np_i q_j} - 2\sum_i\sum_j k_{ij} + n\sum_i\sum_j p_i q_j = \frac{1}{n}\sum_i\sum_j \frac{k_{ij}^2}{p_i q_j} - 2n + n =$$

$$= \frac{1}{n}\sum_i\sum_j \frac{k_{ij}^2}{p_i q_j} - n.$$

Becsléses függetlenségvizsgálatnál

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(k_{ij} - \frac{f_{i.}f_{.j}}{n}\right)^2}{f_{i.}f_{.j}} =$$

$$= n\sum_i\sum_j \frac{k_{ij}^2}{f_{i.}f_{.j}} - 2\sum_i\sum_j k_{ij} + \frac{1}{n}\sum_i\sum_j f_{i.}f_{.j} =$$

$$= n\sum_i\sum_j \frac{k_{ij}^2}{f_{i.}f_{.j}} - 2n + \frac{1}{n}\underbrace{\sum_i f_{i.}}_{n}\underbrace{\sum_j f_{.j}}_{n} = n\sum_i\sum_j \frac{k_{ij}^2}{f_{i.}f_{.j}} - n.$$

Végiggondolható, hogy az *a)* és *c)* esetekben valóban kisebb tárigényű, rövidebb számítógépi futású algoritmusokat kaptunk, a *b)* esetben azonban nem. A *b)* esetben a gyakorlati számítást egyszerűsíti, ha előbb kiszámítjuk az

$$\omega = \frac{m}{m+n}, \quad \omega_i = \frac{f_i}{f_i + k_i} \quad (i = 1, 2, \ldots, r)$$

mennyiségeket, amelyek segítségével:

$$\chi^2 = \frac{1}{\omega(1-\omega)}\left(\sum_{i=1}^{r} f_i\omega_i - m\omega\right).$$

4. Tekinthető-e szabályosnak az a játékkocka, amelyet $n = 1200$-szor feldobva, az egyes számok gyakoriságára az alábbi táblázatban látható eredményeket kaptuk?

Dobott szám	Gyakoriság, k_i	Relatív gyakoriság, k_i/n	$\frac{k_i}{n} - p_i$
1-es	195	0,1625	−0,0042
2-es	210	0,1750	0,0083
3-as	190	0,1583	−0,0083
4-es	204	0,1700	0,0033
5-ös	205	0,1708	0,0042
6-os	196	0,1633	−0,0033
Összesen:	$1200 = n$		

Ha a kocka szabályos, akkor bármely szám dobása egyenlő valószínűségű. Nullhipotézisünk:

$$P(A_i) = P(X = x_i) = p_i = \frac{1}{6} \quad (i = 1, 2, \ldots, 6).$$

Kíváncsiságból megmértük a táblázatban a relatív gyakoriságoknak a feltételezett valószínűségtől való eltérését, a $\left(\frac{k_i}{n}-p_i\right)$ értékeket. Annyira kicsik, hogy azt sejtjük: a kocka szabályos. Nem tudjuk azonban, hogy milyen eltérések azok, amelyek még nem szignifikánsak, melyek azok, amelyek már statisztikailag jelentősek a nullhipotézis elvetéséhez.

Oldjuk meg a feladatot a χ^2-próbával! Tiszta illeszkedésvizsgálatról van szó, mert X eloszlását ismerjük (egyenletes eloszlás, amelyben nincs becsülendő paraméter). Használjuk az előző feladatban kapott képletet!

$$\chi^2=\sum_{i=1}^{r}\frac{(k_i-np_i)^2}{np_i}=\frac{1}{n}\sum_{i=1}^{r}\frac{k_i^2}{p_i}-n=$$
$$=\frac{1}{1200}\cdot 6\,\Sigma k_i^2-1200=$$
$$=\frac{1}{200}(195^2+210^2+190^2+204^2+205^2+196^2)-1200=$$
$$=\frac{240\,282}{200}-1200=1{,}41.$$

A szabadságfokok száma: $r-1=5$, a kapott

$$1{,}41<\chi^2_{95}=11{,}1,$$

a nullhipotézis elfogadható, a kocka szabályosnak tekinthető.

Az $np_i=1200\cdot\frac{1}{6}=200$ érték minden i-re jóval nagyobb, mint 10, a χ^2-próba nyugodtan alkalmazható.

5. Egy textilüzemben korábbi tapasztalatok azt mutatták, hogy a fonalszakadások száma egy bizonyos géptípus és fonal esetén Poisson-eloszlású, $\lambda=8$ paraméterrel. Vizsgáljuk meg az alábbi újabb adatokkal, fennáll-e most is a Poisson-eloszlás! Az $n=75$ műszakban mért adatok:

A fonalszakadások száma	Gyakoriság, k_i	$\lambda=8$ paraméterhez tartozó Poisson-valószínűségek, $P(x=k)=\frac{\lambda^k}{k!}e^{-\lambda}$
0	–	0,000 33
1	–	2 68
2	–	10 73
3	–	28 63
4	1	57 25
5	4	91 60
6	9	122 14
7	7	139 59
8	8	139 59
9	10	124 08
10	8	99 26
11	10	72 19
12	5	48 13
13	7	29 62
14	2	16 92
15	3	9 03
16	1	4 51
17	–	2 12
18	–	1 60
Összesen:	$75=n$	1,000 00

Gyűjtsük most az adatokat olyan intervallumokba, hogy mindegyik osztályba esés valószínűsége szorozva az adatok számával, n-nel legalább 10-et adjon! Megfelelő például az alábbi osztálybasorolás:

Esemény	Gyakoriság, f_i	Valószínűség, p_i	np_i	$\frac{(f_i-np_i)^2}{np_i}$
$A_1 = \{X \leqq 6\}$	14	0,313 36	23,6	3,905
$A_2 = \{7 \leqq X \leqq 8\}$	15	279 18	20,9	1,666
$A_3 = \{9 \leqq X \leqq 10\}$	18	223 34	16,7	0,101
$A_4 = \{11 \leqq X\}$	28	184 12	13,8	14,612
Összesen:	$75=n$	1,000 00	75,0	20,284

A szabadságfokok száma $r-1 = 3$, mivel tiszta illeszkedésvizsgálatról van szó (λ értékét ismerjük). Mivel

$$\chi^2_{\text{számított}} = 20{,}284 > \chi^2_{95} = 7{,}81,$$
$$> \chi^2_{99,95} = 17{,}7,$$

a nullhipotézist elutasítjuk, a fonalszakadások eloszlását megváltozottnak tekinthetjük.

Az a kérdés, hogy az eloszlás típusa változott-e meg vagy az eloszlás típusa továbbra is Poisson-eloszlás, csak 8-tól különböző paraméterrel.

λ értékét a mintából becsüljük:

$$\hat{\lambda} = \sum_{i=0}^{18} \frac{k_i x_i}{n} = 9{,}48.$$

Próbálkozzunk azzal a feltevéssel, hogy a fonalszakadások száma továbbra is Poisson-eloszlású, csak $\hat{\lambda}=9{,}48$ paraméterrel, vagyis a nullhipotézis:

$$H_0\colon P(X=k) = \frac{\hat{\lambda}^k}{k!} e^{-\hat{\lambda}}.$$

A 3. táblázatban a Poisson-eloszlás valószínűségeit $\lambda=9{,}4$ és $\lambda=9{,}5$ esetén találjuk meg. Lineáris interpolációval ezeket a valószínűségeket $\hat{\lambda}=9{,}48$-ra is megkaphatjuk. Mivel $X \geqq 13$-hoz tartozó np_i is nagyobb most, mint 10, ezért eggyel több osztályba soroltuk az adatokat:

Esemény	Gyakoriság, f_i	Valószínűség, p_i	np_i	$\frac{(f_i-np_i)^2}{np_i}$
A_1	14	0,166 51	12,5	0,18
A_2	15	227 80	17,1	0,26
A_3	18	253 49	19,0	0,05
A_4	15	190 30	14,3	0,04
A_5	13	161 90	12,1	0,08
Összesen:	$75=n$	1,000 00	75,0	0,61

Mivel $s=1$ becsült paraméterünk van, a χ^2-eloszlás paramétere:

$$r-1-s = 5-1-1 = 3,$$

a számított χ^2 érték (0,61) jóval kisebb, mint a táblázatbeli $\chi^2_{95}=7{,}81$, elfogadjuk, hogy a fonalszakadások száma Poisson-eloszlású, csak megnövekedett a λ paraméter. Mivel a paraméter a várható értékkel egyezik meg ennél az eloszlásnál, ezért a fonalszakadások száma várhatóan megnőtt a korábbihoz képest, vagy a gyártási körülmények rosszabbodtak a szóban forgó gépen, vagy a fonal minősége romlott.

6. Munkás- és alkalmazotti családokból kiválasztottunk 220 háztartást, ezekből 120 esetben az asszony nem maradt otthon a szülési szabadság letelte után (A csoport), 100 esetben otthon maradt (B csoport). Azt akarjuk vizsgálni, van-e szignifikáns eltérés a gyerekek számában az otthon maradók és otthon nem maradók között, vagyis: homogén-e a két eloszlás. Nullhipotézisként tegyük fel, hogy a két eloszlás azonos. Ellenhipotézis az, hogy a két eloszlás nem azonos.

A gyerekek száma	Gyakoriság és relatív gyakoriság a mintában				Együtt, f_i+k_i	$\dfrac{\left(\frac{f_i}{n}-\frac{k_i}{m}\right)^2}{f_i+k_i}$
	A csoport		*B* csoport			
	f_i	f_i/n	k_i	k_i/m		
0	36	0,300	28	0,280	64	$6{,}25\cdot 10^{-6}$
1	41	0,341	36	0,360	77	$4{,}37\cdot 10^{-6}$
2	28	0,233	22	0,220	50	$3{,}56\cdot 10^{-6}$
3	11	0,092	8	0,080	19	$7{,}16\cdot 10^{-6}$
4	3	0,025	4	0,040	7	$40{,}91\cdot 10^{-6}$
5 vagy több	1	0,008	2	0,020	3	$45{,}37\cdot 10^{-6}$
Összesen:	$120 = n$		$100 = m$		220	$107{,}62\cdot 10^{-6}$

$$\chi^2 = mn\sum_{i=1}^{r}\frac{\left(\frac{f_i}{n}-\frac{k_i}{m}\right)^2}{f_i+k_i} = 100\cdot 120\cdot 107{,}62\cdot 10^{-6} =$$

$$= 1{,}291\,440.$$

Ez kisebb, mint az $r-1 = 5$ szabadságfokú χ^2-eloszlás valamennyi értéke, amelyet a 7. táblázatban feltüntettünk. Adataink tehát nincsenek ellentmondásban azzal, hogy a két eloszlás azonos.

7. Gyermekgyógyász szívspecialisták foglalkoztak azzal a kérdéssel, hogy a veleszületett rendellenességek (vitiumok) összefüggésben vannak-e az anyát a terhesség első 3 hónapjában sújtó megbetegedésekkel (kivéve a vírusos megbetegedést, amelyeket külön vizsgáltak). Az összes vizsgált, rendellenességgel született gyereknél és a véletlenszerűen választott 100 fős kontrollcsoportban a következő volt az eredmény:

A gyermek	Az anya		Összesen:
	beteg volt	nem volt	
Rendellenes	26	184	210
Nem rendellenes	5	95	100
Összesen:	31	279	310

A kapott 2×2-es kontingenciatáblázat alapján döntsük el, függetlennek tekinthető-e az anyáknál történő megbetegedés a gyermekek veleszületett rendellenességétől.

A 2×2-es kontingenciatáblázat esetén

$$\chi^2 = n\frac{(k_{11}k_{22}-k_{12}k_{21})^2}{f_{1.}f_{2.}f_{.1}f_{.2}} = \frac{310(26\cdot 95-184\cdot 5)^2}{31\cdot 279\cdot 210\cdot 100} = 4{,}10.$$

Mivel az egész (nem ismert) sokaságra vonatkozó adatokat (bekövetkezési valószínűségeket) a mintából becsültük, a χ^2-eloszlás paramétere: $(r-1)(s-1) = 1\cdot 1 = 1$. A 7. táblázat első sorában, 95% biztonsági szintnek megfelelő oszlopában 3,84-et találunk, a számított érték nagyobb, mint a táblázatbeli. A függetlenség hipotézisét 95%-os biztonsági szinten elvetjük, az anya betegsége és a csecsemő öröklött rendellenessége között összefüggés van.

(Még szignifikánsabb volt az eltérés, még erősebb az összefüggés a vírusos megbetegedések és a rendellenességek között.)

8. Csavarok szakítószilárdsága és méretre való megfelelősége közti összefüggés vizsgálatára rendelkezésre áll az alábbi vizsgálati eredmény (egy 2×2-es kontingenciatáblázat):

Függetlennek tekinthető-e a két tulajdonság?

Méretre	Szakítószilárdságra		Összesen:
	megfelelő	selejtes	
Megfelelő	416	23	439
Selejtes	16	5	21
Összesen:	432	28	460

A próbastatisztika számított értéke:

$$\chi^2 = \frac{460(416 \cdot 5 - 16 \cdot 23)^2}{432 \cdot 28 \cdot 439 \cdot 21} = 1{,}209 < \chi^2_{95} = 3{,}84,$$

a két tulajdonság függetlennek tekinthető (a χ^2-eloszlás paramétere $(r-1)(s-1) = 1$).

9. Csapágygyűrűknél fontos minőségi jellemző a külső és belső átmérő (Y és X). Az átmérő nagysága alapján az elkészült gyűrűket három kategóriába soroljuk: jó, javítható, selejtes. Találomra kiválasztunk $n = 200$ db-ot annak ellenőrzésére, hogy a külső és a belső átmérő független-e egymástól. Nullhipotézisünk az, hogy független, ellenhipotézisünk az, hogy nem. Döntsünk χ^2-próbával az alábbi, 3×3-as kontingenciatáblázat alapján:

Belső átmérő	Külső átmérő			Összesen	
	jó	javítható	selejtes		
Jó	169	8	1	178	$f_{1.}$
Javítható	9	4	1	14	$f_{2.}$
Selejtes	1	3	4	8	$f_{3.}$
Összesen:	179 $f_{.1}$	15 $f_{.2}$	6 $f_{.3}$	$200 = n$	

Két képletet is megadtunk a 3. feladatban:

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(k_{ij} - \frac{f_{i.} f_{.j}}{n}\right)^2}{f_{i.} f_{.j}} = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{k_{ij}^2}{f_{i.} f_{.j}} - n.$$

A másodikkal könnyebb számolni, ezzel dolgozzunk.

$$\frac{k_{11}^2}{f_{1.} f_{.1}} = \frac{169^2}{178 \cdot 179},$$

$$\frac{k_{12}^2}{f_{1.} f_{.2}} = \frac{8^2}{178 \cdot 15} \quad \text{stb.}$$

Így jön létre a $\frac{k_{ij}^2}{f_{i.} f_{.j}}$ értékek táblázata:

i	j		
	1	2	3
1	$\frac{169^2}{178 \cdot 179}$	$\frac{8^2}{178 \cdot 15}$	$\frac{1^2}{178 \cdot 6}$
2	$\frac{9^2}{14 \cdot 179}$	$\frac{4^2}{14 \cdot 15}$	$\frac{1^2}{14 \cdot 6}$
3	$\frac{1^2}{8 \cdot 179}$	$\frac{3^2}{8 \cdot 15}$	$\frac{4^2}{8 \cdot 6}$

$$\chi^2_{\text{számított}} = 200 \cdot 1{,}450\,752\,6 - 200 = 90{,}150\,526.$$

Esetünkben $r = s = 3$, a szabadságfokok száma: $(r-1)(s-1) = 2 \cdot 2 = 4$. A kapott χ^2 érték jóval nagyobb a 7. táblázat 4. sorának bármely χ^2 értékénél, ezért a külső és belső átmérő függetlenségére vonatkozó hipotézist elvetjük.

Megjegyzés

Felvethető kérdés ilyenkor: milyen erős a függés? Két módszerrel lehet tisztázni:

a) Kiszámítjuk a két változó korrelációs együtthatójának mintabeli becslését, r-t, ha $|r|$ 1-hez esik közel, a függés erős, ha 0-hoz, akkor a függés gyenge. Ekkor természetesen szükségünk van a mért adatokra.

b) Vezessük be a

$$\varphi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(p_{ij} - p_i q_j)^2}{p_i q_j} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{p_{ij}^2}{p_i q_j} - 1$$

ún. *négyzetes kontingenciát*. Ennek empirikus becslése a számított χ^2 érték n-edrésze,

$$\varphi^2 \approx \frac{\chi^2}{n} = \frac{90{,}150\ 526}{200},$$

az X-re és Y-ra vett osztályok száma közül a kisebb legyen q,

$$q = \min(r, s) = 3,$$

a függőség mértékét a

$$\frac{\varphi^2}{q-1}$$

hányados adja meg. Bizonyítható, hogy

$$0 \leqq \frac{\varphi^2}{q-1} \leqq 1,$$

ez a hányados 0 akkor és csak akkor, ha az $A_1, A_2, \ldots, A_r$ és $B_1, B_2, \ldots, B_s$ eseményrendszerek függetlenek, a hányados 1 akkor és csak akkor, ha függnek egymástól.

Minél közelebb esik a hányados 0-hoz, annál gyengébb, minél közelebb esik 1-hez, annál erősebb a függés a két eseményrendszer között.

Jelen esetben a $\frac{\varphi^2}{q-1}$ hányados mintabeli becslése:

$$\frac{\varphi^2}{q-1} \approx \frac{90{,}150\ 526}{200 \cdot 2} \approx 0{,}225,$$

a két méret egymással gyenge kapcsolatban áll. A függetlenséget elvetettük, mégis (a gyenge függőség alapján), a két méretet célszerű külön-külön ellenőrizni.

10. Független-e egymástól a szem és a haj színe? $n=467$ ember vizsgálatának eredménye az utóbbi 2×2-es kontingenciatáblázatban látható:

Szem	Haj		Összesen
	világos	sötét	
Világos	307	32	339
Sötét	33	95	128
Összesen:	340	127	467

$$\chi^2 = 467 \cdot \frac{(307 \cdot 95 - 32 \cdot 33)^2}{339 \cdot 128 \cdot 340 \cdot 127} = 196{,}931\ 58.$$

Az egész sokaságra vonatkozó, elméleti valószínűségeket becsültük, a χ^2-eloszlás paramétere: $(r-1)(s-1) = 1$. A 7. táblázat első sorában minden χ^2 jóval kisebb, mint a most számított érték, a nullhipotézist (a szem és a haj színének függetlenségét) elvetjük.

Milyen erős a függőség? $r=2$ és $s=2$ közül a „kisebbik" 2, tehát $q=2$, a függés mértékét a

$$\frac{\varphi^2}{q-1} = \frac{\chi^2}{n(q-1)} = \frac{0{,}421\ 695\ 03}{1} \approx 0{,}42$$

hányados adja meg. Minél közelebb esik a hányados 1-hez, annál erősebb a függés, minél közelebb a 0-hoz, annál gyengébb. A 0,42-es hányados közepes függést jelez a haj és a szem színe között.

11. Vizsgáljuk meg Störmer- és egymintás Kolmogorov-próbák együttes alkalmazásával, feltehető-e az alábbi minta alapján, hogy a vizsgált elektrolitkondenzátorok élettartama exponenciális eloszlású (amint azt korábban tapasztalták)!

Minta sorszáma, i	Élettartam (rendezve), X_i^*	Transzformált élettartam, $Y_i^* = F(Y_i^*)$	$F_n(Y_i^*)$	$\lvert F_n(Y_i^*) - F(Y_i^*)\rvert$
1	450	0,2397	0	0,2397
2	650	0,3409	1/19	0,2883
3	950	0,4847	2/19	0,3794 = max.
4	1 000	0,5073	3/19	0,3494
5	1 150	0,5712	4/19	0,3607
6	1 300	0,6312	5/19	0,3680
7	1 350	0,6498	6/19	0,3340
8	1 450	0,6844	7/19	0,3160
9	1 550	0,7164	8/19	0,2953
10	1 700	0,7603	9/19	0,2866
11	1 750	0,7736	10/19	0,2473
12	2 100	0,8575	11/19	0,2786
13	2 200	0,8788	12/19	0,2472
14	2 250	0,8882	13/19	0,2040
15	2 400	0,9121	14/19	0,1753
16	2 550	0,9321	15/19	0,1426
17	2 700	0,9481	16/19	0,1060
18	2 950	0,9680	17/19	0,0733
19	3 400	0,9920	18/19	0,0446
20	3 700	–	–	–
Összesen:	37 550			

A 20 db rendezett X_i^* értékből először 19 db, ugyancsak rendezett Y_k^* értéket állítunk elő a Störmer-transzformációval ($X_0^* = 0$ felvételével):

$$Y_k^* = \frac{\sum_{i=0}^{k-1} (n-i)(X_{i+1}^* - X_i^*)}{\sum_{j=1}^{n} X_j^*},$$

$k=1$ esetén:

$$Y_1^* = \frac{n(X_1^* - X_0^*)}{\sum_{j=1}^{n} X_j^*} = \frac{20(450-0)}{37\,550} = 0{,}2397,$$

$k=2$ esetén:

$$Y_2^* = \frac{n(X_1^* - X_0^*) + (n-1)(X_2^* - X_1^*)}{\Sigma X_j^*} = \frac{20 \cdot 450 - 19(650-450)}{37\,550} =$$
$$= 0{,}3409,$$

és így tovább, $k=19$-ig (l. a táblázatban).

A kapott $0{,}2397 - 0{,}3409 - 0{,}4847 - \ldots - 0{,}9920$ értékekről kell kimutatni, hogy tekinthetők egy (0, 1) intervallumon egyenletes eloszlású Y változóra vonatkozó $n-1 = 19$ elemű mintának. A nullhipotézis tehát:

$$H_0\colon P(Y<y) = F(y) = \begin{cases} 0, & \text{ha} \quad y<0, \\ y, & \text{ha} \quad 0 \leqq y < 1, \\ 1, & \text{ha} \quad y \geqq 1, \end{cases}$$

ezt vizsgáljuk kétoldali ellenhipotézissel szemben 95%-os szinten.

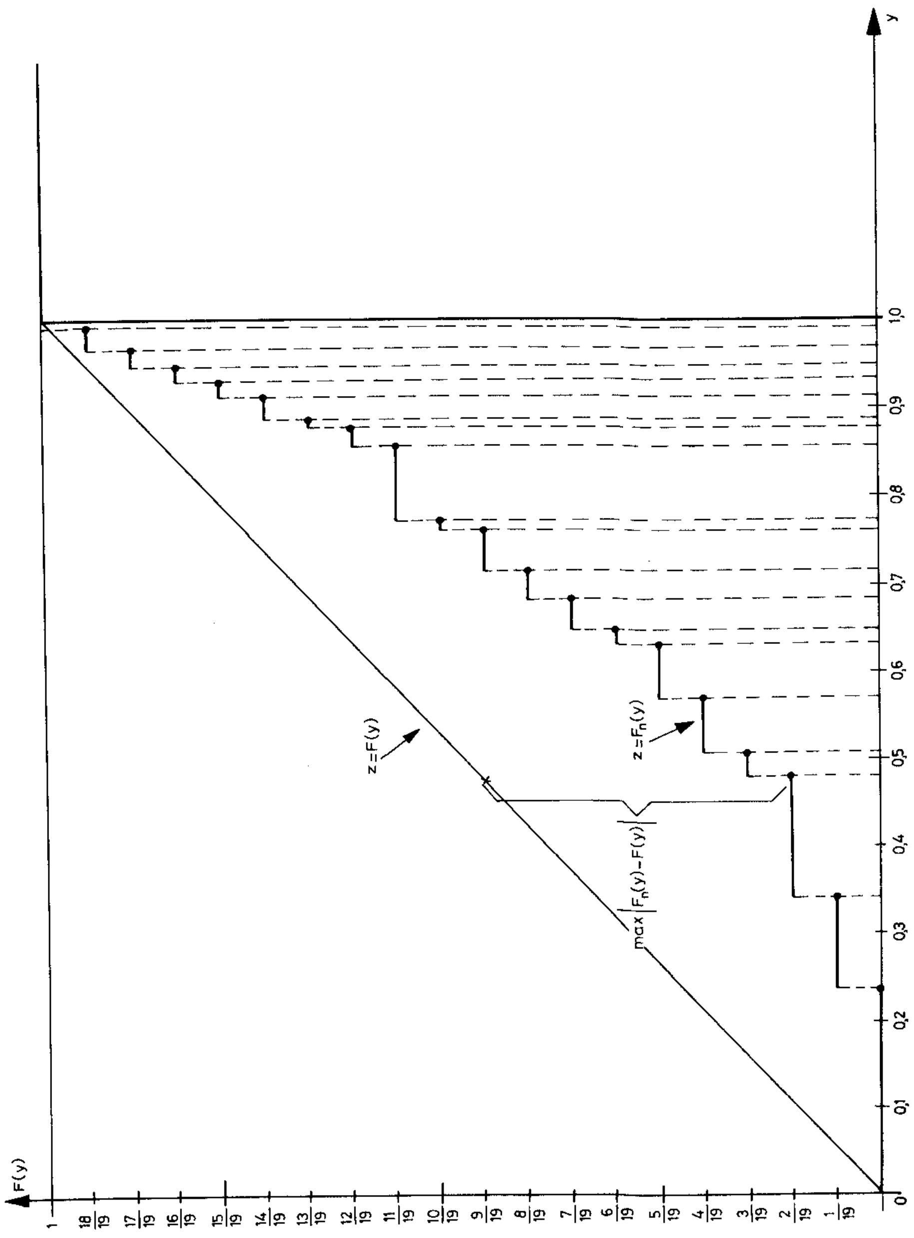
A (0, 1) intervallumon tehát $F(y)$ magával y-nal egyenlő (ezt tüntettük föl a táblázat harmadik oszlopában). Elméleti eloszlásfüggvény-értékek gyanánt tehát magukat a transzformált Y_i^* értékeket kell felvennünk, $F(y) = y$ a (0, 1) intervallumon.

A 43. ábrán elkészítettük $F(y)$ és $F_n(y)$ grafikonját. $F_n(y)$ a (0, 1) intervallumon fölvett minden újabb értéknél 1/19-et ugrik, balról folytonos. A Kolmogorov-féle egymintás próba szerint a nullhipotézis elfogadásáról a tapasztalati és elméleti eloszlásfüggvény eltérése abszolút értékének a maximuma:

$$D_n = \max_{-\infty<y<\infty} \lvert F_n(y) - F(y)\rvert$$

dönt (amint az szemléletesen is sejthető). Ha $D_n = \max\limits_{(y)} \lvert F_n(y) - F(y)\rvert$ meghaladja a 11. táblázatban található kritikus értéket, akkor elutasítjuk azt a hipotézist, hogy a valószínűségi változó eloszlásfüggvénye valóban $F(y)$.

Ha $y<0$ vagy $y \geqq 1$, az elméleti és tapasztalati eloszlásfüggvény különbsége 0. A (0, 1) intervallumon $F(y)$ szigorúan növekvő, $F_n(y)$ viszont „vízszintes" lépcsőkből álló függvény, emiatt elég csak a mintából kapott pontokban kiszámítani $F_n(y)$ és $F(y)$ eltérésének abszolút értékét. Akár számítással, akár rajzról azt kapjuk, hogy



43. ábra

$$D_{19} = \max_{y\text{-ra}} |F_{19}(y) - F(y)| = Y_3^* - F_{19}(Y_3^*) = 0{,}3794.$$

A 11. táblázatban kétoldali ellenhipotézis esetén a kritikus érték (95% biztonság esetén) 0,3014. A számított érték nagyobb, mint a táblázatbeli kritikus érték, Y egyenletes eloszlását és egyúttal X exponenciális eloszlását visszautasítjuk, a kondenzátor-élettartam megváltozott, nem tekintjük továbbra is exponenciálisnak.

12. El akarjuk dönteni, hogy a Tisza Szegednél mért évi maximális vízállásai ugyanazt az eloszlást követték-e 1876–1925 között, mint 1926-tól 1975-ig. Legyen X a maximális vízállás az első 50 évben, Y pedig a második 50 évben. A nullhipotézis az, hogy az eloszlásfüggvények azonosak:

$$H_0\colon P(X<x) = P(Y<x), \quad \text{vagyis} \quad F(x) \equiv G(x),$$

ellenhipotézis az, hogy nem azonosak,

$$H_1\colon F(x) \not\equiv G(x).$$

A méterben megadott adatok osztályba sorolva vannak megadva:

Maximális vízállás, V	Gyakoriság az első 50 évben, f_i	Gyakoriság a második 50 évben, k_i
$V<5$ m	5	10
$5 \leqq V < 6$	11	11
$6 \leqq V < 7$	13	13
$7 \leqq V < 8$	13	10
$8 < V$	8	6
Összesen:	$50 = n$	$50 = m$

a) χ^2-próbás homogenitásvizsgálatot végzünk. Mivel $m=n$, ezért

$$\chi^2 = mn \sum_{i=1}^{r} \frac{\left(\frac{f_i}{n}-\frac{k_i}{m}\right)^2}{f_i+k_i} = \sum_{i=1}^{r} \frac{(f_i-k_i)^2}{f_i+k_i}.$$

Dolgozzunk táblázattal:

f_i	k_i	f_i+k_i	f_i-k_i	$(f_i-k_i)^2$	$\frac{(f_i-k_i)^2}{f_i+k_i}$
5	10	15	−5	25	25/15 = 1,666 666 7
11	11	22	0	0	0
13	13	26	0	0	0
13	10	23	3	9	9/23 = 3,913 043 5 · 10^{-1}
8	6	14	2	4	4/14 = 2,857 142 9 · 10^{-1}
Összesen: 50	50				2,343 685 3

A χ^2 próbastatisztika nagy n és m esetén közelítőleg χ^2-eloszlású, $r-1=4$ paraméterrel. χ^2 értéke nagy valószínűséggel kisebb, mint a 7. táblázatban megadott kritikus érték. Mivel

$$\chi^2_{\text{számított}}=2{,}343\,685\,3 < \chi^2_{95}=9{,}49,$$

95%-os biztonsági szinten nincs okunk elvetni H_0-t, a Tisza Szegednél mért vízállása ugyanolyan eloszlásúnak tekinthető az első 50 évben, mint a második 50 év alatt. (A Tisza kanyarjait átvágó, nagymértékű szabályozások már 1876 előtt megtörténtek. Az azóta történt szabályozás nem változtatta lényegesen a maximális árhullámokat.)

b) A homogenitásvizsgálatot most a Kolmogorov–Szmirnov-féle kétmintás próbával végezzük el.

A 12. táblázat csak $n=m=30$-ig közli a kritikus értékeket, tehát most nem használható. Mivel n és m elég nagyok, a Kolmogorov-féle $K(z)$ függvény 14. táblázatát használjuk:

z	$K(z)$
1,35	0,9477
1,36	0,9505

Ha a nullhipotézis teljesül, kb. 0,95 valószínűséggel állíthatjuk, hogy

$$\sqrt{\frac{nm}{n+m}} \max_{x\text{-re}} |F_n(x)-G_n(x)| < z = 1{,}36.$$

Az eloszlásfüggvények értékeit a gyakoriságok összegezésével megkaphatjuk:

f_i	k_i	$F_n(x)=\frac{\Sigma f_i}{n}$	$G_m(x)=\frac{\Sigma k_i}{m}$	$\lvert F_n(x)-G_m(x)\rvert$
5	10	5/50	10/50	5/50
11	11	16/50	21/50	5/50
13	13	29/50	34/50	5/50
13	10	42/50	44/50	2/50
8	6	50/50	50/50	0

$$\sqrt{\frac{nm}{n+m}} \max_{x\text{-re}} |F_n(x)-G_m(x)| = \sqrt{\frac{50\cdot 50}{50+50}}\cdot\frac{5}{50} = \frac{25}{50} = 0{,}5.$$

A számított érték (0,50) kisebb, mint a táblázatbeli kritikus érték (1,36), tehát nincs alapunk H_0 elutasítására, az első 50 év alatt a maximális vízállás ugyanolyan eloszlásúnak vehető, mint a második 50 év alatt.

13. A Tiszán Tokajnál az 1945–1975 közötti időszakban a maximális vízállás-túllépéseket (az 500 cm fölötti maximális vízállás-többleteket) mérték 6–6 alkalommal:

- az I. negyedévben (X változó),
- a II. negyedévben (Y változó).

A mért vízállás-túllépések:

$$X_1 = 65, \quad X_2 = 13, \quad X_3 = 164, \quad X_4 = 88, \quad X_5 = 71, \quad X_6 = 45,$$
$$Y_1 = 194, \quad Y_2 = 257, \quad Y_3 = 123, \quad Y_4 = 132, \quad Y_5 = 55, \quad Y_6 = 258.$$

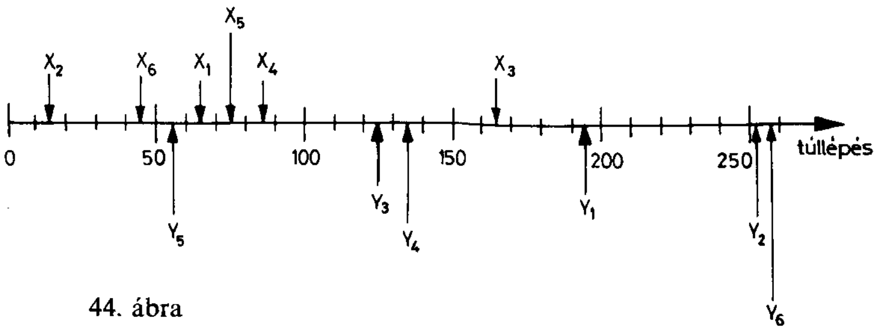
Vizsgáljuk Wilcoxon-próbával, hogy a

$$H_0\colon \; P(X < Y) = \frac{1}{2}$$

nullhipotézis vagy az egyoldali

$$H_1\colon \; P(X < Y) > \frac{1}{2}$$

ellenhipotézis érvényes-e!



44. ábra

A 44. ábrán ábrázoltuk a vízállás-túllépéseket (fölül az X_i értékeket, alul az Y_i értékeket). X_1 a nagyság szerinti rangsorban a negyedik ($r_1 = 4$), X_2 az első ($r_2 = 1$), X_3 a kilencedik ($r_3 = 9$), X_4 a hatodik ($r_4 = 6$), X_5 az ötödik ($r_5 = 5$), X_6 a második ($r_6 = 2$).

A rangszámok összege:

$$\sum_{i=1}^{6} r_i = 4+1+9+6+5+2 = 27,$$

a Wilcoxon-statisztika értéke:

$$W_{n,m} = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2} = 27 - \frac{6 \cdot 7}{2} = 6,$$

mivel

$$W_{\text{számított}} = 6 < \frac{mn}{2} = \frac{6 \cdot 6}{2} = 18,$$

ezért a most számított W-t hasonlítjuk össze a 13. táblázatban található $W_{\frac{\varepsilon}{2}}$-lel. (Vigyázat: az elfogadási tartomány $W_{\frac{\varepsilon}{2}} \leq W$!)

	m
$n=6$	6
	5

Mivel $W_{\text{számított}} = 6 > W_{\text{táblázatbeli}} = 5$, vagyis a próbastatisztika értéke az elfogadási tartományra esik, a H_0 hipotézis elvetésére 0,975 szinten nincs alapunk. Elfogadjuk tehát, hogy átlagosan ugyanannyi második negyedévinél kisebb első negyedévi árvíztúllépés van, mint amennyi nagyobb, $P(X < Y) = P(X > Y) = \frac{1}{2}$ igaz.

14. Vizsgáljuk most azt a kérdést, hogy a Tisza árvíztúllépései Tokajnál ugyanolyan eloszlásúak-e az első negyedévben, mint a második negyedévben! X a maximális túllépés az első negyedévben, $F(x)$ eloszlásfüggvénnyel, Y a maximális túllépés a második negyedévben, $G(x)$ eloszlásfüggvénnyel.

$$H_0\colon \; F(x) \equiv G(x), \qquad H_1\colon \; F(x) \not\equiv G(x)$$

(kétoldali ellenhipotézis).

Az $m = 36$, $n = 29$ adatot nagysági rangsorba szedték, megállapították az X_i elemek rangjainak (nagyság szerinti sorszámainak) az összegét, majd ebből a Wilcoxon-statisztika értékét. Az eredmény:

$$W_{29,36} = 510.$$

Mivel n és m elég nagyok, közelíthetjük a Wilcoxon-statisztikát normálissal. A várható érték:

$$M(W)=\frac{mn}{2}=\frac{36\cdot 29}{2}=522,$$

a szórás:

$$D(W)=\sqrt{\frac{mn(m+n+1)}{12}}=\sqrt{\frac{36\cdot 29\cdot 66}{12}}\approx 76.$$

W standardizáltja $N(0, 1)$-eloszlású, a próbastatisztika aktuális értéke standardizálva

$$\frac{W-M(W)}{D(W)}=\frac{510-522}{76}\approx -0{,}16.$$

Egy $N(0, 1)$-eloszlású változó viszont 0,95 valószínűséggel $-1{,}96$ és $+1{,}96$ közé esik (l. a Student-eloszlás táblázatát, $n\to\infty$ sor), a minta alapján $-0{,}16$ az elfogadási tartományon van, nincs okunk H_0 elutasítására. Az első és második negyedéves maximális árvíztúllépések eloszlása azonosnak tekinthető.

Megjegyzés

Wilcoxon-próbát használunk két várható érték egyezésének vizsgálatára is. Az X, Y változókra n, ill. m nagy elemszámú mintát veszünk. Egyesítjük a két mintát és növekvő sorrend szerint rendezünk.

Ezután a szokásos módon kiszámítjuk a Wilcoxon-statisztika $W_{m,n}$ (vagy röviden W) értékét. A

$$H_0\colon M(X)=M(Y)$$

nullhipotézis, a

$$H_1\colon M(X)\neq M(Y)$$

kétoldali ellenhipotézis vizsgálatára kiszámítjuk W standardizáltját:

$$\frac{W-M(W)}{D(W)},$$

és u-próbát hajtunk végre. Ha

$$-u_\varepsilon\leqq\frac{W-M(W)}{D(W)}\leqq u_\varepsilon,$$

akkor a nullhipotézist elfogadjuk, különben elvetjük. Az $1-\varepsilon$ szinthez tartozó u_ε értékét vehetjük a Student-eloszlás táblázatából, az $n\to\infty$ sorból is.

Például $W=953$, $M(W)=480$, $D(W)=80$ esetén $1-\varepsilon=0{,}95$ szinthez $u_\varepsilon=1{,}96$ tartozik.

$$\frac{W-M(W)}{D(W)}=\frac{953-480}{80}=6{,}2875.$$

Mivel ez $(-1{,}96;\ 1{,}96)$ elfogadási tartományon kívül esik, nincs alapunk a két változó várható értékének egyezésére.

3.4. Korreláció- és regresszióelemzés

1. Már utaltunk arra, hogy *ha egy Y véletlen változó* sztochasztikus kapcsolatban van egy X véletlen változóval, *akkor Y adott x értéke alapján Y értékét célszerű az $M(Y/X=x)$ feltételes várható értékkel becsülni.* Ezt a becslést minden x-re elvégezve az

$$y=m_2(x)=M(Y/X=x)$$

elsőfajú regressziós függvényt kapjuk.

Ha X és Y együttes eloszlása normális, akkor az *elsőfajú regressziós függvény képe egyenes.*

A legkisebb négyzetek elve alapján Y értékeit legjobban közelítő függvény az elsőfajú regressziós függvény.

Mivel a kétdimenziós normál eloszlás a gyakorlatban sűrűn előfordul vagy jól közelíti a valóságos együttes eloszlást, ezért szívesen keresnek a két változó között lineáris kapcsolatot.

Ha a két változó együttes eloszlása nem ismert, akkor is szokás Y-t a legkisebb négyzetek elve alapján az

$$y=R\frac{\sigma_2}{\sigma_1}(x-m_1)+m_2$$

másodfajú regressziós egyenessel becsülni.

Ha az elsőfajú regressziós függvény egyenes, akkor egybeesik a másodfajúval.

Megjegyezzük, hogy két valószínűségi változó közötti regressziós görbe akkor is lehet egyenes, ha együttes eloszlásuk nem kétváltozós normál eloszlás.

A lineáris kapcsolat mérőszáma *a korrelációs együttható:*

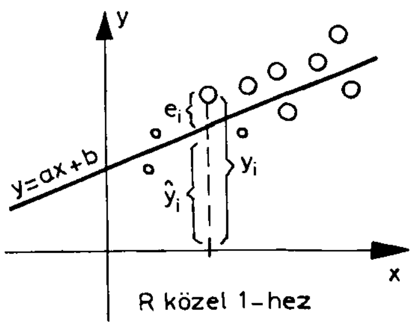
$$R=\frac{M\{[X-M(X)]\,[Y-M(Y)]\}}{D(X)D(Y)}=\frac{M(XY)-M(X)M(Y)}{D(X)D(Y)},$$

$$-1\leqq R\leqq 1,$$

minél közelebb van $|R|$ az 1-hez, annál szorosabb a lineáris kapcsolat, minél közelebb áll 0-hoz, annál lazább. $R=0$ esetén a két változó korrelálatlan, a lineáris kapcsolat laza, egyéb függvénykapcsolat azonban jól írhatja le a változók összefüggését.

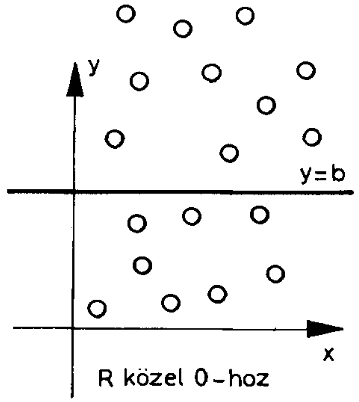
Kétdimenziós normál eloszlásnál, vagy ha X és Y a 0 és 1 értékeket felvevő indikátorváltozók, a korrelálatlanságból a függetlenség is következik.

2. Ha a két változóra csak egy n-elemű $(X_1,\ Y_1)$, $(X_2,\ Y_2)$, ..., $(X_n,\ Y_n)$ pontokból álló minta ismert, akkor a pontokat egy $(X,\ Y)$ koordináta-rendszerben ábrázolva, megállapítjuk, hogy azok egy egyenes, egy parabola, egy exponenciális görbe, egy hiperbola stb. körül csoportosulnak-e (45*a–c* ábra).



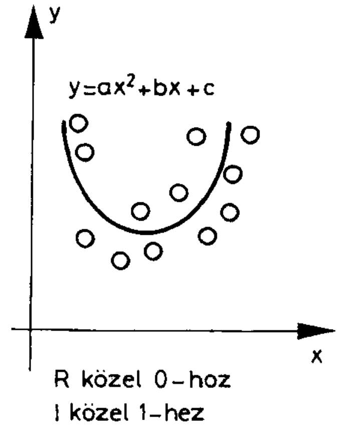
a)

45*a* ábra



b)

45*b* ábra



c)

45*c* ábra

Ha szemmelláthatóan egyenes körül, akkor *lineáris korrelációt* (véletlentől függő kapcsolatot) tételezünk fel a két változó között.

Az R elméleti korrelációs együttható mintabeli becslése r. A mintából kapott értékekkel becsülve:

$$M\{[X-M(X)]\,[Y-M(Y)]\}\approx\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n},$$

$$D(X)\approx\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n}},\qquad D(Y)\approx\sqrt{\frac{\Sigma(y_i-\bar{y})^2}{n}},$$

ezért

$$r=\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}},\tag{I}$$

ill. ugyanígy az R-re adott második képlet alapján:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. \qquad \text{(II)}$$

Az

$$y = R\frac{\sigma_2}{\sigma_1}(x-m_1)+m_2$$

másodfajú regressziós egyenes egyenletének mintabeli becslésére szintén két alakot kaphatunk:

$$y = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\,\Sigma(y_i-\bar{y})^2}}\frac{\sqrt{\frac{\Sigma(y_i-\bar{y})^2}{n}}}{\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n}}}(x-\bar{x})+\bar{y},$$

vagyis egyszerűsítve:

$$y = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x-\bar{x})+\bar{y}; \qquad \text{(I)}$$

hasonlóképpen:

$$y = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}(x-\bar{x})+\bar{y}. \qquad \text{(II)}$$

Számítástechnikailag kisebb tárigényű, gyorsabb futási idejű számításokat tesz lehetővé a (II) képletcsoport. Azonban, ha az x_i, y_i értékek nagyok, akkor könnyebb az (I) képletcsoporttal számolni, mert az $x_i-\bar{x}$, $y_i-\bar{y}$ értékek már kisebbek.

A számításokat érdemes kézzel, az alábbi táblázatos megoldással végezni:

I.	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
–	–	–	–	–	–	–	
Összesen:							
II.	x_i	y_i	x_iy_i	x_i^2	y_i^2		
	–	–	–	–	–		
Összesen:							

A szükséges szummák az egyes oszlopok alján megtalálhatók.

Zsebszámológépre, ill. személyi számítógépre alkalmas számításhoz a szükséges programokat a gyakorló feladatok közt megadjuk.

Ha az (x_1, y_1), (x_2, y_2), …, (x_n, y_n) pontpárokat „legjobban közelítő” másodfajú regressziós egyenest keresünk

$$y = ax+b$$

alakban, akkor ezt annak a követelménynek az alapján tesszük, hogy a mért és számított y ordináták négyzetes eltérésének összege a lehető legkisebb legyen:

$$Q(a, b) = \sum_{i=1}^{n}(y_i-ax_i-b)^2 \to \min,$$

ez mintabeli becslése az

$$M[(Y-aX-b)^2] \to \min$$

követelménynek (gondoljunk a *legkisebb négyzetek elvére!*). Mivel a kapott szumma a és b függvénye, minimuma ott keresendő, ahol elsőrendű parciális deriváltjai nullával egyenlők:

$$Q'_a = \Sigma 2(y_i - ax_i - b)(-x_i) = 0,$$
$$Q'_b = \Sigma 2(y_i - ax_i - b)(-1) = 0.$$

A rendezéssel kapható

$$\Sigma x_i y_i = a\,\Sigma x_i^2 + b\,\Sigma x_i,$$
$$\Sigma y_i = a\,\Sigma x_i + b\,\Sigma 1$$

normálegyenletekből a korrelációs egyenes egyenletének (II) alakját kaphatjuk. A normálegyenletek formális megjegyzése egyszerű, az

$$y_i = ax_i + b$$

egyenletből összegzéssel jön létre a második egyenlet, x_i-vel való szorzás után az

$$x_i y_i = ax_i^2 + bx_i$$

egyenletből szummázással jön létre az első egyenlet. (Ez nem tekinthető pontos levezetésnek, csupán a könnyebb megjegyzéshez útmutató.)

3. Ha az (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) mérési pontpárokat grafikonra felvéve azt tapasztaljuk, hogy azok inkább egy másodfokú parabola körül csoportosulnak (l. a 45*c* ábrát), akkor a pontokat a legkisebb négyzetek elve alapján legjobban közelítő

$$y = ax^2 + bx + c$$

parabola a, b, c együtthatóit az előbbihez hasonló módon megjegyezhető *normálegyenletekből* kaphatjuk:

$$\Sigma y_i = a\,\Sigma x_i^2 + b\,\Sigma x_i + c\,\Sigma 1,$$
$$\Sigma x_i y_i = a\,\Sigma x_i^3 + b\,\Sigma x_i^2 + c\,\Sigma x_i,$$
$$\Sigma x_i^2 y_i = a\,\Sigma x_i^4 + b\,\Sigma x_i^3 + c\,\Sigma x_i^2.$$

A legkisebb négyzetek elve itt:

$$Q(a, b, c) = \Sigma(y_i - ax_i^2 - bx_i - c)^2 \to \min.$$

Akárhányad fokú racionális egészfüggvénnyel (polinommal) való közelítésnél az ismeretlen együtthatókat az előbbi módon ismertetett normálegyenletek felírásával kiszámíthatjuk.

4. Az eddigieket *általánosítjuk*. Ha a mérési pontpárokból adódó ponthalmazt ábrázolva azt gondoljuk, hogy azt a „legjobban közelítő függvény" az

$$y = f(x)$$

függvény, akkor ez azt jelenti, hogy olyan $f(x)$ függvényt keresünk, amelyre a mért és az egyenlettel számított, becsült ordináták négyzetes eltérésének az összege minimális:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 \to \min.$$

Az ismeretlen paraméterek szerinti első deriváltak zérushelye alapján kiszámítjuk az ismeretlen paramétereket, így megkapjuk a keresett $y = f(x)$ függvényt.

A korreláció szorosságát a *korrelációs index* méri:

$$I(X, Y) = \sqrt{1 - \frac{D^2(Y - f(x))}{D^2(Y)}}, \quad 0 \leqq I \leqq 1.$$

Ennek mintabeli becslése:

$$i = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}, \quad 0 \leqq i \leqq 1,$$

ahol

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Minél közelebb van a korrelációs index 1-hez, annál pontosabban közelíti meg a ponthalmazt az $y=f(x)$ függvény, minél közelebb van a korrelációs index 0-hoz, annál rosszabb a közelítés, annál lazább a feltételezett kapcsolat. Ennek belátása egyszerű. Az Y változó adott x behelyettesítésével kapott becslése

$$\hat{Y} = f(X),$$

a mért és regresszióval számolt ordináták különbsége, az $Y-\hat{Y}$ az ún. *reziduum (maradék)*. Akkor jó a közelítés, ha a maradék szórásnégyzete közeljár 0-hoz, ekkor ($D^2(Y)>0$ miatt):

$$I = \sqrt{1-\frac{D^2(Y-\hat{Y})}{D^2(Y)}}$$

közeljár 1-hez. Akkor rossz a közelítés, ha a maradék szórásnégyzete közeljár az Y szórásnégyzetéhez, tehát ha

$$I = \sqrt{1-\frac{D^2(Y-\hat{Y})}{D^2(Y)}}$$

közeljár 0-hoz.

Ugyanez vonatkozik I mintabeli becslésére, i-re is.

$D^2(Y-\hat{Y})$ tulajdonképpen

$$M[(Y-f(X))^2],$$

amit minimalizáltunk a legkisebb négyzetek elve alapján.

$I=0$ esetén nem biztos, hogy a két változó független, csak *korrelálatlan* (az $Y=f(X)$ kapcsolat laza, a pontok rosszul közelítik az $y=f(x)$ függvényt, összevissza és erősen szóródnak körülötte).

Lineáris kapcsolat esetén I megegyezik az R korrelációs együtthatóval.

A gyakorlatban szokásos közelítések (regressziós becslések):

$y = ax+b$ lineáris kapcsolat,

$y = ax^2+bx+c$ másodfokú kapcsolat,

$y = \dfrac{a}{x+b}$ hiperbolikus kapcsolat,

$y = ae^{bx}$ exponenciális kapcsolat,

de más, bonyolultabb függvényekkel is közelítik a ponthalmazt.

Némelyik kapcsolatot könnyű visszavezetni lineárisra. Például exponenciális kapcsolat esetén:

$$\ln y = bx+\ln a,$$

ami azt mutatja, hogy x és $\ln y$ között lineáris a kapcsolat. Az $(x_1, \ln y_1), (x_2, \ln y_2), \ldots, (x_n, \ln y_n)$ pontokat már egyenessel közelíthetjük.

5. Térjünk vissza a lineáris regresszióra! A megfigyelt (mért) y_i ordináta és a regresszióval becsült (számított) $\hat{y}_i$ ordináta közötti eltérés az e_i véletlen hatás okozta hiba:

$$y_i = \hat{y}_i+e_i$$

(l. a 45*a* ábrát).

Általában az Y változó is a regressziós becslés ($\hat{Y}$) és egy $\hat{Y}$-tól független, véletlen hiba vagy más hatás okozta hiba összege:

$$Y = \hat{Y}+E.$$

Mivel független véletlen változók összegének szórásnégyzete a változók szórásnégyzetének összege, ezért

$$D^2(Y) = D^2(\hat{Y})+D^2(E),$$

vagyis: *az Y szórásnégyzete (teljes szórásnégyzet) felbontható a regresszióval megmagyarázott $\hat{Y}$ becslés szórásnégyzetének (külső szórásnégyzetnek) és az egyéb hatás, a véletlen okozta E hiba szórásnégyzetének (belső szórásnégyzetnek) az összegére.*

Ez az összefüggés a mintából adódó becslésekre is érvényes:

$$s^{*2}(y_i) = s^{*2}(\hat{y}_i)+s^{*2}(e_i),$$

$$\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{n-1} + \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-1}.$$

Hányad része a regresszióval számított értékek szórásnégyzete a mintából kapott értékek szórásnégyzetének? A regresszióból megmagyarázott külső szórásnégyzet és a teljes szórásnégyzet aránya a determinációs együttható:

$$D = \frac{D^2(\hat{Y})}{D^2(Y)}.$$

A determinációs együttható azt mutatja, hogy az Y értékek szórásnégyzetének hányad részét magyarázhatjuk a feltételezett korrelációval.

Alakítsuk át D-t:

$$D = \frac{D^2(\hat{Y})}{D^2(Y)} = \frac{D^2(Y) - D^2(E)}{D^2(Y)} = 1 - \frac{D^2(E)}{D^2(Y)} = 1 - \frac{D^2(Y-\hat{Y})}{D^2(Y)}.$$

Mivel lineáris kapcsolat esetén a korrelációs index a korrelációs együttható abszolút értékével egyenlő,

$$I = \sqrt{1 - \frac{D^2(Y-\hat{Y})}{D^2(Y)}} = |R|,$$

ezért a determinációs együttható a korrelációs együttható négyzete:

$$D = R^2.$$

Mivel $R^2 \geqq 0$, ezért $D \geqq 0$.

$$D = 1 - \frac{D^2(Y-\hat{Y})}{D^2(Y)} \geqq 0,$$

ebből következik, hogy a maradékok szórásnégyzete nem lehet nagyobb az Y szórásnégyzeténél,

$$D^2(Y-\hat{Y}) \leqq D^2(Y).$$

Az is észrevehető, hogy a nemnegatív D úgy keletkezik, hogy 1-ből egy nemnegatív számot vonunk le, ebből viszont következik, hogy

$$0 \leqq R^2 = D \leqq 1,$$

ebből viszont látható, hogy a korrelációs együttható nem kisebb (-1)-nél, nem nagyobb $(+1)$-nél:

$$-1 \leqq R \leqq 1.$$

Ugyanezek érvényesek D és R mintából kapható becsléseire is.

A külső-belső és teljes szórásnégyzetekről, determinációs együtthatóról mondottak érvényesek általános

$$Y = f(X)$$

kapcsolat esetén is. Itt $D = I^2$ lesz.

6. A lineáris kapcsolatra felírt

$$\sum_{i=1}^{n} y_i = a\Sigma x_i + b,$$
$$\sum_{i=1}^{n} x_i y_i = a\Sigma x_i^2 + b\Sigma x_i$$

normálegyenlet-rendszer megoldása a-ra, b-re *mátrixokkal is történhet*. Az induló adatokat az alábbi két mátrixba helyezzük el:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ - \\ y_n \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ - & - \\ 1 & x_n \end{bmatrix}.$$

Az ismeretlen

$$\mathbf{b} = \begin{bmatrix} b \\ a \end{bmatrix}$$

mátrixot $\mathbf{x}$ transzponáltja $(\mathbf{x}^*)$, $\mathbf{x}^*\mathbf{x}$ inverze: $(\mathbf{x}^*\mathbf{x})^{-1}$ segítségével így kapjuk meg:

$$\mathbf{b} = (\mathbf{x}^*\mathbf{x})^{-1}\mathbf{x}^*\mathbf{y}.$$

Bizonyítható, hogy *2-nél több változós lineáris regressziónál is hasonló módszerrel számíthatók ki az ismeretlen együtthatók.*

7. A lineáris korrelációnál talált

$$D = 1-\frac{D^2(E)}{D^2(Y)} = R^2$$

összefüggésből:

$$\frac{D^2(E)}{D^2(Y)} = 1-R^2,$$

$$D^2(E) = D^2(Y)\,(1-R^2).$$

Ha tehát az R korrelációs együtthatót és az Y szórásnégyzetét ismerjük, akkor ebből a képlettel a maradékok szórásnégyzete, $D^2(E)$ is kiszámítható. Sőt R és $D^2(Y)$ ismeretében a harmadik szórásnégyzet, $D^2(\hat{Y})$ is kiszámítható:

$$D^2(\hat{Y}) = D^2(Y)-D^2(E) = D^2(Y)-D^2(Y)\,(1-R^2),$$
$$D^2(\hat{Y}) = D^2(Y)R^2.$$

Ugyanezek az eredmények érvényesek a mintából kapható szórásnégyzet becslésekre is. *Ha a tapasztalati korrelációs együttható*

$$r = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)\,(\Sigma y_i^2 - n\bar{y}^2)}},$$

és az y értékek szórásnégyzete

$$s^2(y) = \frac{1}{n}\Sigma(y_i-\bar{y})^2 = \frac{1}{n}\Sigma y_i^2-\bar{y}^2$$

ismertek, a másik két szórásnégyzet ezekből egyszerűen számítható:

$$s^2(e) = s^2(y)\,(1-r^2),$$
$$s^2(\hat{y}) = s^2(y)r^2.$$

A korrigálatlan szórásnégyzet helyett persze dolgozhatunk a korrigáltakkal is.

8. A lineáris korrelációnál számított tapasztalati korrelációs együttható, r a mintaelemek függvénye, tehát maga is véletlen változó

(ahányszor veszünk n-elemű mintát, mindannyiszor különböző r értéket kapunk). Mit tudunk mondani r várható értékéről és szórásáról? Bizonyítható, hogy

$$M(r) = R+O\left(\frac{1}{n}\right), \qquad D^2(r) = \frac{(1+R^2)^2}{n}+O\left(\frac{1}{n^{3/2}}\right),$$

itt O (nagy ordó) olyan mennyiséget jelent, amelyet a mögötte zárójelben álló kifejezéssel $\left(\frac{1}{n}\text{-nel, ill. } \frac{1}{n^{3/2}}\text{-nel}\right)$ osztva, a hányados korlátos marad. Vagyis, ha $n \to \infty$, akkor $O\left(\frac{1}{n}\right)$, ill. $O\left(\frac{1}{n^{3/2}}\right)$ legalább olyan gyorsan tart 0-hoz, mint $\frac{1}{n}$, ill. $\frac{1}{n^{3/2}}$. $O\left(\frac{1}{n}\right)$ tehát legalább $\frac{1}{n}$ nagyságrendben 0-hoz tartó mennyiséget jelent, ha $n \to \infty$ (ordó = nagyságrend). Ez szemléletesen azt jelenti, hogy nagy n-re:

$$M(r) \approx R, \qquad D^2(r) \approx \frac{(1+R^2)^2}{n},$$

vagyis *az r tapasztalati korrelációs együttható az R elméleti korrelációs együtthatónak aszimptotikusan torzítatlan és aszimptotikusan erősen konzisztens becslése* (hiszen ha $n \to \infty$, $D^2(r) \to 0$).

Bizonyítható, hogy

$$r^* = r\left(1+\frac{1-r^2}{2(n-4)}\right)$$

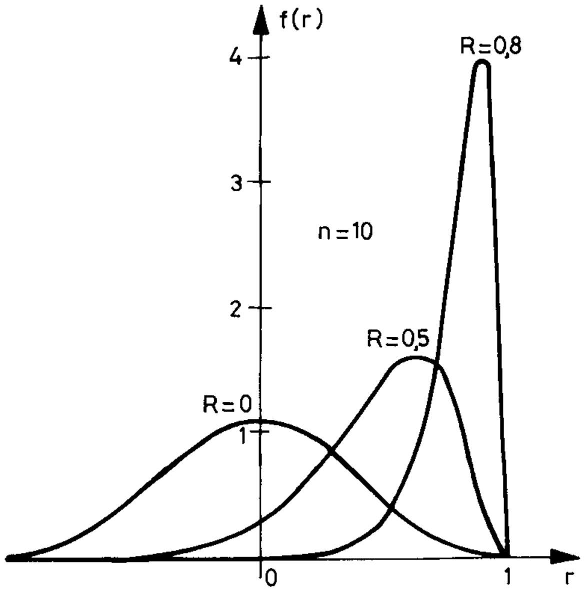
torzítatlan becslése R-nek. A gyakorlatban azonban inkább r-rel számolnak, ami nem okoz nagy kárt, hiszen $r^* > r$, ha $n > 4$.

Milyen r eloszlása? Ez X és Y együttes eloszlásától függ.

Ha X és Y együttes eloszlása kétdimenziós normál eloszlás, akkor r sűrűségfüggvénye:

$$f(r) = \frac{n-2}{\pi}(1-R^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}}\int_0^1 \frac{x^{n-2}\,dx}{(1-Rxr)^{n-1}\sqrt{1-x^2}},$$

természetesen $-1 \leqq r \leqq 1$ az értelmezési tartomány. Az $n=10$, $R=0$, $R=0{,}5$, $R=0{,}8$ értékeknek megfelelő sűrűségfüggvényeket tartalmazza a 46. ábra. Feltűnő, hogy az $R=0$ esetben (ami kétváltozós normál eloszlás esetén X és Y függetlenségét jelenti) mennyire hasonlít a görbe a haranggörbéhez.



46. ábra

Amikor nullhipotézisünk a két változó függetlensége,

$$H_0\colon\ P(X<x,\ Y<y) = P(X<x)P(Y<y),$$

vagyis

$$H_0\colon\ R=0,$$

akkor az ábráról is látható, hogy legnagyobb valószínűséggel r-re is 0 körüli értéket kapunk, de nem biztos, hogy éppen nullát. ***Hogyan tudjuk tesztelni, hogy mekkora r tapasztalati korrelációs együttható esetén fogadható el a függetlenség, vagyis hogy az egész sokaságban az elméleti R = 0?***

Ha X és Y együttes eloszlása kétdimenziós normál eloszlás, r sűrűségfüggvénye egyszerűsödik:

$$f(r) = \frac{1}{\sqrt{\pi}}\,\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}}.$$

Ennek segítségével megmutatható, hogy *kétdimenziós normál eloszlás, $R=0$ esetén a*

$$t = \frac{\sqrt{n-2}\,r}{\sqrt{1-r^2}}$$

statisztika Student-eloszlású, $n-2$ paraméterrel, a Student-eloszlás táblázatából meghatározható az adott $1-\varepsilon$ szinthez olyan $(-t_\varepsilon, t_\varepsilon)$ intervallum, amelybe t nagy, $1-\varepsilon$ valószínűséggel esik. Ha t aktuális értéke az elfogadási $(-t_\varepsilon, t_\varepsilon)$ tartományra esik, az $R=0$ hipotézist elfogadjuk ($r \neq 0$ csak véletlen), ha az elfogadási tartományon kívül esik, az $R=0$ feltevést (a függetlenséget) elutasítjuk. Tehát szabályszerű t-próbát hajtunk végre.

Kétdimenziós normál eloszlás esetén az előbbi próba helyett használhatjuk a 15. táblázatot, amelyben az van megadva, hogy ha a nullhipotézis $R=0$, akkor maximálisan mekkora lehet r a feltevés elfogadásához. Ha ennél a kritikus értéknél a mintabeli korrelációs együttható abszolút értéke kisebb, az $R=0$ feltevés elfogadható, ha nagyobb, elutasítható. Ha például $n=18$ esetén

$$|r| \leqq 0{,}468,$$

akkor az elméleti korrelációs együtthatónk az $R=0$ feltevés (a két változó korrelálatlansága) elfogadható; ha

$$|r| > 0{,}468,$$

az $R=0$ feltevés elvethető 95%-os biztonsági szinten, l. a 15. táblázat részletét:

n	95%
18	0,468

Ha X és Y együttes eloszlása nem kétdimenziós normál eloszlás, akkor célszerű elvégezni az R. A. Fisher által bevezetett

$$Z = \frac{1}{n} \ln \frac{1+r}{1-r}$$

transzformációt. A Z valószínűségi változó nagy n-re eléggé általános feltételek mellett közel normál eloszlású,

$$M(Z) \approx \frac{1}{2} \ln \frac{1+R}{1-R} + \frac{R}{2(n-1)}$$

várható értékkel,

$$D^2(Z) \approx \frac{1}{n-3)}$$

szórásnégyzettel. *Az $R=0$ feltevés vizsgálatára közelítőleg megbízható eredményt kapunk* a normál eloszlás alkalmazásával.

9. Gyakori probléma, hogy két mérési helyen vagy két alkalommal végzünk ellenőrzést arra, hogy két változó között milyen szoros a korreláció. A két mintából kapunk egy r_1, ill. r_2 korrelációs együtthatót. Ezeknek mekkora eltérése esetén vetjük el azt a feltevést, hogy az elméleti korrelációs együtthatók megegyeznek, vagyis hogy a korreláció ugyanolyan szoros a két esetben?

A problémát kissé általánosítva fogalmazzuk meg. *Két tapasztalati korrelációs együtthatót így hasonlíthatunk össze:*

Az (X, Y) változópárra vett n_1-elemű minta:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_{n_1}, Y_{n_1}),$$

az (U, V) változópárra vett n_2-elemű minta:

$$(U_1, V_1), (U_2, V_2), \ldots, U_{n_2}, V_{n_2}).$$

Legyen az (X, Y) változópár korrelációs együtthatója R_1, ennek mintabeli becslése r_1, az (U, V) változópár korrelációs együtthatója R_2, ennek mintabeli becslése r_2.

Vizsgáljuk a H_0: $R_1 = R_2$ hipotézist!

A Fisher-féle transzformáció alapján:

$$Z_1 = \frac{1}{2} \ln \frac{1+r_1}{1-r_1}, \qquad Z_2 = \frac{1}{2} \ln \frac{1+r_2}{1-r_2}$$

valószínűségi változók jó közelítésben normál eloszlásúak

$$D^2(Z_1) = \frac{1}{n_1-3} \quad \text{és} \quad D^2(Z_2) = \frac{1}{n_2-3}$$

szórásnégyzettel, a nullhipotézis teljesülése esetén a várható értékek azonosak, így ezért a

$$W = \frac{Z_1 - Z_2}{\sqrt{\dfrac{1}{n_1-3} + \dfrac{1}{n_2-3}}}$$

véletlen változó standard normál eloszlású. Ha tehát

$$-1{,}96 \leqq W \leqq 1{,}96,$$

akkor a H_0 hipotézist 95%-os biztonsági szinten elfogadhatjuk; különben elvetjük.

Kettőnél több tapasztalati korrelációs együtthatót így hasonlíthatunk össze: Legyenek az $n_1, n_2, \ldots, n_k$ elemszámú mintákból számított tapasztalati korrelációs együtthatók: $r_1, r_2, \ldots, r_k$, a megfelelő elméleti korrelációs együtthatók: $R_1, R_2, \ldots, R_k$. A

$$H_0\colon R_1 = R_2 = \ldots = R_k$$

nullhipotézis ellenőrzésére számítsuk ki a

$$\chi^2 = \sum_{i=1}^{n} (n_i - 3)(Z_i - \bar{Z})^2$$

próbastatisztika aktuális értékét, ahol

$$Z_i = \frac{1}{2}\ln\frac{1+r_i}{1-r_i}, \qquad \bar{Z} = \frac{\sum_{i=1}^{k}(n_i-3)Z_i}{\sum_{i=1}^{k}(n_i-3)}.$$

A próbastatisztika $k-1$ paraméterű, χ^2-eloszlású. A nullhipotézis vizsgálatára χ^2-próbát végezhetünk. Ha χ^2 aktuális értéke nem nagyobb, mint az adott $1-\varepsilon$ szinthez tartozó kritikus érték, a nullhipotézist (a korrelációk szorosságának azonos voltát) elfogadjuk; ellenkező esetben elvetjük.

k értékének az összes mintadarabszámhoz képest kicsinek kell lennie, mert Z várható értékének $\bar{Z}$ becslésében az $\frac{R}{2(n-1)}$ korrekciós tagot elhanyagoltuk.

10. Kétváltozós lineáris regresszió esetén az elméleti regressziós függvény:

$$Y = \alpha X + \beta + E,$$

itt $\hat{Y} = \alpha X + \beta$ a regresszió által megmagyarázott rész, E a véletlen hatás (vagy más változó) okozta hiba (maradék, reziduum). A regresszió által megmagyarázott rész mintából nyert becslőfüggvénye:

$$\hat{y} = ax + b.$$

Az elméleti lineáris regressziós egyenes α iránytangensének torzítatlan becslése a mintából:

$$a = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

az egyenes y tengelymetszetének, β-nak mintabeli torzitatlan becslése:

$$b = -a\bar{x} + \bar{y},$$

$\hat{y}$ pedig torzítatlan becslése Y feltételes várható értékének ($y = \hat{y} + e$, az e maradék várható értéke 0).

Az a, b, $\hat{y}$, e maga is véletlen változó (a mintaelemek függvénye). *Keressük meg a lineáris regressziónál szereplő véletlen változók pontosságának jellemzésére ezek szórásnégyzetét!*

Y szórásnégyzetének mintából kapható becslése:

$$D^2(Y) \approx s^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2;$$

a regressziós becslés szórásnégyzetének mintából nyert becslése:

$$D^2(\hat{Y}) \approx s^2(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \frac{1}{n}\Sigma\hat{y}_i^2 - \bar{y}^2;$$

a mért és számított értékek eltérésének, az ún. reziduumoknak (maradékoknak) a szórásnégyzetbecslése:

$$D^2(E) \approx s^2(e) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

Bizonyítottuk, hogy ha $s^2(y)$ és az r tapasztalati korrelációs együttható ismert, akkor a másik két szórásnégyzet ezekből kiszámítható:

$$s^2(e) = s^2(y)(1-r^2),$$
$$s^2(\hat{y}) = s^2(y)r^2.$$

Az X szórásnégyzetének mintából eredő becslése:

$$D^2(X) \approx s^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2.$$

Bizonyítható, hogy $s^2(e)$ mintából kapott becslés az elméleti $D^2(e)$-re torzított; torzitatlan becslést így kapunk:

$$s^2(e) = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}.$$

Nagy n-nél ennek nincs jelentősége, kis n-nél van.

Az $\hat{y} = ax+b$ egyenletnél az a iránytangens szórásnégyzete:

$$s^2(a) = \frac{s^2(e)}{ns^2(x)}.$$

Tudjuk, hogy ha X és Y függetlenek, akkor a regressziós függvény:

$$y=\beta \text{ (állandó)},$$

ez az x tengellyel párhuzamos egyenes, iránytangense $\alpha=0$. Nem biztos, hogy ilyen esetben az iránytangens mintabeli becslése, a is 0 lesz (csak várhatóan 0 körüli érték).

Milyen konfidencia-intervallum tartalmazza az elméleti iránytangenst nagy valószínűséggel? A

$$t = \frac{a-\alpha}{s(a)}$$

változó Student-eloszlású, $n-2$ szabadságfokkal. Az $1-\varepsilon$ valószínűséggel fennálló

$$-t_\varepsilon \leqq \frac{a-\alpha}{s(a)} \leqq t_\varepsilon$$

egyenlőtlenségből

$$a-t_\varepsilon s(a) \leqq \alpha \leqq a+t_\varepsilon s(a)$$

konfidencia-intervallum kapható az α *elméleti iránytangensre.* Ez a mérési eredményekből kapható a-ra szimmetrikus intervallum.

– Hasonlóképpen az y tengelymetszet b szórásnégyzete

$$s^2(b) = \frac{s^2(e)\sum_{i=1}^{n} x_i^2}{n^2s^2(x)},$$

ennek segítségével *az alábbi,* $1-\varepsilon$*-szintű konfidencia-intervallum kapható az elméleti y tengelymetszetre,* β*-ra:*

$$b-t_\varepsilon s(b) \leqq \beta \leqq b+t_\varepsilon s(b),$$

ahol t_ε a Student-eloszlás $n-2$ szabadságfokhoz, $(1-\varepsilon)\cdot 100\%$ biztonsági szinthez tartozó kritikus értéke.

Mindkét konfidencia-intervallumból t-próbát is csinálhatunk.

– Az y értékeit a mintából nyert regressziós egyenletből számított

$$\hat{y} = ax+b$$

értékkel becsüljük. *Egy adott* x_i*-hez tartozó* $M_i = M(Y/X=x_i)$ *feltételes várható értéket* $1-\varepsilon$ *valószínűséggel tartalmazza az* $\hat{y}_i$*-re szimmetrikus, alábbi konfidencia-intervallum:*

$$\hat{y}_i - t_\varepsilon s(e)\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{ns^2(x)}} \leqq M_i \leqq \hat{y}_i + t_\varepsilon s(e)\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{ns^2(x)}},$$

ahol t_ε a Student-eloszlás táblázatában $(1-\varepsilon)\cdot 100\%$ biztonsági szinthez, $n-2$ szabadságfokhoz tartozó kritikus érték.

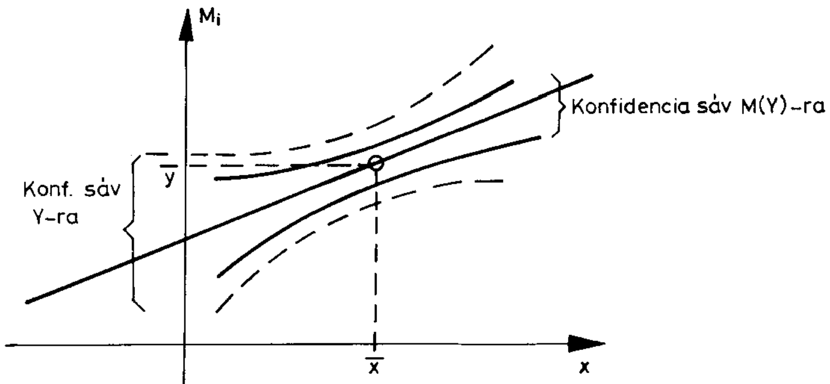
Az adott x_i*-hez tartozó* Y_i *egyedi értéket tartalmazza az alábbi,* $\hat{y}_i$*-ra szimmetrikus konfidencia-intervallum* $1-\varepsilon$ *valószínűséggel:*

$$\hat{y}_i - t_\varepsilon s(e)\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{ns^2(x)}} \leqq Y_i \leqq \hat{y}_i + t_\varepsilon s(e)\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{ns^2(x)}},$$

itt t_ε az előbbi módon a Student-eloszlás táblázatából nyerhető érték.

Minden egyes x_i értékhez kiszámíthatjuk mindkét fajta konfidencia-intervallumot. A különböző x_i értékeknél számított pontok *egy-egy konfidenciasávot,* egy-egy hiperbolát alkotnak a regressziós egyenes körül (47. ábra), hiszen például az első esetben

$$\frac{(M_i-y_i)^2}{\left(\frac{t_\varepsilon s(e)}{\sqrt{n}}\right)^2} - \frac{(x_i-\bar{x})^2}{s^2(x)} = 1$$



47. ábra

egy hiperbola egyenlete az (M_i, x_i) koordináta-rendszerben. A szélesebb sáv nyilván a második, az egyedi értékre kapott. A konfidencia-intervallum szélessége $x_i = \bar{x}$ értéknél éri el minimumát. Mivel a regressziós becslés átmegy az $(\bar{y}, \bar{x})$ ponton az előbbi koordináta-rendszerben (l. az $y = a(x-\bar{x})+\bar{y}$ alakú egyenletet), ez azt jelenti, hogy a regresszióval kapott becslés a legmegbízhatóbb az $(\bar{y}, \bar{x})$ súlypont körül; ha x_i távolodik $\bar{x}$-tól jobbra és balra, a sáv szélesedik, a becslés kevésbé megbízható.

11. Az eddigiek során két változót tartalmazó regressziós feladatról beszéltünk. Most *a kettőnél több változót tartalmazó regressziót* tárgyaljuk meg. (A feladat elméleti részét az I. rész 7. fejezete tartalmazza.)

Azt kell megvizsgálnunk, hogy az ott szereplő mennyiségeket hogyan számítjuk ki az $X_1, X_2, \ldots, X_n$ változókra kapott $n \times s$ elemű minta segítségével:

A mérési eredmények:

$$\begin{array}{r} X_1\text{-re:} \\ X_2\text{-re:} \\ --- \\ X_n\text{-re:} \end{array} \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1s} \\ x_{21} & x_{22} & \ldots & x_{2s} \\ - & - & - & - \\ x_{n1} & x_{n2} & \ldots & x_{ns} \end{bmatrix}$$

Akár elsőfajú, akár másodfajú

$$x_1 = a_1 + b_{12}x_2 + b_{13}x_2 + \ldots + b_{1n}x_n$$

regressziós „sík" esetén teljesülnie kell a legkisebb négyzetek elvének:

$$M[(X_1 - a_1 - b_{12}X_2 - \ldots - b_{1n}X_n)^2] \to \min.$$

Ha X és Y együttes eloszlása kétdimenziós normál eloszlás, az első- és másodfajú regressziós sík azonos.

A minimumfeladat megoldásaként az alábbi „hipersíkot" (regressziós síkot) kapjuk.

I. A regressziós sík kovarianciákkal kifejezett alakja:

$$x_1 = m_1 - \sum_{k=2}^{n} \frac{C_{1k}}{C_{11}}(x_k - m_k),$$

II. a regressziós sík korrelációs együtthatókkal kifejezett alakja:

$$x_1 = m_1 - \sum_{k=2}^{n} \frac{\sigma_1}{\sigma_k} \frac{R_{1k}}{R_{11}}(x_k - m_k).$$

Itt $m_k = M(X_k)$, $\sigma_k = D(X_k)$, C_{ik} és R_{ik} pedig az $X_1, X_2, \ldots, X_n$ változók

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ - & - & - & - \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{bmatrix}$$

kovarianciamátrixának, ill.

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ r_{21} & r_{22} & \ldots & r_{2n} \\ - & - & - & - \\ r_{n1} & r_{n2} & \ldots & r_{nn} \end{bmatrix}$$

korrelációmátrixának c_{ik}, ill. r_{ik} eleméhez tartozó előjeles aldeterminánsai. Az X_i és X_k változók c_{ik} *kovarianciájának mintaelemekből kapható torzítatlan becslése* (a $c_{ik} = M[(X_i - m_i)(X_k - m_k)] = M(X_i X_k) - m_i m_k$ képletekből könnyen fejben tartható mindkét alak):

$$\text{I. } c_{ik} \approx \frac{1}{s-1}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k) \text{ vagy}$$

$$\text{II. } c_{ik} \approx \frac{1}{s-1}\sum_{j=1}^{s} x_{ij}x_{kj} - \frac{s}{s-1}\bar{x}_i\bar{x}_k.$$

Az X_i és X_k változók totális korrelációs együtthatójának mintából kapható becslése abból kapható meg, hogy

$$r_{ik} = \frac{c_{ik}}{\sigma_i\sigma_k} = \frac{c_{ik}}{\sqrt{c_{ii}}\sqrt{c_{kk}}},$$

c_{ik}, c_{ii} és c_{kk} mintabeli becslései pedig az előbbi képlettel megkaphatók. Természetesen $r_{ii}=1$. Vagy pedig az elméleti korrelációs együttható két alakjából, az

$$r_{ik} = \frac{M[(X_i-m_i)(X_k-m_k)]}{D(X_i)D(X_k)} = \frac{M(X_iX_k)-m_im_k}{D(X_i)D(X_k)}$$

képletekből könnyen megjegyezhető r_{ik} *torzított becslésének két alakja:*

$$\text{I. } r_{ik} \approx \frac{\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k)}{\sqrt{\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\sum_{j=1}^{s}(x_{kj}-\bar{x}_k)^2}} \text{ vagy}$$

$$\text{II. } r_{ik} \approx \frac{\sum_{j=1}^{s}x_{ij}x_{kj}-s\bar{x}_i\bar{x}_k}{\sqrt{\left(\sum_{j=1}^{s}x_{ij}^2-s\bar{x}_i^2\right)\left(\sum_{j=1}^{s}x_{kj}^2-s\bar{x}_k^2\right)}}.$$

Számítástechnikailag a II. képletekkel gyorsabb és kisebb tárigényű a számolás.

Önkényes, hogy éppen az első változót fejeztük ki a többi segítségével. Ugyanígy kapható, hogy *bármelyik X_i változó regressziós becslése a többi változó segítségével:*

$$\hat{X}_i = m_i - \sum_{\substack{k=1\\k\neq i}}^{n}\frac{C_{ik}}{C_{ii}}(X_k-m_k) = m_i - \sum_{\substack{k=1\\k\neq i}}^{n}\frac{\sigma_i}{\sigma_k}\frac{R_{ik}}{R_{ii}}(X_k-m_k).$$

Az $Y_i = X_i-\hat{X}_i$ maradék pozitívan korrelált X_i-vel, korrelációs együtthatójuk:

$$R(Y_i, X_i) = \sqrt{\frac{|C|}{c_{ii}C_{ii}}} > 0,$$

ugyanakkor az $Y_i = X_i-\hat{X}_i$ maradék korrelálatlan a többi $(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ változóval. Itt $|C|$ a kovarianciamátrix determinánsát jelenti.

Az $X_i-\hat{X}_i$ maradék szórásnégyzete:

$$D^2(X_i-\hat{X}_i) = \frac{|C|}{C_{ii}} = \sigma_i^2\frac{|R|}{R_{ii}}$$

méri az X_i változó $\hat{X}_i$ alapján történő becslésének hatásosságát. $|R|$ a korrelációs együtthatók $\mathbf{R}$ mátrixának determinánsa.

A most bemutatott mennyiségek mintából kapható becslése a c_{ik}, ill. r_{ik} értékek mintabeli becslése alapján elkészíthető.

Az r_{ik} *totális (teljes) korrelációs együttható* az X_i és X_k közötti *teljes közvetett kapcsolat erősségének a mérőszáma.* Ebbe a kapcsolatba belejátszik a többi változó is. Az X_i és X_k változót elvéve megállapítjuk ezek $\hat{X}_i$ és $\hat{X}_k$ legjobb (regressziós) becslését. Az $X_i-\hat{X}_i$ maradék korrelálatlan a két változó elvétele után megmaradt többi változóval, és pozitívan korrelált X_i-vel. Ugyanígy az $X_k-\hat{X}_k$ maradék korrelálatlan a két változó elvétele után megmaradt többi változóval, és pozitívan korrelált X_k-val. Ezért az $X_i-\hat{X}_i$ és $X_k-\hat{X}_k$ maradékok korrelációs együtthatója X_i és X_k *közvetlen kapcsolatát méri, ez az ún. parciális korrelációs együttható:*

$$\varrho_{ik} = -\frac{C_{ik}}{\sqrt{C_{ii}C_{kk}}} = -\frac{R_{ik}}{\sqrt{R_{ii}R_{kk}}}.$$

Itt C_{ik} a $\mathbf{C}$ kovarianciamátrix c_{ik} eleméhez tartozó előjeles aldeterminánsa, R_{ik} az $\mathbf{R}$ korrelációmátrix r_{ik} eleméhez tartozó előjeles aldeterminánsa. Az aldeterminánsok elemeinek mintából való becslését már megadtuk.

X_1, X_2, X_3 változó esetén a parciális korrelációs együtthatók képletét áttekinthető, egyszerű alakra lehet rendezni (l. a „Többdimenziós eloszlások; korreláció" c. fejezetet). Például (használva az ott említett Yule-féle jelölést):

$$r_{12\cdot 3} = \varrho_{12} = \frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \quad \text{stb.}$$

Hasonló egyszerű képlet adható meg még 4 változó esetén is.

Készítsük el X_i regressziós becslését a többi változóval (most egyszerűbben jelöljük a regressziós együtthatókat):

$$\hat{X}_i = a+b_1X_1+b_2X_2+\ldots+b_{i-1}X_{i-1}+b_{i+1}X_{i+1}+\ldots+b_nX_n.$$

X_i és $\hat{X}_i$ korrelációs együtthatója *a többszörös korrelációs együttható:*

$$\varrho_i = \sqrt{1-\frac{|C|}{\sigma_i^2 C_{ii}}} = \sqrt{1-\frac{|R|}{R_{ii}}}.$$

Ez mutatja X_i változó kapcsolatának szorosságát a többi ($X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$) változók együttesével, vagyis a regressziós becslés jóságát. Minél közelebb esik 0-hoz, annál lazább a kapcsolat, minél közelebb esik 1-hez, annál szorosabb.

Mivel a ϱ_i többszörös korrelációs együttható úgy is megadható, mint a regresszióval számított és a mért értékek szórásának az aránya, ezért ϱ_i más alakban is megadható. Ugyanis

$$X_i = \hat{X}_i+E,$$
$$D^2(X_i) = D^2(\hat{X}_i)+D^2(E),$$
$$\varrho_i = \frac{D(\hat{X}_i)}{D(X_i)} = \frac{\sqrt{D^2(X_i)-D^2(E)}}{D(X_i)},$$

vigyük be a gyökjel alá az $\frac{1}{D(X_i)}$ szorzót (a gyökjel alatt a négyzetével kell szorozni), így kapjuk a többszörös korrelációs együttható harmadik szokásos alakját:

$$\varrho_i = \sqrt{1-\frac{D^2(E)}{D^2(X_i)}}.$$

$D^2(E)$ és $D^2(X_i)$ szórásnégyzeteket a mintából becsülve, a fenti képlet megadja ϱ_i torzítatlan becslését.

ϱ_i negatív nem lehet, hiszen $X_i = \hat{X}_i+E$ pozitív korrelációt mutat X_i és $\hat{X}_i$ között.

A többszörös korrelációs együttható négyzete a már megismert *determinációs együttható.*

Számítógépes megoldás szempontjából a determinánsok és aldeterminánsok számítása nem könnyű, és nem rövid program. Az X_i regressziós becslésére kapott regressziós együtthatókat szokták a normálegyenlet-rendszer mátrixos megoldásával, mátrixaritmetikával megkeresni (különösen, ha mátrixaritmetika szubrutin a számítógépbe huzalozva van). A többszörös korrelációs együtthatók szintén számíthatók mátrixaritmetikával. Ugyanis az **R** korrelációmátrix inverze:

$$\mathbf{R}^{-1} = \begin{bmatrix} \frac{R_{11}}{|R|} & \frac{R_{12}}{|R|} & \cdots & \frac{R_{1n}}{|R|} \\ \hline \frac{R_{n1}}{|R|} & \frac{R_{n2}}{|R|} & \cdots & \frac{R_{nn}}{|R|} \end{bmatrix}.$$

Innen pl. X_1 és $\hat{X}_1$ becslés korrelációs együtthatója (az egyik többszörös korrelációs együttható):

$$\varrho_1 = \sqrt{1-\frac{|R|}{R_{11}}},$$

ha 1-ből kivonjuk a fenti inverz mátrix bal felső elemének reciprokát és a különbségből négyzetgyököt vonunk, ϱ_1-et kapjuk meg. Ugyanígy kapható a többi, többszörös korrelációs együttható is.

Ha sok változó van, a vizsgálatba esetleg nem érdemes minden változót bevonni. Melyik változót érdemes bevonni a regresszióba? Azt, amelyiknek nagyobb a hatása a kiemelt változóra, amelyiknek nagyobb a súlya. De mivel mérjük ezt a hatást (súlyt)? *A maradék szórásnégyzetével.* Ugyanis minél kisebb ez a szórásnégyzet, annál nagyobb a regresszió által megmagyarázott rész, annál lényegesebb az újonnan bevont változó hatása a kiszemelt változóra.

Ha például az $X_1, X_2, \ldots, X_n$ változók közül X_i-re akarunk regressziós becslést készíteni, akkor először gondolatban elkészítjük X_i és X_1, X_i és $X_2, \ldots, X_i$ és X_n kétváltozós regresszióit, kiszámítva mindegyikkel a

$$D^2(X_i - \hat{X}_i) = \frac{|C|}{C_{ii}} = \sigma_i^2 \frac{|R|}{R_{ii}}$$

maradék szórásnégyzetet. Legnagyobb hatású az a változó, amelyiknél ez a szórásnégyzet a legkisebb az $n-1$ darab szórásnégyzet közül. Legyen ez mondjuk X_1. A gép ezekután ténylegesen megállapítja X_i regressziós becslőfüggvényét X_1-gyel.

Most megpróbál újabb változót bevonni a megmaradt $n-2$ változó közül. Elkészíti az $X_i - \hat{X}_i$ maradék szórásnégyzetét, mind az $(n-2)$-t. Amelyiknek legkisebb a szórásnégyzete, azt vonja be másodiknak a regresszióba. Legyen ez mondjuk X_2.

Ezekután a gép kiszámítja X_i regressziós becslését X_1-gyel és X_2-vel, és így tovább, fokozatosan bevonva az újabb változókat. Legutoljára a legkisebb súlyú (hatású) változót vonja be, megadva a most már n változót tartalmazó regressziós becslőfüggvényt, és az összes jellemzőt.

A kiírt maradék szórásnégyzetek alapján magunk választhatjuk ki, meddig megyünk el a változók bevonásában.

Mivel

$$D^2(X_i) = D^2(\hat{X}_i) + D^2(X_i - \hat{X}_i),$$

ezért, ha $D^2(X_i - \hat{X}_i)$ minimális, akkor a regresszió által megmagyarázott szórásnégyzet, $D^2(\hat{X}_i)$ maximális. Ez utóbbinak $D^2(X_i)$-hez való aránya éppen a többszörös korrelációs együttható négyzete, vagyis a determinációs együttható:

$$\varrho_i^2 = \frac{D^2(\hat{X}_i)}{D^2(X_i)}.$$

Minél nagyobb a determinációs együttható, annál nagyobb a regresszió által megmagyarázott rész (a külső szórásnégyzet) a teljes

szórásnégyzethez viszonyítva. *A változó súlyát mérhetjük a többszörös korrelációs együttható négyzetével (a determinációs együtthatóval), amelyik X_i mellé bevont új változó determinációs együtthatója (vagy annak mintabeli becslése) nagyobb, annak a hatása (súlya) nagyobb, azt érdemesebb bevonni.*

A most vázolt elv alapján működnek a STEP-WISE MULTIPLE REGRESSION (lépésenkénti többváltozós regresszió) programcsomagok. Az IBM 370/145 gép szoftverjében megtalálható STEP-WISE MULTIPLE REGRESSION programcsomag használatát ismertetjük a most következő gyakorló feladatok közt.

A maradék szórásnégyzete tulajdonképpen a legkisebb négyzetek elvénél kitűzött, minimális várható értéket adja meg, ezért jellemzi a becslés hatásos voltát.

12. Vizsgáljuk meg, hogy *a többváltozós regressziónál a legkisebb négyzetek elvét a mintából kapott mérési eredmények felhasználásával hogyan valósítjuk meg, és a regressziós együtthatókra milyen normálegyenletrendszert kapunk!*

Legyen X_1 regressziós becslése a többi változó lineáris függvénye, azaz

$$\hat{X}_1 = a_1 + b_2 X_2 + b_3 X_3 + \ldots + b_n X_n.$$

Adott az n darab változóra s darab megfigyelési pont az n-dimenziós térben:

A mérés sorszáma:

$$\begin{array}{r c}
 & \begin{array}{cccc} 1 & 2 & \ldots & s \end{array} \\
\begin{array}{r} X_1\text{-re} \\ X_2\text{-re} \\ - \; - \\ X_n\text{-re} \end{array} & \left[\begin{array}{cccc} x_{11} & x_{12} & \ldots & x_{1s} \\ x_{21} & x_{22} & \ldots & x_{2s} \\ - & - & - & - \\ x_{n1} & x_{n2} & \ldots & x_{ns} \end{array}\right].
\end{array}$$

Ez részletesen azt jelenti, hogy a következő egyenletrendszer adott:

$$\begin{array}{ll}
x_{11} = a_1 + b_2 x_{21} + \ldots + b_n x_{n1} & (1.\ \text{mérés}), \\
x_{12} = a_1 + b_2 x_{22} + \ldots + b_n x_{n2} & (2.\ \text{mérés}), \\
- - - - - - - - - - - - & \\
\underline{x_{1s} = a_1 + b_2 x_{2s} + \ldots + b_n x_{ns}} & (s\text{-edik mérés}).
\end{array}$$

Az egyenletrendszer i-edik egyenlete:

$$x_{1i} = a_1 + b_2 x_{2i} + \ldots + b_n x_{ni} \quad (i\text{-edik mérés, } i = 1, 2, \ldots, s).$$

Olyan $a_1, b_2, b_3, \ldots, b_n$ regressziós együtthatókat keresünk, amelyekre a mért és számított értékek eltéréséből vett várható érték (vagyis a maradék szórásnégyzete) a legkisebb:

$$M[(X_1 - \hat{X}_1)^2] = M[(X_1 - a_1 - b_2 X_2 - \ldots - b_n X_n)^2] \to \min.$$

A várható értéket pontosan kiszámítani nem tudjuk, mert nem ismert annak a változónak az eloszlása, amelynek a várható értékét minimalizáljuk. Vesszük a várható érték mintabeli becslését, ez egy összeg osztva az összeadandók számával, s-sel. A kapott érték s-szeresét elég minimalizálni. Ez az összeg az ismeretlen $a_1, b_2, \ldots, b_n$ regressziós együtthatók függvénye:

$$Q(a_1, b_2, \ldots, b_n) = \sum_{i=1}^{s} (x_{1i} - a_1 - b_2 x_{2i} - \ldots - b_n x_{ni})^2 \to \min.$$

Ez most a legkisebb négyzetek elvének mintabeli becslésből adódó alakja.

A többváltozós függvénynek szélsőértéke ott lehet, ahol parciális deriváltjai nullák. Biztosan van, ha ugyanott a Hesse-determinánssal kapcsolatos feltételek teljesülnek. Bizonyítható (mint azt a kétváltozós regressziónál megtettük), hogy a parciális deriváltak zérushelyei valóban a kívánt minimumot adó $a_1, b_2, \ldots, b_n$ regressziós együtthatókat szolgáltatják (a következőkben a szummázás mindig $i = 1$-től s-ig megy, egyesével):

$$\frac{\partial Q}{\partial a_1} = 2\sum (x_{1i} - a_1 b_2 x_{2i} - \ldots - b_n x_{ni})(-1) = 0,$$
$$\frac{\partial Q}{\partial b_2} = 2\sum (x_{1i} - a_1 - b_2 x_{2i} - \ldots - b_n x_{ni})(-x_{2i}) = 0,$$
$$\cdots$$
$$\frac{\partial Q}{\partial b_n} = 2\sum (x_{1i} - a_1 - b_2 x_{2i} - \ldots - b_n x_{ni})(-x_{ni}) = 0.$$

Ezekből átrendezéssel kapható az ún. *Gauss-féle normál-egyenletrendszer:*

$$\sum x_{1i} = a_1 \sum 1 + b_2 \sum x_{2i} + \ldots + b_n \sum x_{ni},$$
$$\sum x_{1i} x_{2i} = a_1 \sum x_{2i} + b_2 \sum x_{2i} x_{2i} + \ldots + b_n \sum x_{ni} x_{2i},$$
$$\cdots$$
$$\sum x_{1i} x_{ni} = a_1 \sum x_{ni} + b_2 \sum x_{2i} x_{ni} + \ldots + b_n \sum x_{ni} x_{ni}.$$

A kapott normál-egyenletrendszer formális megjegyzése igen egyszerű. Az

$$x_{1i} = a_1 + b_2 x_{2i} + \ldots + b_n x_{ni}$$

alapegyenletből a normál-egyenletrendszer első egyenlete úgy keletkezik, hogy minden tag helyett az összegét vesszük $i = 1$-től s-ig, az alapegyenletet x_{2i}-vel végigszorozva az

$$x_{1i} x_{2i} = a_1 x_{2i} + b_2 x_{2i}^2 + \ldots + b_n x_{ni} x_{2i}$$

egyenletből a már említett összegzéssel kapjuk a normál-egyenletrendszer második egyenletét, és így tovább, az alapegyenletet x_{ni}-vel szorozva az így nyert

$$x_{1i} x_{ni} = a_1 x_{ni} + b_2 x_{2i} x_{ni} + \ldots + b_n x_{ni}^2$$

egyenletből „szummázással" kapjuk a normál-egyenletrendszer n-edik egyenletét (mint a kétváltozós esetben).

Az egyenletrendszert bármilyen módszerrel számítógéppel megoldhatjuk (determinánsokkal, mátrixaritmetikával, Gauss-féle eliminációs módszerrel stb.). Használhatjuk a kovariancia-, ill. korrelációmátrix elemeiből adódó előbbi képleteket is. Utóbbiak egyikének kiszámítása azért is jó, mert a totális, parciális és többszörös együtthatókat is, a maradék szórásnégyzet becslését is megkaphatjuk a segítségükkel.

Az egyenletrendszer megoldható mátrixok segítségével is (l. a következő gyakorló feladatokat).

A kapott $a_1, b_2, \ldots, b_n$ értékek szolgáltatják a minimumot adó regressziós becslőfüggvényt, *a kapott*

$$\hat{x}_1 = a_1 + b_2 x_2 + \ldots + b_n x_n$$

regressziós függvény segítségével becslést készíthetünk X_1 feltételes várható értékére, adott $X_2 = x_2, \ldots, X_n = x_n$ értékek mellett (extrapolálás, interpolálás).

Igazolható, hogy *a mintából kapott becslés minden regressziós együtthatóra torzítatlan, és az összes lineáris torzítatlan becslések közül a legkisebb szórásnégyzetű.*

A b_i regressziós együtthatók jelentése: az

$$\hat{X}_1 = a_1 + b_2X_2 + \ldots + b_iX_i + \ldots + b_nX_n$$

egyenletben $\hat{X}_1$-et X_2 szerint parciálisan deriválva (tehát a többi változót fixen tartva):

$$\frac{\partial \hat{X}_1}{\partial X_2} = b_2,$$

általában

$$\frac{\partial \hat{X}_1}{\partial X_i} = b_i.$$

Ez a többi változó adott értéke esetén azt mutatja, hogy X_i-t változtatva mekkora az $\hat{X}_1$ változási sebessége. Másképpen; a többi változó adott értéke esetén X_i egységnyi növekedésére $\hat{X}_1$ átlagosan b_i-t változik.

13. Az

$$Y = a_0 + a_1X + a_2X^2 + \ldots + a_nX^n$$

polinomiális regresszió együtthatói kereshetők ugyanilyen normálegyenletrendszerrel, vagy az

$$X = Z_1, X^2 = Z_2, \ldots, X^n = Z_n$$

helyettesítéssel

$$Y = a_0 + a_1Z_1 + a_2Z_2 + \ldots + a_nZ_n$$

lineáris regresszióra visszavezetve. Ez esetben az a_i regressziós együtthatók megkereshetők számítógéppel, lineáris regressziós programcsomag segítségével, de a totális korrelációs együtthatóknak nem tulajdoníthatunk különösebb jelentőséget.

14. Egyetemi hallgatókat egyértelmű nagyság szerinti sorrendbe tudunk állítani, mért magasságaik alapján. A szépségverseny győzteseinél azonban a szépséget mérni nem lehet, mégis, az összes bíráló megad egy rangsort. *Mennyire függenek össze egymással, mennyire korrelálnak egymással két bíráló által adott rangsorok?* Ez is egyfajta véletlentől való függés, neve *rangkorreláció.*

A rangsorolásban való egyetértésnek többféle korrelációs együttható a mérőszáma, mi ezek közül a *Spearman-féle rangkorrelációs együtthatóval* foglalkozunk. Ennek értéke:

$$R = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n},$$

ahol n a rangsorolt egyedek száma; a rangsor a két bírálónál:

$$1.\ \text{bíráló rangsora: } \frac{1.\ 2.\ \ldots\ i\text{-edik}\ \ldots\ n\text{-edik egyed}}{x_1\ x_2\ \ldots\ x_i\ \ldots\ x_n},$$

$$2.\ \text{bíráló rangsora: } y_1\ y_2\ \ldots\ y_i\ \ldots\ y_n,$$

$\sum d_i^2 = \sum_{i=1}^{n} (x_i - y_i)^2$ a rangkülönbségek négyzetösszege.

Ha a két rangsorolás azonos, akkor $R = 1$. Ha a két rangsorolás éppen egymás ellentettje (fordítottja), akkor $R = -1$. Az $R = 0$ körüli érték rossz korrelációt jelez a két rangsor között.

Két konkrét bírálói rangsor tekinthető egy n-elemű mintának az egész sokaságból (akiket nem is rangsoroltak). *Hogyan vizsgáljuk meg a rangkorrelációs együttható szignifikanciáját?* Vizsgáljuk azt a nullhipotézist, hogy a kapott $R \neq 0$ érték ellenére a bírálók között nincs sem pozitív, sem negatív egyetértés, azaz

H_0: az egész sokaságban $R = 0$.

Ha $n \geq 10$, akkor a

$$t = R\sqrt{\frac{n-2}{1-R^2}}$$

próbastatisztika Student-eloszlású $n-2$ szabadságfokkal, és t-próbát végezhetünk H_0 elfogadására vagy elvetésére.

Néhány egyéb alkalmazás. Mennyire korrelál egymással a két bíráló véleménye, ha

– két bíráló n terméket rangsorba állít, különböző bírálati szempontok figyelembevételével;
– valakinek a súlyérzékét próbáljuk ki azáltal, hogy a valódi nagyság szerinti sorrendet összehasonlítjuk az illető rangsorolásával; vagy
– a keverés hatékonyságát ellenőrizhetjük, ha minden összekeverendő tárgyra sorszámot írunk fel. Megvizsgáljuk, keverés után milyen sorszámokat kapunk. Ha a keverés jó, akkor a rangkorrelációs együttható 0 körül jár. Ha nem jó, akkor 1 vagy -1 körül, és így tovább. Nagyon sok alkalmazása lehet a rangkorrelációnak.

Igen gyakran nemcsak két ítélet megegyezésével foglalkozunk. Közvéleménykutatást végzünk arról, hogyan ízlenek az újonnan gyártandó gyümölcslevek a nagyközönségnek. *Kettőnél több „bíráló" rangsora áll rendelkezésre. Azt szeretnénk tudni, a bírák együttesét tekintve van-e a közöttük levő egyetértésnek szignifikáns mértéke,* mennyire tekinthető homogénnek a bírálati rangsor, *milyen az egyetértés (konkordancia) a rangsorolók együttesében.* Ha m „bíráló" volt és n „tárgy", az egyetértés mértékét a *W konkordancia-együttható* jellemzi:

$$W=\frac{12S}{m^2(n^3-n)}.$$

Adjunk magyarázatot S kiszámítására! Ha egyáltalán nincs egyetértés a véleményükben, a várt rangösszeg:

$$Z=\frac{m(n+1)}{2}.$$

S az egy-egy tárgyban megfigyelt rangösszeg és a várt rangösszeg eltéréseinek a négyzetösszege.

Értékelés:

$$0\leq W\leq 1.$$

Minél közelebb van W az 1-hez, annál tökéletesebben egyezik a bírálók véleménye, minél közelebb van W a 0-hoz, annál kevésbé van egymással korrelációban a bírálók véleménye.

Az egész sokaságbeli konkordancia-együttható 0 voltát F-próbával tesztelhetjük, ekkor bizonyítható, hogy

$$F=\frac{(m-1)W_{\text{korr}}}{1-W_{\text{korr}}}$$

F-eloszlású változó, $n-1-\frac{2}{m}$ és $(m-1)\left(n-1-\frac{2}{m}\right)$ szabadságfokkal. A konkordancia-együttható korrigált értéke:

$$W_{\text{korr}}=\frac{S-1}{\dfrac{m^2(n^3-n)}{12}+2}.$$

A

$$H_0\colon W_{\text{elméleti}}=0$$

feltevés vizsgálatára F-próba végezhető.

A bírálati vélemények homogenitásának vizsgálatára bevezethető *az átlagrangsorral való rangkorreláció vizsgálata.* Ha az i-edik bíráló véleményének (rangsorának) az átlagrangsorral való rangkorrelációs együtthatója r_i, akkor

$$\bar{R}=\frac{\sum_{i=1}^{m}|r_i|}{m}$$

0-hoz, ill. 1-hez közel eső volta utal az átlagrangsorral való laza, ill. szoros korrelációra. A most mondott r_i rangkorrelációs együtthatók szórásának aránya abszolút értékük átlagához, $\bar{R}$-hoz mutatja, milyen erősen szóródnak az átlaguk körül az r_i értékek. Ha az így kapott relatív szórás magas százalékú, akkor a bírálói gárda véleményadás szempontjából nem eléggé homogén, ha alacsony százalékú, akkor az egyedi vélemények eléggé megegyeznek az átlagossal.

Különböző új technológiák eredményességét az eredeti (kontroll) technológiával így hasonlíthatjuk össze. Pl. n munkás teljesítményét rangsoroljuk az összes (kontroll + új) technológiák esetén (p esetben). A rangsorok:

Munkás	Teljesítményrangok a			
	kontroll-	2.	3.	p-edik technológiánál
1.	r_{11}	r_{12}	r_{13}	r_{1p}
2.	r_{21}	r_{22}	r_{23}	r_{2p}
n-edik	r_{n1}	r_{n2}	r_{n3}	r_{np}
Rang-összegek	s_1	s_2	s_3	s_p

Annak a feltevésnek a tesztelésére, hogy a technológiák nem hoznak szignifikáns javulást, elkészítjük a

$$\chi^2 = \frac{12}{np(p+1)} \sum_{i=1}^{p} S_p^2 - 3n(p+1)$$

próbastatisztikát, amely $p-1$ szabadságfokkal χ^2-eloszlású. A nullhipotézis elfogadásáról χ^2-próbával döntünk.

Gyakorló feladatok

Az alábbi két feladatban először a lineáris kétváltozós, ill. háromváltozós regresszióra mutatunk egy-egy olyan feladatot, amelyben *az egyik változó* (az év) *nem sztochasztikus, hanem determinisztikus változó.* Ilyenkor is szoktak regressziót számolni, korrelációról, korrelációs együtthatókról azonban igazában nem lehet beszélni.

A számítógép ilyenkor is számolja a korrelációs együtthatókat, úgy tekinti a szóban forgó determinisztikus változóra kapott mintát, mint ha sztochasztikus változóra kaptuk volna.

A formailag kiszámított korrelációs együttható mégis mutatja, hogy a kapott *idősor* pontjai mennyire szóródnak a regressziós egyenes körül ($r \approx 0$), vagy mennyire tapadnak hozzá ($r \approx \pm 1$).

Gyakorlásra igen alkalmas a most következő három feladat, amelynek kapcsán a számítástechnikai fogásokat a kezdő Olvasó is jól elsajátíthatja.

1. *A)* Az alábbi táblázat a termelt mennyiség változását mutatja az egyes években.

Év, x_i	Termelt mennyiség, y_i
1950	1
1955	2
1960	1,5
1965	2,5
1970	3

Az 1950. évi érték a bázisérték (a 100%), ehhez viszonyítva adjuk meg a többi évben termelt mennyiséget. A felírt y értékek tehát ún. bázisviszonyszámok.

A pontok látszólag egy egyenes körül csoportosulnak. A két változó között keressünk lineáris regressziót! Keressük a regressziós egyenes egyenletét, másképpen a ponthalmaz pontjait legjobban közelítő egyenes egyenletét!

A pontokat legjobban megközelítő regressziós egyenes egyenlete:

$$y = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} (x - \bar{x}) + \bar{y}.$$

100 darab $(x_i; y_i)$ mérési pont esetén már számítógépre kell bízni ezt a számítást. A termelt mennyiségre vonatkozó 5 pont esetén azonban elvégezzük a számítást (a nagyobb áttekintés érdekében táblázat segítségével). Figyeljük a képletet! Szükség lesz:

a) az x_i értékek eltérésére az átlaguktól ($x_i - \bar{x}$ feliratú oszlop); például $x_1 - \bar{x} = 1950–1960 = -10$ stb.;

b) az y_i értékeknek az átlaguktól való eltérésére ($y_i - \bar{y}$ oszlop), például $1–2 = -1$, $2–2 = 0$ stb.;

c) az *a)* és *b)* alatti megfelelő értékek szorzatára, vagyis az $(x_i - \bar{x})(y_i - \bar{y})$ vegyesszorzatokra, és az oszlop alján ezek összegére (pl. $(x_1 - \bar{x})(y_1 - \bar{y}) = (-10)(-1) = +10$ stb.);

d) az x_i értékek átlaguktól való eltérésének négyzetére (pl. $(x_1 - \bar{x})^2 = (-10)^2 = 100$ stb.), a rovat alján ezek összege a nevezőben álló $\Sigma(x_i - \bar{x})^2$;

e) itt nem, de a korrelációs együtthatónál szükség lesz az $(y_i - \bar{y})^2$ értékek összegére, ezért az utolsó oszlopban ezt is kiszámítjuk.

Év, x_i	Termelt mennyiség, y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1950	1	-10	-1	$(-10)(-1) = 10$	100	1
1955	2	-5	0	$(-5)\ 0 = 0$	25	0
1960	1,5	0	$-0,5$	0	0	0,25
1965	2,5	$+5$	$+0,5$	2,5	25	0,25
1970	3	$+10$	$+1$	10	100	1
Összesen: $\bar{x} = 1960$	$\bar{y} = 2$			$22,5 = \Sigma(x_i - \bar{x})(y_i - \bar{y})$	$250 = \Sigma(x_i - \bar{x})^2$	$2,5 = \Sigma(y_i - \bar{y})^2$

A pontokat legjobban megközelítő regressziós egyenes egyenlete:

$$y = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}(x - \bar{x}) + \bar{y} = \frac{22,5}{250}(x - 1960) + 2,$$

$$y = 0,09(x - 1960) + 2.$$

Az egyenes iránytangense (0,09) az egy év emelkedésre jutó átlagos y növekedést jelzi. Egy év alatt tehát várhatóan az 1950. évi bázis 9%-ával növekszik a termelés (hiszen ez az egyenlet becslés az egész sokaságra vonatkozó feltételes várható értékre, az elméleti regresszióra). Az iránytangens pozitív, ez a termelési tendencia átlagosan növekvő jellegét mutatja (az egyes években az előzőhöz viszonyítva felfelé vagy lefelé változott a termelt mennyiség, az egyenes egyenletének ismeretében a tendencia mégis növekvőnek mutatkozik).

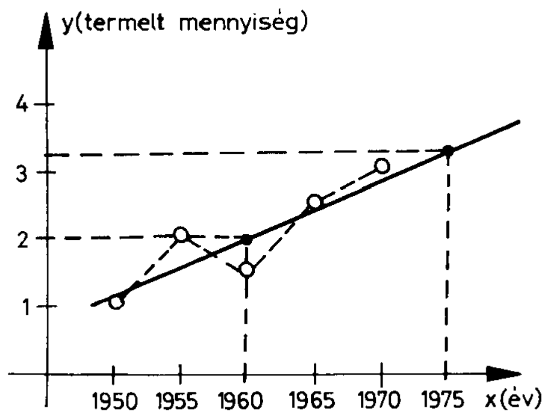
Ábrázoljuk az egyenest! Egyszerű például $x = 1960$-nál kiszámítani az y értéket:

$$\hat{y} = 0,09\underbrace{(x - 1960)}_{0} + 2 = 2,$$

$x = 1975$-ben pedig:

$$\hat{y} = 0,09\ (1975–1960) + 2 = 3,35.$$

A 48*a* ábrán az egyenest az (1960; 2), (1975; 3,35) pontok segítségével ábrázoltuk. Az 1970-es évben az 5 évvel előre számított 3,35 voltaképpen előrejelzés volt (empirikus becslés az elméleti várható értékre). Nem bizonyos, hogy a tényleges érték 1975-ben 3,35, csak e körül ingadozik (hiszen 3,35 csak egy várható érték, sőt annak is csak a becslése).



a)

48*a* ábra

Feladatunkban a korrelációs együttható:

$$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2 \cdot \Sigma(y_i-\bar{y})^2}} = \frac{22{,}5}{\sqrt{250 \cdot 2{,}5}} = 0{,}9.$$

(L. a táblázat alján szereplő szummákat.)

A kapott tapasztalati korrelációs együttható 1-hez közel esik, ez arra utal, hogy a két változó szoros összefüggésben van egymással, a pontok jól tapadnak a regressziós egyenesre. Ilyenkor a regressziós becslésben, előrejelzésben megbízhatunk. Az együttható előjele pozitív; ez is igazolja azt, hogy a bázisviszonyszámok a hullámzás ellenére átlagosan növekednek.

Torzítatlan becslés az egész sokaság korrelációs együtthatójára:

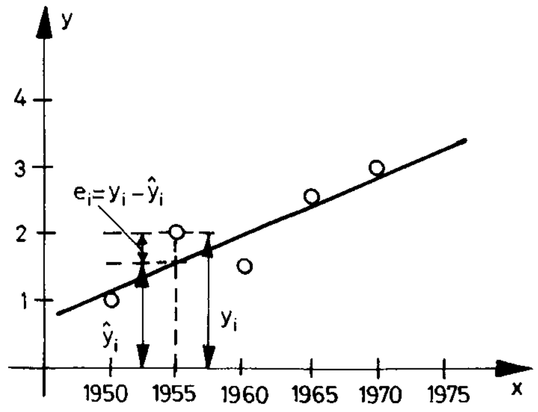
$$r^* = r\left(1+\frac{1-r^2}{2(n-4)}\right) = 1{,}095r = 0{,}9855.$$

B) Szórásnégyzetek; determinációs együttható. Az 1950-es évben a regressziós becslés (az egyenlettel számított ordináta):

$$\hat{y}_1 = 0{,}09(1950–1960) = +1{,}1,$$

ez természetesen nem azonos a tényleges $y_1 = 1$-gyel (hiszen az $\hat{y}$ egy várható értéknek a becslése csupán). Az eltérés $e_1 = y_1-\hat{y}_1 = -0{,}1$.

Ha általában $e_i = y_i-\hat{y}_i$, akkor (l. a 48*b* ábrát):



b)

48*b* ábra

y_i =	$\hat{y}_i$ +	e_i
Megfigyelt, valódi, mért ordináta	Regresszióval becsült, számított ordináta	A regresszióval meg nem magyarázott, véletlen eltérés (egyéb hatás okozta eltérés)

Vezessük be a következő szórásnégyzeteket:

$$s^{*2}(y_i) = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1},$$ a mért értékek átlagtól való eltérését méri,

$$s^{*2}(\hat{y}_i) = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{n-1},$$ a számított értékek átlagtól való eltérését méri,

$$s^{*2}(e_i) = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-1},$$ a mért és számított értékek eltérését (az ún. maradékokat, reziduumokat) jellemzi.

Könnyen bizonyítható a fenti szórásnégyzetek között fennálló következő összefüggés:

$s^{*2}(y_i)$ =	$s^{*2}(\hat{y}_i)$ +	$s^{*2}(e_i)$
y szórásnégyzete: teljes szórásnégyzet	*y* szórásnégyzetének a regresszió által megmagyarázott része (*x* változásával megmagyarázható része); külső szórásnégyzet	Az *y* változását jellemző szórásnégyzetnek az *x* hatás „kikapcsolása" után fennmaradó része; belső szórásnégyzet

(A szórásnégyzetekre – és természetesen nemcsak a felírt tapasztalati, hanem az elméleti szórásnégyzetekre is – az $y_i = \hat{y}_i + e_i$ összefüggésnek megfelelő összefüggés áll fenn.)

Térjünk vissza a feladatra!

Év, x_i	Mért ordináta, y_i	Számított ordináta, $\hat{y}_i$	$y_i-\hat{y}_i=e_i$	$(y_i-\hat{y}_i)^2=e_i^2$	$\hat{y}_i-\bar{y}$	$(\hat{y}_i-\bar{y})^2$
1950	1	1,1	−0,1	0,0100	−0,90	0,8100
1955	2	1,55	+0,45	0,2025	−0,45	0,2025
1960	1,5	2	−0,5	0,2500	0,00	0,0000
1965	2,5	2,45	+0,05	0,0025	+0,45	0,2025
1970	3	2,9	+0,1	0,0100	+0,90	0,8100
Összesen: $\bar{x}=1960$	$\bar{y}=2$			0,4750		2,0250

$$s^{*2}(y_i)=\frac{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}{n-1}=\frac{2,5}{4}=0,625,$$

a mért értékek szórásnégyzete 0,625. Az

$$s^{*2}(\hat{y}_i)=\frac{2,0250}{4}=0,50625,$$

a számított értékek szórásnégyzete tehát 0,506 25, ez a regresszió által megmagyarázott rész. Az y_i értékek változásának meg nem magyarázott (egyéb hatás, más független változó vagy a véletlen okozta) része:

$$s^{*2}(e_i)=\frac{0,4750}{4}=0,118\,75.$$

Valóban:

$$s^{*2}(y_i)=s^{*2}(\hat{y}_i)+s^{*2}(e_i),$$

$$0,625=0,506\,25+0,118\,75.$$

(Elég lett volna tehát a három szórásnégyzet közül csak kettőt kiszámítani.)

Hányad része a regresszióval számított értékek szórásnégyzete a mintából kapott értékek szórásnégyzetének?

$$\frac{s^{*2}(\hat{y})}{s^{*2}(y)}=\frac{0,506\,25}{0,625}=0,81.$$

Az y értékek szórásnégyzetét 81%-ban megmagyarázza a regresszió, 19%-ban nem, 19% a véletlen vagy egyéb változó okozta hatás. Érdekes megfigyelni, hogy a kapott ún. determinációs együttható a korrelációs együttható négyzete:

$$D=\frac{s^{*2}(\hat{y})}{s^{*2}(y)}=r^2=0,9^2.$$

(Az összefüggés igazolását az Olvasóra bízzuk.)

A determinációs együttható azt mutatja, hogy az y_i értékek szórásnégyzetének hányad részét magyarázhatjuk a feltételezett korrelációval (regresszióval).

A kapott összefüggés segítségével igen könnyű igazolni, hogy $-1\leqq r\leqq 1$. Ugyanis:

$$r^2=\frac{s^{*2}(\hat{y})}{s^{*2}(y)}=\frac{s^{*2}(y)-s^{*2}(e)}{s^{*2}(y)}=1-\frac{s^{*2}(e)}{s^{*2}(y)},$$

mivel 1-ből egy nemnegatív számot kell levonni, ezért $0\leqq r^2\leqq 1$, amiből $-1\leqq r\leqq 1$ következik.

Az r^2 annál közelebb esik az 1-hez, minél közelebb van a maradékok szórásnégyzete, $s^{*2}(e)$ a 0-hoz, vagyis: minél szorosabb az x és y kapcsolata. (Ha a maradékok szórásnégyzete 0, akkor a kapcsolat jellege már függvény, a pontok mind az egyenesen sorakoznak.)

Az r^2 annál közelebb esik 0-hoz, minél közelebb van a maradékok szórásnégyzete a mért értékek szórásnégyzetéhez, vagyis az y_i értékek szórásnégyzetét a regresszió nem magyarázza meg. Tehát a két változó vagy független, vagy más kapcsolat áll fenn köztük.

A regressziós becslés jellemzői a mért és számított értékek eltérései (a maradékok, reziduumok, az $e_i=y_i-\hat{y}_i$ értékek). Az ezekből az értékekből alkotott szórásnégyzet és az y_i értékek szórásnégyzete között a következő összefüggés áll fenn:

$$s^{*2}(e)=s^{*2}(y)\,(1-r^2).$$

A feladatban szereplő adatok esetén:

Év, x_i	Mért ordináta, y_i	Számított ordináta, $\hat{y}_i$	$e_i=y_i-\hat{y}_i$	$e_i^2=(y_i-\hat{y}_i)^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$
1950	1	1,1	−0,1	0,0100	−1	1
1955	2	1,55	+0,45	0,2025	0	0
1960	1,5	2	−0,5	0,2500	−0,5	0,25
1965	2,5	2,45	+0,05	0,0025	+0,5	0,25
1970	3	2,9	+0,1	0,0100	+1	1
Összesen				0,4750		2,5

$$s^{*2}(y) = \frac{2{,}5}{4} = 0{,}625, \quad s^{*2}(e) = \frac{0{,}4750}{4} = 0{,}118\,75,$$

$$s^{*2}(e) = 0{,}625(1-0{,}9^2) = 0{,}118\,75.$$

Megjegyzés

Hangsúlyozzuk, hogy az előzőekben a tapasztalati értékekre felírt összefüggések a megfelelő elméleti értékek között is fennállnak!

C) Számítógépes programcsomag használata. 1. A következőkben a feladat számítógépes megoldását mutatjuk be.

A STEP-WISE MULTIPLE REGRESSION (lépésenkénti többváltozós regresszió) programcsomag használatával az IBM 370/145 gépen lefuttatott eredményeket mutatjuk be az *a),–b),–c)* pontokban. Csak a legfontosabb eredményeket mutatjuk.

```
STEP-WISE MULTIPLE REGRESSION ........... LINREG
NUMBER OF OBSERVATIONS      5
NUMBER OF VARIABLES         2
NUMBER OF SELECTIONS        2
CONSTANT TO LIMIT VARIABLES 0.0
VARIABLE   MEAN       STANDARD
  NO.                 DEVIATION
   1       2.00000    0.79057
   2    1960.00000    7.90569
```


a)

```
CORRELATION MATRIX
ROW  1
  1.00000  0.90000
ROW  2
  0.90000  1.00000
```

b)

```
SELECTION...........1
DEPENDENT VARIABLE .......................... 1
NUMBER OF VARIABLES FORCED ................. 1
NUMBER OF VARIABLES DELETED ................ 0
```

c)

```
STEP 1
VARIABLE ENTERED ........... 2
  (FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP ..........     2.025
PROPORTION REDUCED IN THIS STEP ..............     0.810
CUMULATIVE SUM OF SQUARES REDUCED.............     2.025
CUMULATIVE PROPORTION REDUCED ................     0.810
                                              OF 2.500
FOR 1 VARIABLES ENTERED ......................
  MULTIPLE CORRELATION COEFFICIENT............     0.900
    (ADJUSTED FOR D.F.) ......................     0.900
F-VALUE FOR ANALYSIS OF VARIANCE..............    12.789
STANDARD ERROR OF ESTIMATE....................     0.398
    (ADJUSTED FOR D.F.) ......................     0.398

  VARIABLE     REGRESSION      STD.ERROR OF     COMPUTED
   NUMBER      COEFFICIENT     REG.COEFF.       T.VALUE
     2          0.09000         0.02517          3.576
INTERCEPT –    174.39995
```

Magyarázat

a)-hoz:

NUMBER OF OBSERVATIONS = a megfigyelések száma = 5,
VARIABLE 1 = az y változó,
MEAN = az y változó átlaga = 2,
STANDARD DEVIATION = az y változó szórása = Ø,79.
VARIABLE 2 = az x változó,
az x változó átlaga = 196Ø,
az x változó szórása = 7,9.

A korrelációs mátrix egyes elemei az y és x, x és y, y és y, x és x közötti kapcsolat korrelációs együtthatóit adják meg, ilyen sorrendben:

Korrelációs együtthatók	y	x
y	1	0,9
x	0,9	1

Az x és x, y és y közötti kapcsolat korrelációs együtthatója természetesen 1, az y és x közötti, valamint az x és y közötti kapcsolat korrelációs együtthatója is természetesen egyenlő, 0,9. (Jó a feltételezett lineáris regresszió!)

b)-hez: Ez a részlet azt mutatja, hogy az első választáskor (SELECTION) az 1. változó (az y) mellé egy újabb változót vesz be a program (NUMBER OF VARIABLES FORCED...1):

c)-hez: A bevont változó (VARIABLE ENTERED...2) az x.

A számított értékek átlagtól való eltérésének, az $(y_i-\bar{y})^2$ értékeknek négyzetösszege: 2,0250 (SUM OF SQUARES REDUCED IN THIS STEP), ez a mintából kapott $(y_i-\bar{y})^2$ értékek összegének a 81%-a (PROPORTION REDUCED IN THIS STEP...81). (Alatta az OF utáni rész mutatja, hogy ez a 2,0250 a 2,5-nek a 81%-a.) Ez a 0,81 egyúttal a külső és a teljes szórásnégyzet aránya is, vagyis a determinációs együttható.

Alulról olvasható a regressziós együttható:

VARIABLE NUMBER	REGRESSION COEFFICIENT	
2	0,09	= az x változó együtthatója
INTERCEPT	$-174{,}39$	= az állandó,

vagyis a korrelációs egyenes egyenlete:

$$y = 0{,}09x - 174{,}39$$

(ezt kapnánk az $y = 0{,}09(x-1960)+2$ egyenlet rendezéséből).

Észrevehető, hogy még egyéb értékeket is kiszámít a gépi program. Nem magyarázunk el mindent, de a t érték (COMPUTED T-VALUE = 3,576) magyarázatát megadjuk.

A korrelációs együttható értékét a mintából becsültük. Valóban eltér-e ez a 0-tól, vagy csak egy rossz mintavétel miatt véletlenül kaptunk 0,9-et?

Ha a két változó együttes eloszlása normális, akkor az $R=0$ hipotézis esetén a

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

változó $n-2$ szabadságfokú, Student-eloszlású (n a megfigyelések száma). Az $R=0$ feltétel fennállásának vizsgálatára tehát t-próbát végezhetünk. Ha

$$\left.\begin{array}{ll} |t_{\text{számított}}| \leqq t_{\text{táblázatbeli}}, & \text{elfogadjuk} \\ |t_{\text{számított}}| > t_{\text{táblázatbeli}}, & \text{elvetjük} \end{array}\right\} \text{az } R=0 \text{ feltevést.}$$

Jelen esetben a szabadságfokok száma: $f = n-2 = 3$, a táblázatbeli t érték:

$n = f+1$	Statisztikai biztonság	
	95%	99%
4	3,182	5,841

Mivel $3{,}182 = t_{95} < 3{,}576 = |t_{\text{számított}}| < t_{99} = 5{,}841$, az $R=0$ feltevés el is fogadható, el is vethető; lehet, hogy csak véletlenül (rossz mintavétel alapján) kaptuk az $r=0{,}9$-et, lehet, hogy nem. Kevés a mérésszám, a végleges döntéshez több mérésre lenne szükség.

2. Egy termékre vonatkozóan a termelt mennyiség változását jellemző bázisviszonyszámokat mutatja az alábbi táblázat az évek és a kutatásra–fejlesztésre fordított összeg függvényében.

Termelt mennyiség, y	Év, x_1	Kutatásra–fejlesztésre fordított összeg, millió Ft, x_2
1	1950	2
2	1955	2,5
1,5	1960	1,8
2,5	1965	2,2
3	1970	3,5
3,5	1975	4,0

Az y változót itt két független változó lineáris függvényeként akarjuk meghatározni,

$$y = b_0 + b_1 x_1 + b_2 x_2$$

alakban. Olyan b_0, b_1 b_2 regressziós együtthatókat kell keresni, hogy a fenti egyenlettel megadott regressziós sík a megadott pontokat a „legjobban" megközelítse. A legjobban közelítő sík itt is a legkisebb négyzetek elve alapján határozandó meg (a mért és számított y értékek eltérésének négyzetösszege legyen minimális).

Igazolható, hogy a regressziós függvénnyel kapott y érték egyúttal az x_1, x_2 adott értékénél szóba jöhető, adott eloszlású y értékek feltételes várható értéke (regressziója), pontosabban az adatokból kapott függvény ennek az elméleti regressziós függvénynek a becslése, így a számított y érték is egy becslés lesz az y feltételes várható értékére.

Nézzük a kitűzött feladat listáját!

```
STEP-WISE MULTIPLE REGRESSION ......... LINREG
NUMBER OF OBSERVATIONS           6
NUMBER OF VARIABLES              3
NUMBER OF SELECTIONS             3
CONSTANT TO LIMIT VARIABLES      Ø.Ø

  VARIABLE      MEAN       STANDARD
     NO.                   DEVIATION
     1.          2,25ØØØ    Ø.93541
     2.       1962.5ØØØØ    9.35414
     3.          2.66667    Ø.88468
```



```
CORRELATION MATRIX
ROW  1
  1.ØØØØØ    Ø.94286    Ø.89421
ROW  2
  Ø.94286    1.ØØØØØ    Ø.8Ø962
ROW  3
  Ø.89421    Ø.8Ø962    1.ØØØØØ

SELECTION ............ 1
DEPENDENT VARIABLE ........................... 1
NUMBER OF VARIABLES FORCED................. 2
NUMBER OF VARIABLES DELETED ................ Ø
STEP 1
VARIABLE ENTERED ........... 2
  (FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP ..........        3.889
PROPORTION REDUCED IN THIS STEP ...............       Ø.889
CUMULATIVE SUM OF SQUARES REDUCED..........           3.889
CUMULATIVE PROPORTION REDUCED ..............          Ø.889
                                                   OF 4.375
FOR 1 VARIABLES ENTERED
  MULTIPLE CORRELATION COEFFICIENT...........         Ø.943
    (ADJUSTED FOR D. F.)..........................    Ø.943
  F-VALUE FOR ANALYSIS OF VARIANCE...........        32.Ø29
  STANDARD ERROR OF ESTIMATE.................         Ø.348
    (ADJUSTED FOR D. F.)..........................    Ø.348

  VARIABLE    REGRESSION    STD. ERROR OF    COMPUTED
   NUMBER     COEFFICIENT    REG. COEFF.      T-VALUE
     2          Ø.Ø9429        Ø.Ø1666         5.659
  INTERCEPT  -182.78561

STEP 2
VARIABLE ENTERED ........... 3
  [FORCED VARIABLE]
SUM OF SQUARES REDUCED IN THIS STEP ..........        Ø.217
PROPORTION REDUCED IN THIS STEP ...............       Ø.Ø5Ø
CUMULATIVE SUM OF SQUARES REDUCED..........           4.1Ø7
CUMULATIVE PROPORTION REDUCED ..............          Ø.939
                                                   OF 4.375
```

FOR 2 VARIABLES ENTERED

MULTIPLE CORRELATION COEFFICIENT...........	Ø.969
[ADJUSTED FOR D. F.]...........................	Ø.961
F-VALUE FOR ANALYSIS OF VARIANCE............	22.962
STANDARD ERROR OF ESTIMATE.................	Ø.299
[ADJUSTED FOR D. F.]...........................	Ø.334

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
2	Ø.Ø6353	Ø.Ø2436	2.6Ø8
3	Ø.4Ø16Ø	Ø.25755	1.559
INTERCEPT	− 123.5Ø784		

Rövid értékelés:

a) A STEP 1 (1. lépés) mutatja, hogy a 2. számú változót (az x_1 évet) vonta be először a program, a SUM OF SQUARES REDUCED és a PROPORTION REDUCED IN THIS STEP itt a legnagyobb. A kapott regressziós összefüggés:

$$y = 0{,}09429x_1 - 182{,}78561,$$

egy év növekedésre átlagban 9,4% termelésnövekedés jut, ha csak y és x_1 értékei alapján vizsgáljuk a korrelációt. A MULTIPLE CORRELATION COEFFICIENT 0,943; ez jó korrelációra mutat.

A második lépésben a program bevonta x_2-t is, a regresszió becslése:

$$y = \underset{\substack{\uparrow \\ \text{év}}}{0{,}06353x_1} + \underset{\substack{\uparrow \\ \text{kutatásra–fejlesztésre} \\ \text{fordított összeg}}}{0{,}40160x_2} - 123{,}50784,$$

évente átlagosan 6,4% a termelésnövekedés, 1 millió Ft fejlesztési összegre 40% termelésnövekedés jut.

b) A páronkénti totális korrelációs együtthatók az y és x_1, y és x_2 közvetett kapcsolatát mutatják (a szimmetria miatt elég csak a mátrix alsó felét feltüntetni):

$$\begin{array}{c} \\ y \\ x_1 \\ x_2 \end{array} \begin{array}{c} \begin{array}{ccc} y & x_1 & x_2 \end{array} \\ \begin{bmatrix} 1 & & \\ 0{,}94286 & 1 & \\ 0{,}89421 & 0{,}80962 & 1 \end{bmatrix} \end{array}.$$

Eszerint a termelés az évekkel szoros (pozitív) korrelációban van, $r_{y1} = 0{,}94$, míg valamivel kevésbé, de még eléggé függ a kutatásra–fejlesztésre fordított összegtől, $r_{y2} = 0{,}89$. A kutatásra fordított összeg is függ az évek számától, bár nem annyira, mint a termelés, $r_{12} = 0{,}81$.

c) Az x_2 hatásának kiszűrése után az y és x_1 közvetlen kapcsolatát jellemző parciális korrelációs együttható:

$$r_{y1\cdot 2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}} = \frac{0{,}94 - 0{,}89 \cdot 0{,}81}{\sqrt{(1-0{,}89^2)(1-0{,}81^2)}} = 0{,}819,$$

ez valamivel kisebb, mint a totális korrelációt jelző $r_{y1} = 0{,}94$, de x_2 kiszűrése után is szoros a kapcsolat a termelés és az évek között.

$$r_{y2\cdot 1} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{(1-r_{y1}^2)(1-r_{12}^2)}} = \frac{0{,}89 - 0{,}94 \cdot 0{,}81}{\sqrt{(1-0{,}94^2)(1-0{,}81^2)}} = 0{,}643,$$

tehát az évek hatásának kiszűrése után lényegesen rosszabb korreláció mutatkozik a termelés és a kutatásra–fejlesztésre fordított összeg között. (Nagyobb mértékű a fejlődés az eltelt időben szerzett tapasztalatok révén.) A gyengülést magyarázza, hogy az évek és a kutatási összeg között elég jó korreláció mutatkozott a totális korrelációs együtthatónál is.

d) A többszörös korrelációs együttható $R = 0{,}969$, tehát a második lépésben kapott

$$y = 0{,}063\,53x_1 + 0{,}401\,60x_2 - 123{,}507\,84$$

regressziós összefüggésben összességében jól megbízhatunk, a feltételezett korreláció jó, a pontok jól „tapadnak" a síkra. A determinációs együttható, $R^2 = 0{,}938\,961$, az y szórásnégyzetének majdnem 94%-át megmagyarázza a regresszió (kb. 6% a véletlen, egyéb hatás következménye).

3. Forgácsolásnál a fő forgácsolóerő (Y) és a forgácsolószélesség (X) kapcsolatának megállapítására adottak az alábbi összetartozó pontpárok:

Forgácsoló szélesség, x_i, mm	Fő forgácsolóerő, y_i, kN
1	1
2	2
4	3
5	6

Állapítsuk meg a regressziós egyenes egyenletét és a korrelációs együtthatót mindkét képletcsoport segítségével!

I.	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
	1	1	−2	−2	4	4	4
	2	2	−1	−1	1	1	1
	4	3	1	0	0	1	0
	5	6	2	3	6	4	9
Összesen:	12 $\bar{x}=3$	12 $\bar{y}=3$			11	10	14

$$y = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}(x-\bar{x})+\bar{y} = \frac{11}{10}(x-3)+3,$$

$$r = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}} = \frac{11}{\sqrt{10\cdot 14}} = 0,93,$$

II.	x_i	y_i	x_iy_i	x_i^2	y_i^2
	1	1	1	1	1
	2	2	4	4	4
	4	3	12	16	9
	5	6	30	25	36
Összesen:	12 $\bar{x}=3$	12 $\bar{y}=3$	47	46	50

$$y = \frac{\sum x_iy_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}(x-\bar{x})+\bar{y} = \frac{47-4\cdot 3\cdot 3}{46-4\cdot 9}(x-3)+3 = 1{,}1(x-3)+3;$$

$$r = \frac{\sum x_iy_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i - n\bar{y}^2)}} = \frac{47-4\cdot 3\cdot 3}{\sqrt{(46-4\cdot 9)(50-4\cdot 9)}} = 0{,}93.$$

4. Vizsgáljuk az őszi csapadék és a talajvízállás közötti kapcsolatot! Tegyük fel, hogy a kapcsolat

$$Y = a+bX$$

alakú! Az $n=18$ évi mérésből származó (x_i, y_i) adatpár és a szükséges x_iy_i, x_i^2, y_i^2 értékek összegét az alábbi táblázat tartalmazza.

Talajvízszint, y_i, mm	Csapadék, x_i, cm	x_iy_i	x_i^2	y_i^2
1,25	10,36			
1,40	8,94			
2,13	13,21			
1,19	15,80			
1,65	11,18			
1,89	13,64			
1,68	19,53			
1,77	24,56			
1,28	11,48			
1,16	7,77			
0,94	11,30			
3,69	28,13			
3,51	30,18			
3,14	23,14			
1,22	15,88			
2,29	19,76			
4,42	35,36			
2,90	25,40			
Összeg: 37,51 Átlag: 2,1	325,61 18,1	809,17	7002	96,1

A b iránytangens mintabeli becslése:

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{809{,}17 - 18 \cdot 2{,}1 \cdot 18{,}1}{7002 - 18 \cdot 18{,}1^2} = 0{,}113.$$

Az ordinátatengely-metszet becslése:

$$a = \bar{y} - b\bar{x} = 2{,}1 - 0{,}113 \cdot 18{,}1 = 0{,}040.$$

Így a regressziós egyenes egyenlete:

$$y = a + bx = 0{,}040 + 0{,}113x.$$

Az X változó szórásnégyzetének becslése:

$$s^*(x)^2 = \frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n-1} = \frac{7002 - \frac{(325{,}61)^2}{18}}{17} = 65{,}409.$$

Az Y változó szórásnégyzetének (a teljes szórásnégyzetnek) becslése:

$$s^*(y)^2 = \frac{\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}}{n-1} = \frac{96{,}1 - \frac{(37{,}51)^2}{18}}{17} = 1{,}053.$$

A korrelációs együttható becslése:

$$r = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)(\Sigma y_i^2 - n\bar{y}^2)}} = b\frac{s^*(x)}{s^*(y)} = 0{,}113 \cdot \frac{8{,}09}{1{,}026} = 0{,}89,$$

ez jó (pozitív) korrelációra utal a két változó között. Ha a két változó együttes eloszlása kétdimenziós normál eloszlás, akkor az $R=0$ hipotézis (a két változó korrelálatlansága a teljes sokaságban) vizsgálatára a 15. táblázatból látható, hogy

$$r_{\text{krit}} = 0{,}468,$$

$f=n-2$	95%
16	0,468

mivel

$$|r| = 0{,}89 > r_{\text{krit}} = 0{,}468,$$

az $R=0$ feltevést elvetjük, nem tekintjük korrelálatlannak a két változót az egész sokaságban sem.

Ha a két változó együttes eloszlása nem normális eloszlás, akkor a

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

változó elég nagy n-re közel normál eloszlású, várható értéke:

$$M(Z) \approx \frac{1}{2}\ln\frac{1+R}{1-R} + \frac{R}{2(n-1)},$$

szórásnégyzete:

$$D^2(Z) \approx \frac{1}{n-3}.$$

Tegyük fel, hogy $R=0$. Akkor az

$$M(Z) \approx \frac{1}{2}\ln 1 + 0 = 0, \qquad D^2(Z) \approx \frac{1}{n-3} = \frac{1}{15} = 0{,}067,$$

0 várható értékű, $\sqrt{0{,}067} = 0{,}258$ szórású, normál eloszlású változó értékei 95% valószínűséggel

$$0 - 1{,}96 \cdot 0{,}258 = -0{,}506 \quad \text{és} \quad 0 + 1{,}96 \cdot 0{,}258 = 0{,}506$$

közé esnek. Mivel

$$Z = \frac{1}{2}\ln\frac{1+0{,}89}{1-0{,}89} = 1{,}422$$

bőven az előbbi konfidencia-intervallumon kívül esik, az $R=0$ feltevés elvethető.

A maradékok szórásnégyzetének (a belső szórásnégyzetnek) a becslése:

$$s^*(e)^2 = s^*(y)^2(1-r^2) = 1{,}053(1-0{,}89^2) = 0{,}219,$$

ez $D^2(E)$-re torzított becslés. Torzítatlan becslést akkor kapunk, ha $n-1$ helyett $(n-2)$-vel osztunk, vagyis a torzítatlan becslés a most kapott érték $\frac{n-1}{n-2}$-szerese:

$$s^*(e)^2 = \frac{17}{16} \cdot 0{,}219 = 0{,}232.$$

Mivel a mintaelemszám nem túl nagy, dolgozzunk ezzel a becsléssel!

A szórásnégyzet regresszió által megmagyarázott részének (a külső szórásnégyzetnek) a becslése:

$$s^*(\hat{y})^2 = s^*(y)^2 r^2 = 1{,}053 \cdot 0{,}89 = 0{,}937.$$

Az

$$y = a + bx$$

egyenletben az iránytangens szórásnégyzetének becslése:

$$s^*(b)^2 = \frac{s^*(e)^2}{ns^*(x)^2} = \frac{0{,}232}{18 \cdot 65{,}409} = 0{,}000\,197,$$
$$s^*(b) = 0{,}014.$$

Az iránytangens szórása meglehetősen kicsi.

Az elméleti β iránytangest $1-\varepsilon$ valószínűséggel tartalmazza a

$$b - t_\varepsilon s^*(b) \leqq \beta \leqq b + t_\varepsilon s^*(b)$$

konfidencia-intervallum. Itt t_ε a Student-eloszlás táblázatából vehető, $n-2 = 16$ szabadságfok esetén:

$n=f+1$	99%
17	2,921

A 99%-os szintű konfidencia-intervallum:

$$0{,}113 - 2{,}921 \cdot 0{,}014 \leqq \beta \leqq 0{,}113 + 2{,}921 \cdot 0{,}014$$
$$0{,}072 \leqq \beta \leqq 0{,}154.$$

Ha tehát azt tesszük fel, hogy a mintabeli $b = 0{,}113$ ellenére az elméleti $\beta = 0$ (a regressziós egyenes párhuzamos az y tengellyel), akkor ezt a feltevést 99%-os biztonsági szinten elvethetnénk, b szignifikánsan (statisztikailag jelentősen) eltér a 0-tól.

Az y tengelymetszet szórásnégyzetének becslése:

$$s^*(a)^2 = \frac{s^*(e)^2 \Sigma x_i^2}{n^2 s^*(x)^2} = \frac{0{,}232 \cdot 7002}{18^2 \cdot 65{,}409} = 0{,}077,$$
$$s^*(a) = 0{,}277.$$

Ez sem túl nagy. Ennek segítségével az alábbi, $1-\varepsilon$-szintű megbízhatósági intervallum kapható az elméleti y tengelymetszetre, α-ra:

$$a - t_\varepsilon s^*(a) \leqq \alpha \leqq a + t_\varepsilon s^*(a),$$

itt t_ε a Student-eloszlás $n-2 = 16$ szabadságfokhoz, $(1-\varepsilon)\cdot 100\%$ biztonsági szinthez tartozó kritikus értéke. 99% biztonsági szinten az előző $t_\varepsilon = 2{,}921$ szorzóval dolgozva:

$$0{,}040 - 2{,}921 \cdot 0{,}277 \leqq \alpha \leqq 0{,}040 + 2{,}921 \cdot 0{,}277,$$
$$-0{,}769 \leqq \alpha \leqq 0{,}849.$$

Azt a feltevést, hogy az elméleti

$$Y = \alpha + \beta X$$

egyenes átmegy az origón (Y tengelymetszete 0), itt most nem vetnénk el, hiszen $\alpha = 0$ a kapott konfidencia-intervallumra esik.

Az adott x_i-hez tartozó Y_i egyedi értéket tartalmazza az alábbi, $\hat{y}_i$-re szimmetrikus konfidencia-intervallum, mégpedig $1-\varepsilon$ valószínűséggel:

$$\hat{y}_i - t_\varepsilon s^*(e)\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{ns^*(x)^2}} \leqq Y_i \leqq \hat{y}_i +$$
$$+ t_\varepsilon s^*(e)\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{ns^*(x)^2}},$$

itt t_ε a Student-eloszlás táblázatában $(1-\varepsilon)\cdot 100\%$ biztonsági szinthez, $n-2$ szabadságfokhoz tartozó kritikus érték. Az így kapható konfidenciasáv a legkeskenyebb az $x_i = \bar{x} = 18{,}1$ értéknél (ahol $\hat{y}_i = \bar{y} = 2{,}1$):

$$2{,}1 - 2{,}921 \cdot \sqrt{0{,}232}\sqrt{1 + \frac{1}{18}} \leqq y \leqq 2{,}1 + 2{,}921 \cdot \sqrt{0{,}232}\sqrt{1 + \frac{1}{18}},$$
$$0{,}646 \leqq y \leqq 3{,}545.$$

Számítsuk ki az $x = 8$-hoz és az $x = 28{,}2$-hez tartozó konfidencia-intervallum határokat!

$x = 8$ esetén $\hat{y} = a + bx = 0{,}040 + 0{,}113 \cdot 8 = 0{,}944$,

a regresszióval becsült $\hat{y}$ értéktől való eltérés:

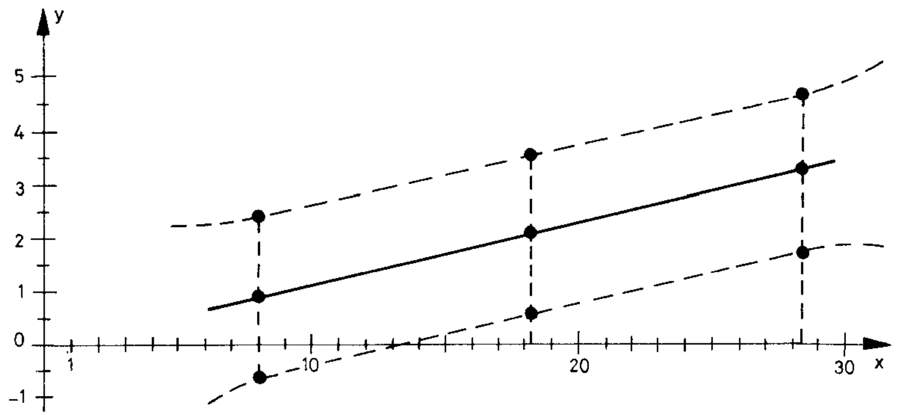
$$2{,}921 \cdot \sqrt{0{,}232}\sqrt{1 + \frac{1}{18} + \frac{(8 - 18{,}1)^2}{18 \cdot 65{,}409}} = 1{,}504,$$

így $x=8$-nál a konfidencia-intervallum:

$$-0{,}560 \leqq y \leqq 2{,}448.$$

$x=28{,}2$ esetén $\hat{y} = a+bx = 0{,}040+0{,}113 \cdot 28{,}2 = 3{,}227$, itt a konfidencia-intervallum (kihasználjuk, hogy $(8-18{,}1)^2 = (28{,}2-18{,}1)^2$ miatt a becsült $\hat{y}$ értéktől való eltérés szintén az előbbi 1,504):

$$1{,}723 \leqq y \leqq 4{,}731.$$



49. ábra

A 49. ábrán szemléltetett *konfidenciasávba* esnek az egyedi y értékek 99% biztonsággal. (A konfidenciasávot két hiperbola határolja, a feltüntetett tartományon alig görbülnek, hiszen a regresszióval becsült y-tól

$x=8$-nál és $x=28{,}2$-nél az eltérés 1,504,
$x=18{,}1$-nél az eltérés 1,445.)

5. $n=15$ megfigyelés eredményeit mutatja az alábbi táblázat. Keressünk lineáris regressziót a humusztartalom-mg%-ban (100 mg földben levő humusz-mg-ban) mért értékei (x_i) és a burgonyatermés (y_i) között (utóbbit a 0,6 ha-ra eső 100 kg-ban adtuk meg)!

Humusz, x_i, mg %	Termés, y_i, 100 kg/0,6 ha
523	27,3
594	62,8
517	52,0
593	49,6
696	61,5
845	71,8
617	51,0
780	79,6
634	53,9
671	53,6
732	63,1
724	71,9
739	78,9
732	72,2
682	63,0

$$y = 0{,}121x-20{,}49, \quad r=0{,}824.$$

6. A Kara-Kum sivatag földgázvagyonának a kitermeléséről az 1970–74-es években az alábbi adatok álltak rendelkezésre. Keressünk lineáris regressziót a kitermelt mennyiség és az év között! A regressziós becslés alapján mennyi fogy el 1970–71–72–73–74–75-ben? Ha a kitermelés növekedési üteme a jelenlegi marad, várhatóan mikor fogy el a 6000 millió m^3-re becsült földgázvagyon?

Év, x_i	1970	1971	1972	1973	1974
Kitermelt földgáz, y_i, millió m^3	13	24	32	46	57

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$\hat{y}_i$
1970	13	−2	−21,4	42,8	4	457,16	12,4
1971	24	−1	−10,4	10,4	1	108,16	23,4
1972	32	0	− 2,4	0	0	5,76	34,4
1973	46	+1	+11,6	11,6	1	134,56	45,4
1974	57	+2	+22,6	45,2	4	510,76	56,4
Összesen: $\bar{x}=1972$	$\bar{y}=34{,}4$			110,0	10	1217,20	172,0

A regressziós egyenes egyenlete:

$$y = 11(x-1972)+34{,}4,$$

az x együtthatója 11, az évi átlagos növekedés a kitermelésben 11 millió m^3. Így az 1975-ös becslés $56{,}4+11 = 67{,}4$, a 6 év alatt összesen $172+67{,}4 = 239{,}4$ millió m^3 kitermelése várható.

Az évi átlagos 11 millió m^3 növekedés alapján az évente várható kitermelt mennyiségek számtani sorozatot alkotnak, az első tag $a_1=12{,}4$, a differencia $d=11$, az n év alatt kitermelt mennyiség pedig $S_n=6000$,

$$S_n = \frac{(a_1+a_n)n}{2} = \frac{[12{,}4+12{,}4+(n-1)11]n}{2} = 6000,$$

$$11n^2+13{,}8n-12\,000 = 0,$$

a kapott másodfokú egyenletből csak a pozitív gyök jó, így várhatóan $n=32{,}4$ évre tervezhető a földgázvagyon kiaknázása. A korrelációs együttható $r=0{,}997$, igen szoros a két változó közti lineáris függés, megbízhatunk a tervezésben.

7. Az alábbi táblázatban a magyarországi szén- és villamosenergia-termelés előző évihez mért százalékos változásai (az ún. láncindexek) láthatók.

Év	Széntermelés-emelkedési százalékok	Villamosenergia-emelkedési százalékok
1950	9,4	19,0
1951	13,3	16,8
1952	14,5	19,9
1953	9,3	10,0
1954	1,4	4,4
1955	3,8	12,4
1956	− 8,4	− 4,2
1957	0,3	4,8
1958	12,3	18,9
1959	7,2	9,1

Tételezzünk fel mindegyik változópár között lineáris kapcsolatot!

a) Vizsgáljuk a villamosenergia- és a széntermelés láncindexeinek kapcsolatát!

b) Állapítsuk meg a villamosenergia-változások és az évek közötti kapcsolatot!

c) Írjuk fel a széntermelés-láncindexek és az évek között feltételezett lineáris kapcsolatot!

Igazából az egyik változó, az év, nem véletlentől függő változó, a gyakorlatban viszont annak tekintik. Ez azt jelenti, hogy ***a legkisebb négyzetek elvét alkalmazzák a ponthalmazt legjobban megközelítő egyenes keresésére, a korrelációs együtthatóval pedig jellemzik a pontoknak az egyenesre való tapadását.***

A feladatot számítógéppel megoldva:

a) az egyenlet a v villamosenergia-növekedés és az s széntermelés-növekedés között:

$$v = 1{,}036(s-6{,}310)+11{,}110,$$
$$r=0{,}934,$$

ennek 1-hez közeli volta azt mutatja, hogy a pontok jól tapadnak a feltételezett egyenesre, a korreláció erős a villamosenergia- és a széntermelés-növekedés között. A széntermelés-láncindex egységnyi növelésével a villamosenergia-láncindex átlagosan 1,036-del növekszik.

b) A feltételezett lineáris kapcsolat a villamos energia százalékos növekedése és az *e* év között:

$$v = -1{,}153(e-1954{,}5)+11{,}110,$$
$$r = -0{,}436,$$

tehát a feltételezett „korreláció" közepesnek mondható.

c) $s = -0{,}928(e-1954{,}5)+6{,}310,$
$r = -0{,}389.$

8. Az alábbi táblázattal kapcsolatos kérdéseink: van-e összefüggés a vizsgált két jelenség között? Ha van, milyen szoros ez az összefüggés? Milyen arányban befolyásolja a tudásszintet az osztályba járó értelmiségi származású gyermekek aránya?

Az iskolai osztályokba járó értelmiségi származású gyermekek aránya, x_i, %	Számolási készség minősége az osztályokban (átlag), y_i, %
$x_1 = 6$	$y_1 = 7$
$x_2 = 10$	$y_2 = 10$
$x_3 = 13$	$y_3 = 10$
$x_4 = 14$	$y_4 = 11$
$x_5 = 16$	$y_5 = 9$
$x_6 = 17$	$y_6 = 14$
$x_7 = 19$	$y_7 = 12$
$x_8 = 20$	$y_8 = 18$
$x_9 = 22$	$y_9 = 12$
$x_{10} = 23$	$y_{10} = 17$

$$\bar{x}=16,\ \bar{y}=12,\ r=\frac{131}{10\cdot 5{,}1\cdot 3{,}3}=0{,}78,$$

tehát van lineáris összefüggés az értelmiségi származású gyerekek aránya és a tanulmányi eredmény között. $r^2=0{,}61$, az osztályátlag alakulását az a tényező,

hogy az osztályba hány % értelmiségi gyerek jár, 61%-ban határozza meg, 39% változás más tényezők hatására jön létre.

9. Egy üzemben csöveket vágnak 100 cm-es darabokra, $n=6$ cső hosszát és súlyát lemérik.

Hossz, x_i, cm	101,3	103,7	98,6	99,9	97,2	100,1
Súly, y_i, század N	609	626	586	594	579	605

a) Határozzuk meg a korrelációs együtthatót! Ennek alapján döntsük el, feltételezhető-e a két változó közt lineáris korreláció! Mi lesz a regressziós egyenes egyenlete?

b) Az 1 cm-re eső átlagos csősúlyra a szabványelőírás $6{,}0\pm 0{,}02$ század N/cm. Határozzuk meg az *x* és *y* értékek átlagából, hogy várhatóan teljesül-e a szabványelőírás!

---

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
101,3	609	1,167	9,167			
103,7	626	3,567	26,167			
98,6	586	−1,533	−13,833			
99,9	594	−0,233	−5,833			
97,2	579	−2,933	−20,833			
100,1	605	−0,033	5,167			
600,8	3599	0	0	187,533	25,093	1454,833

a)
$r=0{,}9815$ szoros korrelációra, lineáris regresszióra utal. A regressziós egyenes egyenlete:

$$y=7{,}4735(x-100{,}133)+599{,}833,$$
$$y=7{,}4735x-148{,}51.$$

b) $\dfrac{\bar{y}}{\bar{x}}=\dfrac{599{,}833}{100{,}133}=5{,}9904$, ez a megengedett 5,98 és 6,02 közé esik. Várhatóan teljesül a szabványelőírás.

10. A vízgazdálkodásban fontos feladat a várható vízmagasság előrejelzése. A vízmagasság 5–6 paramétertől is függ, így a többváltozós regresszió számítása komoly számítógépes feladat.

A Duna budapesti tetőző vízállásának becslésére 26 év megfigyelései alapján készítettek regressziós modellt. A csapadék mennyisége és az esőzés kezdetén fennálló budapesti vízállás két fontos változó a vízállás megállapítására. Az adatok:

Tetőző vízállás, x_1, cm	Csapadék, x_2, mm	Budapesti vízállás az esőzés kezdetén, x_3, cm	Tetőző vízállás, x_1, cm	Csapadék, x_2, mm	Budapesti vízállás az esőzés kezdetén, x_3, cm
590	58	405	710	62	560
660	52	450	620	54	420
780	133	350	660	48	620
770	179	285	620	86	390
710	98	330	590	74	350
640	72	400	740	95	570
670	72	550	730	44	710
520	43	480	720	53	700
660	62	450	720	77	580
690	67	610	640	46	700
500	64	380	805	123	560
460	33	460	510	26	370
610	57	425	673	62	430

Készítsük el a vízállás becslésére szolgáló lineáris regressziós modellt!

Tekintsük a megfelelő listát!

```
STEP-WISE MULTIPLE REGRESSION ........... LINREG
NUMBER OF OBSERVATIONS                26
NUMBER OF VARIABLES                    3
NUMBER OF SELECTIONS                   3
CONSTANT TO LIMIT VARIABLES          Ø.Ø

VARIABLE      MEAN       STANDARD
  NO.                    DEVIATION
   1       653.769Ø4     88.35922
   2        7Ø.76923     33.2Ø88Ø
   3       482.11523    121.61542

CORRELATION MATRIX
ROW  1
  1.ØØØØØ     Ø.65645     Ø.27785
ROW  2
  Ø.65645     1.ØØØØØ    -Ø.37291
ROW  3
  Ø.27785    -Ø.37291     1.ØØØØØ

SELECTION ........... 1
DEPENDENT VARIABLE ........................... 1
NUMBER OF VARIABLES FORCED ................ 2
NUMBER OF VARIABLES DELETED ................ Ø
STEP  1
VARIABLE ENTERED ......................................... 2
  (FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP ..........   841Ø9.75Ø
PROPORTION REDUCED IN THIS STEP ...............      Ø.431
CUMULATIVE SUM OF SQUARES REDUCED..........   841Ø9.75Ø
CUMULATIVE PROPORTION REDUCED ..............      Ø.431
                                          OF 195184.ØØØ
FOR 1 VARIABLES ENTERED
  MULTIPLE CORRELATION COEFFICIENT...........      Ø.656
    (ADJUSTED FOR D.F.) ..........................      Ø.656
  F-VALUE FOR ANALYSIS OF VARIANCE............     18.174
  STANDARD ERROR OF ESTIMATE.................     68.Ø3Ø
    (ADJUSTED FOR D.F.) ..........................     68.Ø3Ø
```

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD.ERROR OF REG.COEFF.	COMPUTED T-VALUE
2	1.74663	0.40971	4.263
INTERCEPT	530.16162		

STEP 2
VARIABLE ENTERED 3
(FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP 61928.379
PROPORTION REDUCED IN THIS STEP 0.317
CUMULATIVE SUM OF SQUARES REDUCED.......... 146038.125
CUMULATIVE PROPORTION REDUCED 0.748
OF 195184.000

FOR 2 VARIABLES ENTERED
MULTIPLE CORRELATION COEFFICIENT........... 0.865
(ADJUSTED FOR D. F.) 0.859
F-VALUE FOR ANALYSIS OF VARIANCE 34.173
STANDARD ERROR OF ESTIMATE 46.225
(ADJUSTED FOR D. F.) 47.178

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
2	2.34897	0.30003	7.829
3	0.44106	0.08193	5.384
INTERCEPT	274.89063		

11. Egy vegyipari vállalatnál hetente feljegyezték az előállított termékmennyiséget és a munkaköltséget. Az adatok:

Idő, x_2, hét	Termék-mennyiség, y, 10 000 N/hét	Költség, x_1, Ft/hét	Idő, x_2, hét	Termék-mennyiség, y, 10 000 N/hét	Költség, x_1, Ft/hét
1	182	8980	14	207	8800
2	198	9170	15	220	9550
3	221	9380	16	211	9380
4	205	8850	17	194	8860
5	173	7680	18	186	9290
6	195	8960	19	203	9010
7	188	8890	20	191	8230
8	200	9520	21	183	8390
9	193	8810	22	207	9100
10	199	8880	23	204	9370
11	40	6500	24	201	8800
12	173	8700	25	188	8840
13	183	8530	26	195	8530

Elemezzük a termékmennyiséget mint a költség és az idő lineáris függvényét (keressünk lineáris regressziót)!

Tekintsük a számítógépes megoldást!

STEP-WISE MULTIPLE REGRESSION .... LINREG
NUMBER OF OBSERVATIONS 26
NUMBER OF VARIABLES 3
NUMBER OF SELECTIONS 3
CONSTANT TO LIMIT VARIABLES 0.0

VARIABLE NO.	MEAN	STANDARD DEVIATION
1	190.00000	32.97026
2	8807.69141	628.06519
3	13.50000	7.64853

CORRELATION MATRIX
ROW 1
1.00000 0.86170 0.08312
ROW 2
0.86170 1.00000 0.00691
ROW 3
0.08312 0.00691 1.00000

SELECTION 1
DEPENDENT VARIABLE 1
NUMBER OF VARIABLES FORCED 2
NUMBER OF VARIABLES DELETED 0
STEP 1
VARIABLE ENTERED 2
(FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP 20178.770
PROPORTION REDUCED IN THIS STEP 0.743
CUMULATIVE SUM OF SQUARES REDUCED 20178.770
CUMULATIVE PROPORTION REDUCED 0.743
OF 27175.965

FOR 1 VARIABLES ENTERED

MULTIPLE CORRELATION COEFFICIENT 0.862

(ADJUSTED FOR D. F.) 0.862

F-VALUE FOR ANALYSIS OF VARIANCE 69.212

STANDARD ERROR OF ESTIMATE 17.075

(ADJUSTED FOR D. F.) 17.075

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
2	0.04523	0.00544	8.319
INTERCEPT	−208.41406		

STEP 2

VARIABLE ENTERED 3

(FORCED VARIABLE)

SUM OF SQUARES REDUCED IN THIS STEP 161.808

PROPORTION REDUCED IN THIS STEP 0.006

CUMULATIVE SUM OF SQUARES REDUCED 20340.574

CUMULATIVE PROPORTION REDUCED 0.748

OF 27175.965

FOR 2 VARIABLES ENTERED

MULTIPLE CORRELATION COEFFICIENT 0.865

(ADJUSTED FOR D. F.) 0.859

F-VALUE FOR ANALYSIS OF VARIANCE 34.221

STANDARD ERROR OF ESTIMATE 17.239

(ADJUSTED FOR D. F.) 17.595

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
2	0.04521	0.00549	8.235
3	0.33263	0.45080	0.738
INCERCEPT	−212.65775		

12. Egy konzervgyárban termelési adatokat gyűjtöttek. Évente mérték a következő változók értékeit:

I. 1 munkásra + 1 alkalmazottra jutó termelési érték 1000 Ft-ban,
II. éves átlagbér 1000 Ft-ban,
III. exportra való kiszállítás vagonban,
IV. 1 tonna termék előállítására fordított munkás-munkaóra,
V. nyersanyag-feldolgozás vagonban.

Év	I.	II.	III.	IV.	V.
1965	111	15,20	1500	130	2600
1966	137	16,25	2650	109	4250
1967	152	16,45	3450	96	5500
1968	155	17,45	3450	97	5750
1969	145	17,85	3700	97	5800
1970	150	18,70	3400	93	4650
1971	156	19,30	3850	85	4800
1972	180	20,85	3850	80	6350
1973	209	23,45	4450	67	6750
1974	226	25,22	3650	60	6350

Keressünk lineáris összefüggést a 6 változó között! (Az export legyen adva a többi változó függvényében, $y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6$ alakban!)

Tekintsük a listát!

STEP-WISE MULTIPLE REGRESSION .... KONZER

NUMBER OF OBSERVATIONS 10

NUMBER OF VARIABLES 6

NUMBER OF SELECTIONS 6

CONSTANT TO LIMIT VARIABLES 0.0

VARIABLE NO.	MEAN	STANDARD DEVIATION
1	1969.50000	3.02765
2	162.09999	34.11238
3	19.07199	3.23645
4	3395.00000	805.34668
5	92.00000	19.13110
6	5280.00000	1241.01172

CORRELATION MATRIX

ROW 1					
1.00000	0.90745	0.95805	0.79404	−0.94955	0.76147
ROW 2					
0.90745	1.00000	0.96687	0.72985	−0.94459	0.82773
ROW 3					
0.95805	0.96687	1.00000	0.67913	−0.91955	0.72895
ROW 4					
0.79404	0.72985	0.67913	1.00000	−0.89641	0.90400
ROW 5					
−0.94955	−0.94459	−0.91955	−0.89641	1.00000	−0.88077
ROW 6					
0.76147	0.82773	0.72895	0.90400	−0.88077	1.00000

SELECTION 1
DEPENDENT VARIABLE 4
NUMBER OF VARIABLES FORCED 5
NUMBER OF VARIABLES DELETED 0
STEP 1
VARIABLE ENTERED 6
(FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP 4770306.000
PROPORTION REDUCED IN THIS STEP 0.817
CUMULATIVE SUM OF SQUARES REDUCED.......... 4770306.000
CUMULATIVE PROPORTION REDUCED ..... 0.817 OF 5837250.000
FOR 1 VARIABLES ENTERED
MULTIPLE CORRELATION COEFFICIENT........... 0.904
(ADJUSTED FOR D. F.)........................... 0.904
F-VALUE FOR ANALYSIS OF VARIANCE............ 35.768
STANDARD ERROR OF ESTIMATE.................. 365.196
(ADJUSTED FOR D. F.)........................... 365.196

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
6	0.58665	0.09809	5.981
INTERCEPT	297.51050		

STEP 2
VARIABLE ENTERED 5
(FORCED VARIABLE)

SUM OF SQUARES REDUCED IN THIS STEP 261306.313
PROPORTION REDUCED IN THIS STEP 0.045
CUMULATIVE SUM OF SQUARES REDUCED.......... 5031612.000
CUMULATIVE PROPORTION REDUCED ..... 0.862 OF 5837250.000
FOR 2 VARIABLES ENTERED
MULTIPLE CORRELATION COEFFICIENT........... 0.928
(ADJUSTED FOR D.F.) 0.919
F-VALUE FOR ANALYSIS OF VARIANCE............ 21.859
STANDARD ERROR OF ESTIMATE.................. 339.251
(ADJUSTED FOR D.F.) 359.829

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
6	0.33128	0.19242	1.722
5	−18.80806	12.48216	−1.507
INTERCEPT	3376.20068		

STEP 3
VARIABLE ENTERED 2
(FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP 713541.063
PROPORTION REDUCED IN THIS STEP 0.122
CUMULATIVE SUM OF SQUARES REDUCED.......... 5745153.000
CUMULATIVE PROPORTION REDUCED ..... 0.984 OF 5837250.000
FOR 3 VARIABLES ENTERED
MULTIPLE CORRELATION COEFFICIENT........... 0.992
(ADJUSTED FOR D.F.) 0.990
F-VALUE FOR ANALYSIS OF VARIANCE............ 124.763
STANDARD ERROR OF ESTIMATE.................. 123.893
(ADJUSTED FOR D.F.) 140.482

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
6	0.31825	0.07030	4.527
5	−61.91914	7.79488	−7.944
2	−25.15419	3.68933	−6.818
INTERCEPT	11488.71484		

STEP 4
VARIABLE ENTERED 1
(FORCED VARIABLE)

```
SUM OF SQUARES REDUCED IN THIS STEP ..........   6827.6Ø5
PROPORTION REDUCED IN THIS STEP ...............     Ø.ØØ1
CUMULATIVE SUM OF SQUARES REDUCED.......... 575198Ø.ØØØ
CUMULATIVE PROPORTION REDUCED ..... Ø.985 OF 583725Ø.ØØØ
FOR 4 VARIABLES ENTERED
  MULTIPLE CORRELATION COEFFICIENT...........      Ø.993
   (ADJUSTED FOR D.F.) ..........................  Ø.989
  F-VALUE FOR ANALYSIS OF VARIANCE...........     84.32Ø
  STANDARD ERROR OF ESTIMATE..................   13Ø.591
   (ADJUSTED FOR D.F.) ..........................159.941
 VARIABLE    REGRESSION    STD. ERROR OF   COMPUTED
  NUMBER     COEFFICIENT    REG. COEFF.     T-VALUE
    6          Ø.29Ø88        Ø.Ø858Ø        3.39Ø
    5        -68.12892       12.79948       -5.323
    2        -24.9ØØ51        3.9Ø94Ø       -6.369
    1        -33.76265       53.35995       -Ø.633
 INTERCEPT  78658.875ØØ

STEP 5
VARIABLE ENTERED ........... 3
 (FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP ..........  42179.16Ø
PROPORTION REDUCED IN THIS STEP ...............     Ø.ØØ7
CUMULATIVE SUM OF SQUARES REDUCED.......... 5794159.ØØØ
CUMULATIVE PROPORTION REDUCED ..... Ø.993 OF 583725Ø.ØØØ
FOR 5 VARIABLES ENTERED
  MULTIPLE CORRELATION COEFFICIENT...........      Ø.996
   (ADJUSTED FOR D.F.) ..........................  Ø.993
  F-VALUE FOR ANALYSIS OF VARIANCE...........    1Ø7.571
  STANDARD ERROR OF ESTIMATE..................   1Ø3.792
   (ADJUSTED FOR D.F.) ..........................139.251
 VARIABLE    REGRESSION    STD. ERROR OF   COMPUTED
  NUMBER     COEFFICIENT    REG. COEFF.     T-VALUE
    6          Ø.38935        Ø.Ø8442        4.612
    5        -91.19724       15.47257       -5.894
    2        -5Ø.7Ø19Ø       13.4Ø448       -3.782
    1       -27Ø.42554      126.9ØØ16       -2.131
    3        323.51685      163.49748        1.979
 INTERCEPT 544381.Ø625Ø
```


A STEP 1-ben jól látszik, hogy a legnagyobb súlyú, legnagyobb hatású változó a nyersanyag-feldolgozás vagonban (a regresszióba bevont változó, VARIABLE ENTERED...6), ez a teljes szórásnégyzet 81,7%-át magyarázza meg (PROPORTION REDUCED IN THIS STEP...0,817). Ha ez elegendő, akkor csak ezzel a változóval kifejezve adjuk meg az exportra való kiszállítást.

A további lépésekben ennél kisebb súlyú változókat von be (nézzük végig a PROPORTION REDUCED értékeket). A többszörös korrelációs együttható (MULTIPLE CORRELATION COEFFICIENT) egyre jobb lesz, a STEP 5-ben már 0,996.

A PROPORTION REDUCED az utolsó két lépésben már 0,001 és 0,007, igen kicsi. Megállhatunk tehát a 3. lépésben, mert ekkor a regressziós függvény:

$$y = 0{,}31825x_5 - 61{,}91914x_4 - 25{,}15419x_1 + 11488{,}71484.$$

A regressziós együtthatókat REGRESSION COEFFICIENT néven, a konstanst INTERCEPT néven olvashatjuk ki. A teljes szórásnégyzet regresszió által megmagyarázott része (CUMULATIVE PROPORTION REDUCED) 98,4%, ez a determinációs együttható (a többszörös korrelációs együttható, a 0,992 négyzete).

Az y exportkiszállításra ható legnagyobb súlyú tényezők tehát sorrendbe rakva:

- nyersanyag-feldolgozás (x_5),
- egy munkásra + egy alkalmazottra jutó termelési érték (x_1),
- egy tonna termék előállítására fordított munkaóra (x_4).

A többi változó hatását elhanyagoltuk, mert szórásnégyzetük csak kevéssé járul hozzá a teljes szórásnégyzethez.

13. Fonodai felvételi lapok összesítésével az alábbi adatokat nyerték az eltelt tízpercek sorszámára és a fonálszakadások számának átlagára:

Tízpercek sorszáma, x_i	Szakadásiszám-átlag, y_i	Tízpercek sorszáma, x_i	Szakadásiszám-átlag, y_i
1	11,7		
2	16,4	7	15,3
3	13,1	8	14,6
4	15,7	9	17,2
5	20,6	10	18,9
6	16,0	11	29,7

Ha a pontokat koordináta-rendszerben ábrázoljuk, azok egy parabola körül látszanak csoportosulni. Keressük az x és y közötti regressziós függvényt

$$y = a+bx+cx^2$$

alakban!

x	y	x^2	x^3	x^4	xy	x^2y
1	11,7	1	1	1	11,7	11,7
2	16,4	4	8	16	32,8	65,6
3	13,1	9	27	81	39,3	117,9
4	15,7	16	64	256	62,8	251,2
5	20,6	25	125	625	103,0	515,0
6	16,0	36	216	1 296	96,0	576,0
7	15,3	49	343	2 401	107,0	749,7
8	14,6	64	512	4 096	116,8	934,4
9	17,2	81	729	6 561	154,8	1 393,2
10	18,9	100	1000	10 000	189,0	1 890,0
11	29,7	121	1331	14 641	326,7	3 593,7
Összesen: 66	189,2	506	4356	39 974	1239,7	10 098,0

Így az a, b, c ismeretlen együtthatókra az alábbi normálegyenleteket kapjuk:

$$\begin{aligned} 189{,}2 &= 11a + 66b + 506c, \\ 1\,239{,}7 &= 66a + 506b + 4\,356c, \\ 10\,098{,}0 &= 506a + 4356b + 39\,974c. \end{aligned}$$

Az egyenletrendszert megoldva a regressziós parabola egyenlete:

$$y = 15{,}76 - 1{,}0175x + 0{,}164x^2.$$

A feltételezett regressziós görbére való illeszkedést méri a korrelációs index. Készítsük el az alábbi táblázatot:

y_i	$\hat{y}_i$	$y_i-\hat{y}_i$	$y_i-\bar{y}$	$(y_i-\hat{y}_i)^2$	$(y_i-\bar{y})^2$
11,7	14,9	−3,2	− 5,5	10,24	30,25
16,4	14,4	+2,0	− 0,8	4,00	0,84
13,1	14,2	−1,1	− 4,1	1,21	16,81
15,7	14,3	+1,4	− 1,5	1,96	2,25
20,6	14,7	+5,9	+ 3,4	34,81	11,56
16,0	15,7	+0,3	− 1,2	0,09	1,44
15,3	16,7	−1,4	− 1,9	1,96	3,61
14,6	18,1	−3,5	− 2,6	12,25	6,76
17,2	19,8	−1,6	0,0	2,56	0,00
18,9	21,9	−3,0	+ 1,7	9,00	2,89
29,7	24,4	+5,3	+12,5	28,09	156,25
Összesen:				106,17	232,36

A korrelációs index:

$$i = \sqrt{1-\frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\bar{y})^2}} = \sqrt{1-\frac{106{,}17}{232{,}36}} = 0{,}741.$$

A mérőszám közepesnél jobb illeszkedésre utal.

14. A gyapjúipar alapanyagai különböző mértékben reagálnak a fonás közben fellépő súrlódásra, mialatt statikus elektromosság keletkezik. Ez a feldolgozásnál akadályt jelent, mert megbontja az elemi szálak párhuzamosságát, és így bizonyos kuszáltságot idéz elő. Ez a kuszáltság a gépek lényeges hatásfokcsökkenéséhez vezet, rontja a végtermék minőségét, és gátolja a rendeltetésszerű felhasználást. Mindezek mellett gazdasági kár is keletkezik, romlik az anyagkihozatal. Az előfonalat speciális tárolóhelyen pihentetni szokták, miközben az részben vagy egészben leadja elektromos töltését. A pihentetési idő szükségessége egyértelmű, időtartama azonban vitatott az iparban dolgozó szakemberek körében. A sokéves tapasztalat alapján 6–15 napra becsülik a szükséges pihentetési időt.

A statikus elektromosság megszüntetési ideje és az anyaghozam között sok mérés eredményeképpen a következő kapcsolatot állapították meg. (Ez a kapcsolat sztochasztikus kapcsolat.)

Pihentetési idő, nap	Anyaghozam, %
3,0	91,4
3,5	92,0
4,5	92,3
5,5	92,6
6,0	93,4
7,0	94,2
7,5	95,0
8,0	94,6
9,0	94,4
10,0	94,2
11,0	93,2
12,0	91,8

Az anyaghozam és a pihentetési idő között a lineáris kapcsolat elég megbízhatatlannak mutatkozik (l. az első lépésnél (STEP 1) a többszörös korrelációs együtthatót, ami egyváltozós kapcsolatnál a közönséges korrelációs együtthatóval egyezik: 0,382). Ezt megerősíti, hogy ha az összetartozó adatoknak megfelelő pontokat a derékszögű koordináta-rendszerben ábrázoljuk. Látjuk, hogy ezek a pontok elég jól közelíthetők parabolával. Határozzuk meg a ponthalmazt legjobban közelítő $y = a+bx_1+cx_1^2$ parabolát! A parabola egyenlete alapján határozzuk meg azt a pihentetési időt, amelynél legnagyobb az anyagkihozatal!

Ha a keresett

$$y = a+bx_1+cx_1^2$$

egyenletben az $x_1^2=x_2$ helyettesítést hajtjuk végre, az

$$y = a+bx_1+cx_2$$

többváltozós lineáris kapcsolatra vezettük vissza az egyváltozós parabolikus regressziót. x_2 értékeként tehát az x_1^2 értékeket kell vennünk:

Anyaghozam, y, %	Pihentetési idő, x_1, nap	A pihentetési idő négyzete, x_2, nap
91,4	3,0	9,00
92,0	3,5	12,25
92,3	4,5	20,25
92,6	5,5	30,25
93,4	6,0	36,00
94,2	7,0	49,00
95,0	7,5	56,25
94,6	8,0	64,00
94,4	9,0	81,00
94,2	10,0	100,00
93,2	11,0	121,00
91,8	12,0	144,00

Tekintsük a listát!

```
STEP-WISE MULTIPLE REGRESSION ........... ANYAGH
NUMBER OF OBSERVATIONS        12
NUMBER OF VARIABLES            3
NUMBER OF SELECTIONS           3
CONSTANT TO LIMIT VARIABLES  Ø,Ø

  VARIABLE     MEAN     STANDARD
    NO.                 DEVIATION
     1       93.25826    1.22286
     2        7.25ØØØ    2.89592
     3       6Ø.25ØØØ   43.5791Ø

CORRELATION MATRIX
ROW  1
   1.ØØØØØ    Ø.38186    Ø.22723
ROW  2
   Ø.38186    1.ØØØØØ    Ø.98436
ROW  3
   Ø.22723    Ø.98436    1.ØØØØØ
```

```
SELECTION.......... 1
DEPENDENT VARIABLE........................... 1
NUMBER OF VARIABLES FORCED................. 2
NUMBER OF VARIABLES DELETED ................ Ø

STEP 1
VARIABLE ENTERED........... 2
  (FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP ..........          2.399
PROPORTION REDUCED IN THIS STEP ..............          Ø.146
CUMULATIVE SUM OF SQUARES REDUCED..........             2.399
CUMULATIVE PROPORTION REDUCED .......... Ø.146 OF 16.449
FOR 1 VARIABLES ENTERED
  MULTIPLE CORRELATION COEFFICIENT...........           Ø.382
    (ADJUSTED FOR D. F.)..........................      Ø.382
  F-VALUE FOR ANALYSIS OF VARIANCE.............         1.7Ø7
  STANDARD ERROR OF ESTIMATE.................           1.185
    (ADJUSTED FOR D. F.)..........................      1.185
   VARIABLE     REGRESSION    STD. ERROR OF   COMPUTED
    NUMBER      COEFFICIENT    REG. COEFF.     T-VALUE
      2           Ø.16125        Ø.12341        1.3Ø7
  INTERCEPT      92.Ø8922

STEP 2
VARIABLE ENTERED........... 3
  (FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP ..........         11.7Ø9
PROPORTION REDUCED IN THIS STEP ..............          Ø.712
CUMULATIVE SUM OF SQUARES REDUCED..........            14.1Ø8
CUMULATIVE PROPORTION REDUCED .......... Ø.858 OF 16.449
FOR 2 VARIABLES ENTERED
  MULTIPLE CORRELATION COEFFICIENT...........           Ø.926
    (ADJUSTED FOR D. F.)..........................      Ø.918
  F-VALUE FOR ANALYSIS OF VARIANCE...........          27.117
  STANDARD ERROR OF ESTIMATE.................           Ø.51Ø
    (ADJUSTED FOR D. F.)..........................      Ø.535
   VARIABLE     REGRESSION    STD. ERROR OF   COMPUTED
    NUMBER      COEFFICIENT    REG. COEFF.     T-VALUE
      2           2.15167        Ø.3Ø139        7.139
      3          -Ø.13437        Ø.Ø2ØØ3       -6.7Ø9
  INTERCEPT      85.75443
```


Röviden értékeljük a listát! Az anyaghozam (y) várhatóan így függ a pihentetési időtől (x_1-től):

$$y = -0{,}134\,37x_1^2 + 2{,}151\,67x_1 + 85{,}754\,43.$$

Mivel a négyzetes tag együtthatója negatív, a parabolának maximuma van ott, ahol az első derivált 0:

$$y' = -0{,}268\,74x_1 + 2{,}151\,67 = 0,$$
$$x = 8{,}006.$$

Vagyis várhatóan 8 nap pihentetési idő mellett legnagyobb az anyagkihozatal százaléka.

Mennyire bízhatunk meg a feltételezett kapcsolat jóságában?

A korrelációs mátrixból kiolvashatók a totális korrelációs együtthatók. Az y és x_1 között gyenge a lineáris kapcsolat (a korrelációs együttható 0,381 86). Az x_1 és x_1^2 között még gyengébb (0,227 23 a korrelációs együttható), amint azt előre várni is lehet. A totális korrelációs együtthatók nem mutatják jól a feltételezett másodfokú kapcsolat szorosságát, hiszen ezek a linearitást jellemző számok.

A többszörös korrelációs együttható azonban megnyugtat bennünket: ez összességében mutatja a y függését x_1-től és x_1^2-től: 0,926 jó korrelációra utal. Elég jól megbízhatunk a 8-napos pihentetési idő becslésében.

15. Egy árérzékeny terméknél a piackutató megfigyelte a kereslet és az egységár összefüggését:

Kereslet, y, 10 000 db	Egységár, x_1, font	Egységár reciprokának 100 000-szerese, x, $\frac{10^5}{\text{font}}$
26	114,51	873
28	111,76	896
30	109,38	914
32	107,30	930
34	105,46	949
36	103,83	965
38	102,37	986
40	101,05	990

Az $(x_1; y)$ pontpárokat megrajzolva, látszik, hogy a regressziós görbét

$$y = \frac{a}{x_1} + b$$

hiperbola alakban kereshetjük. Ha alkalmazzuk az $\frac{1}{x_1} = x$ transzformációt, a regressziót $y = ax+b$ lineáris alakra módosítottuk.

Kövessük az eredményeket:

```
STEP-WISE MULTIPLE REGRESSION ........... AREMEL
NUMBER OF OBSERVATIONS       8
NUMBER OF VARIABLES          2
NUMBER OF SELECTIONS         2
CONSTANT TO LIMIT VARIABLES  0.0
  VARIABLE     MEAN       STANDARD
    NO.                   DEVIATION
     1       33.000000    4.89898
     2      937.87500    42.23552

CORRELATION MATRIX
ROW  1
  1.00000   0.99491
ROW  2
  0.99491   1.00000

SELECTION ........... 1
DEPENDENT VARIABLE ............................ 1
NUMBER OF VARIABLES FORCED .................... 1
NUMBER OF VARIABLES DELETED ................... 0

STEP  1
VARIABLE ENTERED ........... 2
  (FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP ..........      166.293
PROPORTION REDUCED IN THIS STEP ...............       0.990
CUMULATIVE SUM OF SQUARES REDUCED .........         166.293
CUMULATIVE PROPORTION REDUCED ......... 0.990 OF 168.000
FOR 1 VARIABLES ENTERED
  MULTIPLE CORRELATION COEFFICIENT ..........        0.995
    (ADJUSTED FOR D. F.) ..........................  0.995
```



```
F-VALUE FOR ANALYSIS OF VARIANCE ..........        584.536
STANDARD ERROR OF ESTIMATE .................         0.533
  (ADJUSTED FOR D. F.) ..........................    0.533
 VARIABLE    REGRESSION     STD. ERROR OF   COMPUTED
  NUMBER     COEFFICIENT    REG. COEFF.     T-VALUE
     2         0.11540        0.00477        24.177
INTERCEPT    -75.23184
```

16. Az egy munkás által kezelt szövőgépek száma (x_i db) és az egy munkásra eső veszteség (y_i Ft/h) között keressünk másodfokú regressziót:

x_i	4	5	6	7	8	9	10	11	12	13
y_i	2,03	1,82	1,65	1,51	1,47	1,33	1,28	1,32	1,45	1,54

```
STEP-WISE MULTIPLE REGRESSION ........... VESZTE
NUMBER OF OBSERVATIONS       10
NUMBER OF VARIABLES           3
NUMBER OF SELECTIONS          3
CONSTANT TO LIMIT VARIABLES  0.0
  VARIABLE     MEAN       STANDARD
    NO.                   DEVIATION
     1        1.54000     0.23678
     2        8.50000     3.02765
     3       80.50000    52.03685

CORRELATION MATRIX
ROW  1
  1.00000   -0.73465   -0.62889
ROW  2
 -0.73465    1.00000    0.98911
ROW  3
 -0.62889    0.98911    1.00000
```

SELECTION 1
DEPENDENT VARIABLE 1
NUMBER OF VARIABLES FORCED 2
NUMBER OF VARIABLES DELETED .... Ø
STEP 1
VARIABLE ENTERED 2
(FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP Ø.272
PROPORTION REDUCED IN THIS STEP Ø.54Ø
CUMULATIVE SUM OF SQUARES REDUCED Ø.272
CUMULATIVE PROPORTION REDUCED .... Ø.54Ø OF Ø.5Ø5
FOR 1 VARIABLES ENTERED
MULTIPLE CORRELATION COEFFICIENT Ø.735
(ADJUSTED FOR D. F.) Ø.735
F-VALUE FOR ANALYSIS OF VARIANCE 9.38Ø
STANDARD ERROR OF ESTIMATE Ø.17Ø
(ADJUSTED FOR D. F.) Ø.17Ø

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
2	−Ø.Ø5745	Ø.Ø1876	−3.Ø63
INTERCEPT	2.Ø2836		

STEP 2
VARIABLE ENTERED 3
(FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP Ø.223
PROPORTION REDUCED IN THIS STEP Ø.441
CUMULATIVE SUM OF SQUARES REDUCED Ø.495
CUMULATIVE PROPORTION REDUCED .... Ø.981 OF Ø.5Ø5
FOR 2 VARIABLES ENTERED
MULTIPLE CORRELATION COEFFICIENT Ø.99Ø
(ADJUSTED FOR D. F.) Ø.989
F-VALUE FOR ANALYSIS OF VARIANCE 178.239
STANDARD ERROR OF ESTIMATE Ø.Ø37
(ADJUSTED FOR D. F.) Ø.Ø4Ø

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
2	−Ø.4Ø647	Ø.Ø2787	−14.585
3	Ø.Ø2Ø53	Ø.ØØ162	12.661
INTERCEPT	3.34229		

17. A magyarországi ügyvitelgépesítés költségeiről rendelkezésre állnak az alábbi adatok:

Gépi költség	20	25	35	63	70
Szervezési költség	90	100	120	150	200
Év	1952	1957	1962	1967	1972

Milyen regressziókat lehet itt számítani?

Vizsgálhatjuk a szervezési költséget mint az év és a gépi költség lineáris függvényét; vizsgálhatjuk a szervezési és gépi költségek lineáris összefüggését stb. Az utóbbiból leolvasható, hogy az egyik egységnyi növekedéshez várhatóan mennyivel nő a másik költség. Az utóbbira lássuk a listát:

STEP-WISE MULTIPLE REGRESSION FELDO2
NUMBER OF OBSERVATIONS 5
NUMBER OF VARIABLES 2
NUMBER OF SELECTIONS 2
CONSTANT TO LIMIT VARIABLES Ø.Ø

VARIABLE NO.	MEAN	STANDARD DEVIATION
1	42.59999	22.61194
2	132.ØØØØØ	44.38467

CORRELATION MATRIX
ROW 1
1.ØØØØØ Ø.955Ø4
ROW 2
Ø.955Ø4 1.ØØØØØ

SELECTION........... 1
DEPENDENT VARIABLE............................ 1
NUMBER OF VARIABLES FORCED.................. 1
NUMBER OF VARIABLES DELETED .... Ø

STEP 1
VARIABLE ENTERED 2
(FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP 1865.426
PROPORTION REDUCED IN THIS STEP Ø.912
CUMULATIVE SUM OF SQUARES REDUCED.......... 1865.426
CUMULATIVE PROPORTION REDUCED .... Ø.912 OF 2Ø45.2ØØ
FOR 1 VARIABLES ENTERED
MULTIPLE CORRELATION COEFFICIENT........... Ø.955
(ADJUSTED FOR D. F.)........................... Ø.955
F-VALUE FOR ANALYSIS OF VARIANCE............ 31.129
STANDARD ERROR OF ESTIMATE................. 7.741
(ADJUSTED FOR D. F.)........................... 7.741

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
2	Ø.48655	Ø.Ø872Ø	5.579
INTERCEPT	−21.62436		

18. Egy terméknél vizsgáljuk meg a bevétel és a termék előállítására fordított összeg változását 12 időszak alatt!

Ráfordítás, eFt	Az időszak sorszáma	Bevétel, eFt
15	1	18,1
20,1	2	25,2
22	3	26,0
30,2	4	38,2
41,1	5	49,7
35,5	6	47,5
41,2	7	50,9
47,6	8	61,1
39,0	9	55,2
48,1	10	62,3
50,2	11	63,5
49,3	12	64,6

Elemezzük a megoldási listát! Milyen regressziót kellene még megállapítani a $\text{költségszint} = \dfrac{\text{ráfordítás}}{\text{bevétel}}$ elemzéséhez és a minimális költségszint megállapításához? (Ha az időt minden határon túl növeljük, határértékként megkapjuk a minimális, el nem érhető, de túl sem léphető költségszintet.)

Bevétel = 1,047 06 · ráfordítás + 1,01553 · idő + 1,926 20. A többszörös korrelációs együttható 0,995, jó a feltételezett összefüggés. A páronkénti korrelációs együtthatók: a bevétel és ráfordítás közötti korrelációs együttható (0,991 39) kicsivel nagyobb, mint a bevétel és idő közötti (0,951 77). A bevétel növekedését a ráfordított összegnek és az idővel kapcsolatos tapasztalatoknak köszönhetjük.

A minimális költségszint megállapításához le kellene még futtatni a ráfordítási idő regressziószámítását. Ha így a $\text{költségszint} = \dfrac{a \cdot \text{idő} + b}{c \cdot \text{idő} + d}$ függvényt kapjuk, akkor elég hosszú időre a költségszint $\dfrac{a}{c}$-hez tart, ami a várhatóan minimális költségszint.

STEP-WISE MULTIPLE REGRESSION........... LINREG
NUMBER OF OBSERVATIONS 12
NUMBER OF VARIABLES 3
NUMBER OF SELECTIONS 3
CONSTANT TO LIMIT VARIABLES Ø.Ø

VARIABLE NO.	MEAN	STANDARD DEVIATION
1	36.6Ø829	12.21328
2	6.5ØØØØ	3.6Ø555
3	46.85829	16.33Ø55

CORRELATION MATRIX
ROW 1
1.ØØØØØ Ø.9291Ø Ø.99139
ROW 2
Ø.9291Ø 1.ØØØØØ Ø.95177

```
ROW  3
  Ø.99139    Ø.95177    1.ØØØØØ


SELECTION........... 1
DEPENDENT VARIABLE........................... 3
NUMBER OF VARIABLES FORCED................... 2
NUMBER OF VARIABLES DELETED ................ Ø
STEP 1
VARIABLE ENTERED........... 1
  (FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP ..........   2883.285
PROPORTION REDUCED IN THIS STEP ..............      Ø.983
CUMULATIVE SUM OF SQUARES REDUCED..........     2883.285
CUMULATIVE PROPORTION REDUCED ........ Ø.983 OF 2933.559
FOR 1 VARIABLES ENTERED
  MULTIPLE CORRELATION COEFFICIENT...........       Ø.991
    (ADJUSTED FOR D. F.)........................    Ø.991
  F-VALUE FOR ANALYSIS OF VARIANCE...........     573.521
  STANDARD ERROR OF ESTIMATE..................      2.242
    (ADJUSTED FOR D. F.)........................    2.242
   VARIABLE     REGRESSION     STD. ERROR OF    COMPUTED
    NUMBER      COEFFICIENT     REG. COEFF.      T-VALUE
       1          1.32561         Ø.Ø5535         23.948
  INTERCEPT      -1.66992


STEP 2
VARIABLE ENTERED........... 2
  (FORCED VARIABLE)
SUM OF SQUARES REDUCED IN THIS STEP ..........     2Ø.17Ø
PROPORTION REDUCED IN THIS STEP ..............      Ø.ØØ7
CUMULATIVE SUM OF SQUARES REDUCED..........     29Ø3.455
CUMULATIVE PROPORTION REDUCED ........ Ø.99Ø OF 2933.559
FOR 2 VARIABLES ENTERED
  MULTIPLE CORRELATION COEFFICIENT...........       Ø.995
    (ADJUSTED FOR D. F.)........................    Ø.994
  F-VALUE FOR ANALYSIS OF VARIANCE...........     434.Ø21
  STANDARD ERROR OF ESTIMATE..................      1.829
    (ADJUSTED FOR D. F.)........................    1.918
```



```
   VARIABLE     REGRESSION     STD. ERROR OF    COMPUTED
    NUMBER      COEFFICIENT     REG. COEFF.      T-VALUE
       1          1.Ø47Ø6         Ø.122Ø9          8.576
       2          1.Ø1553         Ø.41355          2.456
  INTERCEPT       1.9262Ø
```

19. A budapesti gázbekötésekhez szükséges vizsgálatok száma az egyes években a következő volt:

Év, x_i	1960	1962	1964	1966	1968	1970	1972
Vizsgálatok száma, y_i	1460	2500	7000	20 194	22 764	46 480	37 109

Milyen regressziót célszerű feltételezni a két változó között? Ha a regressziós függvény egyenletét ismernénk, mit lehetne ezzel kiszámolni?

A gyors emelkedés miatt kereshetnénk $y = ae^{bx}$ alakú exponenciális regressziót. Ha a, b értéke, valamint a budapesti háztartások száma ismert, akkor megtervezhetjük, hogy várhatóan melyik évben kap gázt az összes fővárosi háztartás. Az előtte levő években nyilván csökkeni fog a gázbekötések száma.

20. Készítsünk számítógép-programot a kétváltozós lineáris regresszióval kapcsolatos számításokra!

Az alábbi képletekben az összegzés mindig $i = 1$-től n-ig megy, egyesével:
– a regressziós egyenes egyenlete:

$$y = \underbrace{\frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}}_{\text{IRTG}}(x-\bar{x})+\bar{y},$$

– a korrelációs együttható:

$$r = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)(\Sigma y_i^2 - n\bar{y}^2)}},$$

– y értékek szórása:

$$s(y) = \sqrt{\frac{\Sigma y_i^2}{n} - \bar{y}^2},$$

– a maradékok szórása:

$$s(e) = s(y)\sqrt{1-r^2},$$

– az $R=0$ hipotézis vizsgálatához szükséges t érték (ha a két változó együttes eloszlása kétdimenziós normál eloszlás):

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}.$$

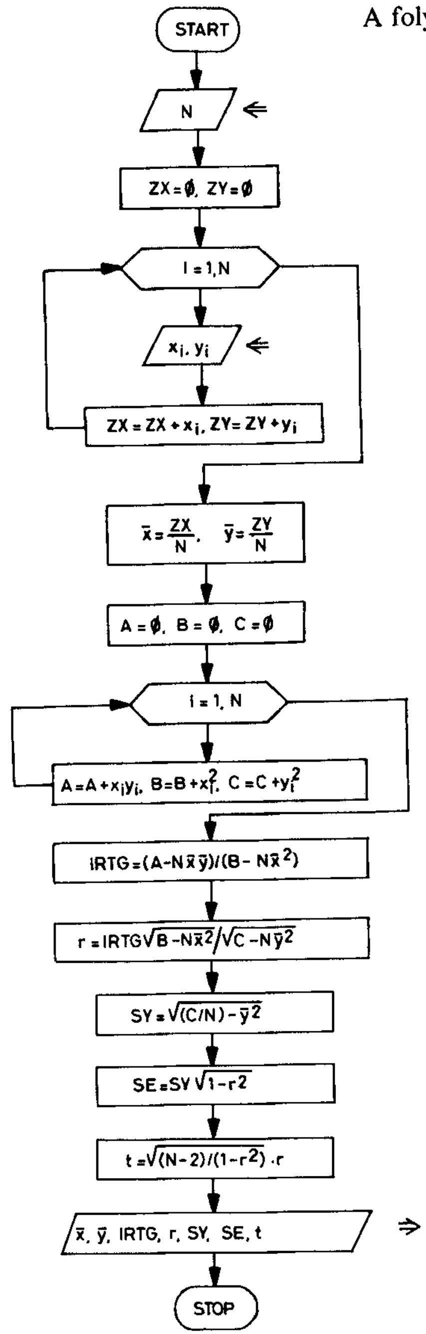
Észrevehető, hogy az r korrelációs együttható összefüggésben van a regressziós egyenes meredekségével (amit az IRTG nevű változóba fogunk elhelyezni):

$$r = \mathrm{IRTG} \cdot \sqrt{\Sigma x_i^2 - n\bar{x}^2} \,/\, \sqrt{\Sigma y_i^2 - n\bar{y}^2}.$$

A következő táblázat összefoglalja a szükséges tárakat és a tárak tartalmát.

Tár (változó)	Tartalom
N	n
ZX	Σx_i
ZY	Σy_i
XA	$\bar{x}$
YA	$\bar{y}$
A	$\Sigma x_i y_i$
B	Σx_i^2
C	Σy_i^2
IRTG	Meredekség
R	r korrelációs együttható
SY	y értékek szórása
SE	Maradékok szórása
T	t-érték

A folyamatábrát az 50. ábra mutatja.



50. ábra

A program BASIC nyelven:

```
5 REM LINEARIS REGRESSZIO
1Ø INPUT "MERESI PONTOK SZAMA=";N
15 DIM X(N),Y(N)
2Ø ZX=Ø: ZY=Ø
25 REM ADATOK BEOLVASASA
3Ø FOR I=1 TO N
4Ø ? "X"; I; "=";: INPUT X(I)
5Ø ? "Y"; I; "=";: INPUT Y(I)
6Ø ZX=ZX+X(I)
7Ø ZY=ZY+Y(I)
8Ø NEXT
85 REM ATLAGOK
9Ø XA=ZX/N: YA=ZY/N
95 REM A SZUKSEGES SZUMMAK KISZAMITASA
1ØØ A=Ø: B=Ø: C=Ø
11Ø FOR I=1 TO N
12Ø A=A+X(I)*Y(I)
13Ø B=B+X(I)↑2: C=C+Y(I)↑2
135 NEXT
137 REM A REGRESSZIOS JELLEMZOK KISZAMITASA
14Ø IRTG=(A-N*XA*YA)/(B-N*XA↑2)
15Ø R=IRTG*SQR(B-N*XA↑2)/SQR(C-N*YA↑2)
16Ø SY=SQR(C/N-YA↑2)
17Ø SE=SY*SQR(1-R↑2)
18Ø T=SQR((N-2)/(1-R↑2))*R
185 REM KIIRAS
19Ø ? TAB(5) "X ATLAG=";XA
195 ? TAB(5) "Y ATLAG=";YA
2ØØ ? TAB(5) "MEREDEKSEG M="; IRTG
2Ø5 ? TAB(5) "KORR.EGYH.=";R
21Ø ? TAB(5) "Y ERT.SZORASA=";SY
215 ? TAB(5) "MARADEKOK SZORASA="; SE
22Ø ? TAB(5) "T ERTEK=";T
23Ø END
```


Megjegyzés

Érdemes a programot a következőkkel bővíteni:

1. Nemcsak az iránytangens számértékét, hanem a regressziós egyenes teljes egyenletét kiíratni;

2. megadni egy olyan programsort, amely adott x-hez y regressziós becslését kiszámítja;

3. megrajzoltatni a számítógéppel a mérési pontokat és az egyenest.

A felhasználónak mindez komoly hasznot jelent. Korszerű statisztikai programokban ilyen jellegű programok találhatók. Már Hewlett–Packard asztali számítógépek is rendelkeznek rajzolóprogramokkal, nemlineáris regresszióhoz is.

21. Készítsünk lineáris regressziós zsebszámológép-programot HT–PTK–1050 zsebszámológépre!

Az n darab (x_i, y_i) „pont" *bevitele* a következőképpen történik:

x_1 x⇄t y_1 2nd Σ^+
x_2 x⇄t y_2 2nd Σ^+
– – – – – – – – – – –
x_n x⇄t y_n 2nd Σ^+

A billentyűzés hatására a *tárak tartalma:*

Tár	Tartalom
0	n
1	Σy_i
2	Σy_i^2
3	Σx_i
4	Σx_i^2
5	$\Sigma x_i y_i$
6	Ø
7	x_n+1

Az átlagokat, korrigálatlan szórásnégyzeteket az alábbi billentyűzéssel hívhatjuk le:

Billentyűzés	Megjelenik
Inv 2nd$\bar{x}$	$\bar{x}$
2nd$\bar{x}$	$\bar{y}$
Inv 2nd σ^2	σ_x^2
2nd σ^2	σ_y^2

A lineáris regressziónál megismert II. képletcsoportot át kell alakítanunk a gép rutinjaival számított mennyiségek függvényévé (átlagok, korrigálatlan szórások függvényévé). A program az egyenes egyenletét

$$y = mx+b$$

alakban számítja, a kijelzőn az m iránytangens és a b ordinátatengely-metszet jelenik meg.

$$m = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} = \frac{\frac{\Sigma x_i y_i}{n} - \bar{x}\bar{y}}{\sigma_x^2},$$

$\Sigma x_i y_i$ az 5. tárban, n a 0. tárban található, $\bar{x}$ az INV 2nd $\bar{x}$ billentyűzés után, $\bar{y}$ a 2nd $\bar{x}$ billentyűzés után jelenik meg a kijelzőn, σ_x^2-et az INV 2nd σ^2 billentyűzés adja.

A regressziós egyenes egyenletéből az ordinátatengely-metszet így képezhető:

$$y = m(x-\bar{x})+\bar{y} = mx+\underbrace{(\bar{y}-m\bar{x})}_{b}.$$

Miután m-et a programmal kiszámítottuk, $b = \bar{y}-m\bar{x}$ megadja az ordinátatengely-metszetet.

A korrelációs együttható II. csoportban megadott képletét átalakítva, az értéket megkapjuk az iránytangens, x és y tapasztalati korrigálatlan szórása segítségével:

$$r = \frac{\frac{\Sigma x_i y_i}{n} - \bar{x}\bar{y}}{\sqrt{\frac{\Sigma x_i^2}{n} - \bar{x}^2}\sqrt{\frac{\Sigma y_i^2}{n} - \bar{y}^2}} = $$

$$= \frac{\frac{\Sigma x_i y_i}{n} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \frac{\sigma_x}{\sigma_x} = \frac{m\sigma_x}{\sigma_y}.$$

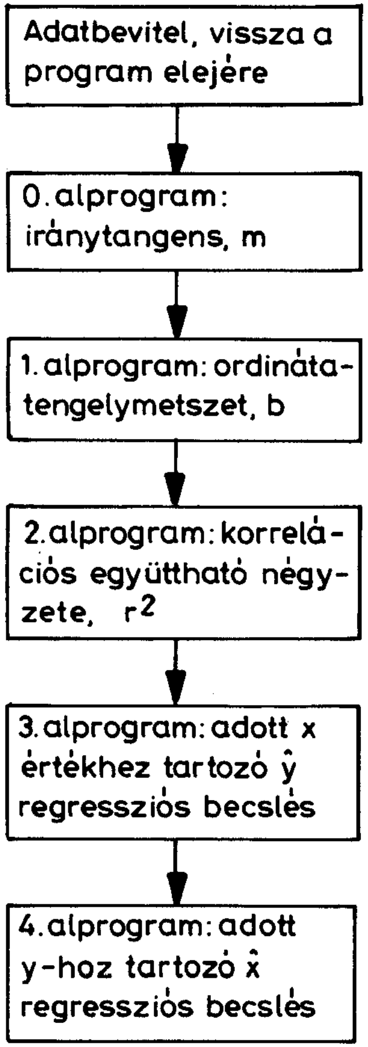
A programban r^2-et számíttatjuk, egyrészt azért, mert az a determinációs együtthatót adja meg, másrészt azért, mert a maximálisan kihasznált 50 programlépés nem engedi meg a négyzetgyökvonást. (Megjegyezzük, hogy kis ügyességgel ki lehet szűrni programlépéseket, és akkor beépíthető még a négyzetgyökvonás is.)

A program 5 alprogram (szubrutin) egymásutánja, az adatok bevitele után SBR s hívással meg kell adni, melyiket hívjuk (s a szubrutin sorszámát, a 0, 1, 2, 3, 4 sorszámok egyikét jelenti).

Az 51. ábrán látható, hogy melyik szubrutin mit csinál. (Figyeljük a tárrekeszek tartalmát, a billentyűzés hatását.)

Az 50-lépéses *program:*

Program	Magyarázat
2nd Σ^+ R/S RST	Adatbevitel
2nd Lbl Ø RCL 5 ÷ RCL Ø − 2nd$\bar{x}$ x	0. alprogram,
INV 2nd $\bar{x}$ = ÷ INV 2nd σ^2 = INV SBR	m (16 lépés)
2nd Lbl 1 SBRØ x INV 2nd$\bar{x}$ +/− +	1. alprogram,
2nd $\bar{x}$ = INV SBR	b (25 lépés)
2nd Lbl 2 SBRØ x² ÷ 2nd σ^2 x	2. alprogram,
INV 2nd σ^2 = INV SBR	r^2 (34 lépés)
2nd Lbl 3 STO 7 SBR Ø 2nd Prd 7	3. alprogram,
SBR1 SUM 7 RCL 7 INV SBR	$\hat{y}$ (42 lépés)
2nd Lbl 4 STO 7 SBR1 INV SUM7	4. alprogram,
SBRØ INV 2nd Prd 7 RCL 7 INV SBR	$\hat{x}$ (50 lépés)



51. ábra

Az *s* szubrutin hívása: 2nd Lbl *s*, a szubrutin vége: INV SBR.
Magyarázat a 3. szubrutinhoz:

$$\hat{y} = mx + b.$$

Magyarázat a 4. alprogramhoz:

$$\hat{x} = \frac{y-b}{m}.$$

Példa

x_i	y_i
1950	1
1955	2
1960	1,5
1965	2,5
1970	3

Adatbevitel (az *y* érték bevitele után szükséges 2nd Σ^+ a programba be van építve):

1950 $x \rightleftarrows t$ 1 R/S
1955 $x \rightleftarrows t$ 2 R/S
1960 $x \rightleftarrows t$ 1,5 R/S
1965 $x \rightleftarrows t$ 2,5 R/S
1970 $x \rightleftarrows t$ 3 R/S

Eredmény:

Billentyűzés	Kijelző
SBRØ	$0{,}09 = m$
SBR1	$-174{,}4 = b$, tehát $y = 0{,}09x - 174{,}4$
SBR2	$0{,}81 = r^2$
$\sqrt{x}$	$0{,}9 = r$
1975 SBR3	$3{,}35 = \hat{y}$
4 SBR4	$1982{,}2222 = \hat{x}$

22. Igazoljuk, hogy lineáris regresszió esetén a korrelációs index a korrelációs együttható abszolút értékével egyenlő!

Az igazolandó állítás:

$$I = \sqrt{1 - \frac{D^2(E)}{D^2(Y)}} = |R|.$$

Az E maradék szórásnégyzete kifejezhető az Y változó szórásnégyzetével és R^2-tel:

$$D^2(E) = D^2(Y)(1-R^2).$$

Ebből az összefüggésből:

$$R^2 = 1 - \frac{D^2(E)}{D^2(Y)},$$

amiből az igazolandó állítás már következik:

$$|R| = \sqrt{1 - \frac{D^2(E)}{D^2(Y)}} = I.$$

23. Igazoljuk n-dimenziós vektorok segítségével a regressziós egyenes egyenletére és a korrelációs együtthatóra adott I. képleteket!

Adott egy X változóra mért n-elemű $(x_1, x_2, \dots, x_n)$ minta, egy vele korrelációban levő másik, Y változóra mért $(y_1, y_2, \dots, y_n)$ minta, az értékpárok összetartozását az azonos index jelzi, az x_i-hez tartozó érték az y_i $(i = 1, 2, \dots, n)$. Legyenek az átlagok:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Vonjuk ki minden egyes mért értékből a megfelelő átlagot, egyesítsük az így kapott n számot egy-egy n-koordinátás (n-dimenziós) vektorba:

$$\mathbf{x} = (x_1-\bar{x}, x_2-\bar{x}, \dots, x_n-\bar{x}),$$
$$\mathbf{y} = (y_1-\bar{y}, y_2-\bar{y}, \dots, y_n-\bar{y}).$$

Jellemezzük most X és Y korrelációját $\mathbf{y}$ *és* $\mathbf{x}$ *lineáris függésével!* Tehát:

a) ha $\mathbf{x}$ merőleges $\mathbf{y}$-ra, akkor nincs közöttük lineáris összefüggés;

b) igen szoros az összefüggés, ha az egyik vektor lineárisan függ a másiktól, egyik a másik konstansszorosa:

$$\mathbf{y} = a\mathbf{x},$$

ahol az a pozitív vagy negatív valós szám (konstans);

c) jó az összefüggés, ha

$$\mathbf{y} = a\mathbf{x} + \mathbf{e},$$

ahol az $\mathbf{e}$ hibavektor abszolút értéke közeljár 0-hoz,

$$|\mathbf{e}| \approx 0.$$

Ekkor

$$\mathbf{y} \approx a\mathbf{x}.$$

Szeretnénk megállapítani az a szorzó értékét. A *b)* esetet véve alapul:

$$\mathbf{y} = a\mathbf{x},$$

kérdés, hogy hányszorosa $\mathbf{y}$ az $\mathbf{x}$-nak. Az $\mathbf{y}$ az $\mathbf{x}$ irányú egységvektornak, vagyis $\frac{\mathbf{x}}{|\mathbf{x}|}$-nek az $|\mathbf{y}|$-szerese, tehát

$$\mathbf{y} = |\mathbf{y}| \frac{\mathbf{x}}{|\mathbf{x}|}, \qquad \text{azaz} \qquad a = \frac{|\mathbf{y}|}{|\mathbf{x}|}.$$

Legyen jelen esetben az $\mathbf{x}$ és $\mathbf{y}$ vektor hajlásszöge 0, ennek koszinusza 1, így

$$\cos\alpha = \cos 0 = \frac{\mathbf{xy}}{|\mathbf{x}|\,|\mathbf{y}|} = 1,$$

ezért

$$\frac{\mathbf{xy}}{|\mathbf{x}|} = |\mathbf{y}|.$$

Helyettesítsük ezt az előbbi egyenletbe:

$$\mathbf{y} = \frac{\mathbf{xy}}{|\mathbf{x}|} \frac{\mathbf{x}}{|\mathbf{x}|} = \frac{\mathbf{xy}}{|\mathbf{x}|^2}\mathbf{x}.$$

Koordinátákra áttéve:

$$\mathbf{y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \mathbf{x}.$$

Ez azt jelenti, hogy $\mathbf{y}$ tetszőleges koordinátája arányos $\mathbf{x}$ hozzátartozó koordinátájával, az arányossági tényező az $\mathbf{x}$ előtt álló tört:

$$y - \bar{y} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}(x - \bar{x}).$$

Ugyanezek az összefüggések érvényesek az x és y koordináta között, ha a két vektor hajlásszöge 180° (csak most az említett arányossági tényező, a meredekség negatív).

c) Minél közelebb jár a két vektor hajlásszöge 0-hoz (minél közelebb van $|\mathbf{e}|$ a 0-hoz), annál jobb közelítésben érvényes x és y között a fenti összefüggés.

Jellemezzük a két vektor közötti lineáris összefüggés erősségét a köztük levő szög koszinuszával! Ekkor:

$$-1 \leqq \cos\alpha = \frac{\mathbf{x}\mathbf{y}}{|\mathbf{x}|\,|\mathbf{y}|} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \leqq 1.$$

a) Ha a két vektor merőleges, $\alpha = 90°$, $\cos\alpha = 0$. A lineáris függetlenséget a mérőszám 0 volta jelzi.

b) Az igen szoros lineáris összefüggést az mutatja, hogy a mérőszám 1 vagy -1 ($\cos\alpha = \cos 0° = 1$ vagy $\cos\alpha = \cos 180° = -1$ alapján).

c) Ha $\alpha \approx 0°$, vagyis $\cos\alpha$ közel van 1-hez, ill. ha $\alpha \approx 180°$, vagyis $\cos\alpha$ közel van (-1)-hez, akkor jó a lineáris függés,

$$\mathbf{y} \approx a\mathbf{x},$$

$$y - \bar{y} \approx \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}(x - \bar{x}),$$

adott x esetén a kapott egyenlettel y-ra jó becslést kaphatunk.

Észrevehető, hogy *a kapcsolat szorosságára kapott mérőszám a tapasztalati korrelációs együtthatóval, r-rel egyezik.*

Észrevehető, hogy *a kapott, y-t becslő egyenlet a regressziós egyenes jól ismert egyenletével egyezik meg.*

A lineáris regressziónál kapott eredményeket tehát három úton is megkaphatjuk: 1. elméleti regresszióval, 2. a legkisebb négyzetek elvével, 3. n-dimenziós vektorokkal.

24. *a)* Mutassuk ki, hogy ha a kétváltozós lineáris regressziónál y-nak a legkisebb négyzetek elve alapján legjobban közelítő lineáris becslése az

$$y = ax + b$$

másodfajú regressziós egyenessel történik, akkor ez az egyenes átmegy a várható értékekből kapható $(M(X), M(Y))$ súlyponton!

b) Ezek alapján keressük egyváltozós differenciálszámítással a „legjobban közelítő" másodfajú regressziós egyenes egyenletét!

Tegyük fel, hogy az Y-nak a legkisebb négyzetek elve alapján legjobban közelítő lineáris becslése $aX + b$. Ekkor

$$Y = aX + b + E,$$

itt az E véletlen hiba olyan, hogy várható értéke $M(E) = 0$.

a) Kimutatjuk, hogy az $(M(X), M(Y))$ súlypont rajta van az $y = ax + b$ regressziós egyenesen:

$$M(Y) = M(aX + b + E) = M(aX) + M(b) + M(E),$$
$$M(Y) = aM(X) + b,$$

ez valóban azt igazolja, hogy a súlypont rajta van a regressziós egyenesen, hiszen koordinátái kielégítik az egyenletét. Ennek alapján kereshetjük a másodfajú regressziós egyenes egyenletét a következő alakban:

$$y = a(x - \underbrace{M(X)}_{m_1}) + \underbrace{M(Y)}_{m_2} = a(x - m_1) + m_2.$$

b) Ekkor a súlyponton átmenő egyenesek közül keressük a legkisebb négyzetek elve alapján a legjobban közelítőt! Olyan a értéket keresünk tehát, amelyre

$$Q(a) = M[(Y - a(X - m_1) - m_2)^2] =$$
$$= M[(Y - m_2)^2] - 2aM[(X - m_1)(Y - m_2)] + a^2 M[(X - m_1)^2]$$

minimális lesz.

A minimum ott lehet, ahol az első derivált 0.

$$\frac{dQ}{da} = -2M[(X-m_1)(Y-m_2)]+2aM[(X-m_1)^2] = 0,$$

$$a = \frac{M[(X-m_1)(Y-m_2)]}{M[(X-m_1)^2]}.$$

Minimum biztosan akkor van, ha a szóban forgó helyen a második derivált pozitív:

$$\frac{d^2Q}{da^2} = 2M[(X-m_1)^2] > 0.$$

Tehát valóban minimum van. A pontokat legjobban közelitő egyenes egyenlete:

$$y = \frac{M[(X-m_1)(Y-m_2)]}{M[(X-m_1)^2]}(x-m_1)+m_2.$$

Ennek az n darab $(x_1, y_1)\,(x_2, y_2)\,\ldots,(x_n, y_n)$ pontokból történő becslése (a számlálóban, nevezőben $\frac{1}{n}$-nel egyszerűsítünk):

$$y = \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}(x-\bar{x})+\bar{y}.$$

Ezzel *a másodfajú regressziós egyenes egyenletének egy egyváltozós differenciálszámítással történő levezetését adtuk meg* (a szokásos kétváltozós helyett). Kétdimenziós normál eloszlásnál ez az egyenes egybeesik az elsőfajú regressziós egyenessel.

25. Az $(x_1; y_1)\,\ldots,(x_n; y_n)$ pontokat ábrázolva, azok látszólag egy egyenes körül csoportosulnak. A két változó között lineáris összefüggést tételezünk fel:

$$y = ax+b.$$

Keressük a korrelációs (regressziós) egyenes egyenletét, másként: a ponthalmazt legjobban közelítő egyenes egyenletét!

Mit értsünk a pontokat legjobban megközelítő egyenesen?

Nem jól jellemezzük a tényleges y_i és a becsült $\hat{y}_i = ax_i+b$ értékek eltérését, ha ezeket az eltéréseket összegezzük, hiszen a

$$\sum_{i=1}^{n}(y_i-ax_i-b)$$

összeg akkor is 0, ha minden eltérés 0, de akkor is, ha a pozitív és negatív eltérések „lerontják", kiegyenlítik egymást. Hogy nemnegatív eltéréseket kaphassunk, célszerű inkább a mért és az egyenlettel becsült y értékek eltérésének négyzetösszegét venni. Az az $y = ax+b$ egyenes közelíti meg legjobban a pontokat, amelyre nézve a mért és az egyenlettel becsült y értékek eltérésének a négyzetösszege minimális:

$$\sum_{i=1}^{n}(y_i-ax_i-b)^2 \to \min.$$

(Ez a Gausstól származó elv a legkisebb négyzetek elve.)

Olyan a és b értékeket kell tehát meghatároznunk, hogy a fenti összeg minimális legyen. Az összeg az a és b paraméterek függvénye:

$$F(a; b) = \sum(y_i-ax_i-b)^2,$$

bizonyítható, hogy ennek minimuma lesz biztosan ott, ahol az elsőrendű parciális deriváltjai 0-k lesznek:

$$F'_a = 2\sum_{i=1}^{n}(y_i-ax_i-b)(-x_i) = 0,$$

$$F'_b = 2\sum_{i=1}^{n}(y_i-ax_i-b)(-1) = 0.$$

Az első egyenletben (-2)-vel osztva, és a tagokat egyenként összegezve:

$$\Sigma x_iy_i-a\,\Sigma x_i^2-b\,\Sigma x_i = 0;$$

a második egyenletben (-2)-vel osztva, és a tagokat egyenként összegezve (az utolsó tagban b-t n-szer kell összeadni):

$$\Sigma y_i-a\,\Sigma x_i-nb = 0,$$

$$\frac{\Sigma y_i-a\,\Sigma x_i}{n} = b,$$

utóbbit az ismeretlen a-ra, b-re kapott első egyenletbe helyettesítve:

$$\Sigma x_i y_i - a\Sigma x_i^2 - \frac{\Sigma y_i - a\Sigma x_i}{n}\Sigma x_i = 0.$$

Gyűjtsük az ismeretlen a-t tartalmazó tagokat az egyik oldalra, a többit a másik oldalra:

$$a\left(\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right) = \Sigma x_i y_i - \frac{\Sigma x_i \, \Sigma y_i}{n},$$

$$a = \frac{n\,\Sigma x_i y_i - \Sigma x_i \, \Sigma y_i}{n\,\Sigma x_i^2 - (\Sigma x_i)^2}.$$

Ha már az a ismert, akkor b kiszámítható:

$$b = \frac{\Sigma y_i - a\,\Sigma x_i}{n}.$$

Ebből az egyenes egyenlete,

$$y = ax + b$$

felírható. Alakítsuk még tovább a kapott eredményt, használjuk fel, hogy $\frac{\Sigma y_i}{n} = \bar{y}$, $\frac{\Sigma x_i}{n} = \bar{x}$:

$$y = ax + \frac{\sum y_i - a\Sigma x_i}{n} = a\left(x - \frac{\Sigma x_i}{n}\right) + \frac{\Sigma y_i}{n},$$

$$y = a(x-\bar{x}) + \bar{y}.$$

Az a értékét alakítsuk úgy, hogy $\bar{x}$, $\bar{y}$ kerüljenek be! Osszunk a számlálóban, nevezőben n-nel:

$$a = \frac{n\,\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{\Sigma x_i y_i - \frac{\Sigma x_i}{n}\Sigma y_i}{\Sigma x_i^2 - \frac{\Sigma x_i}{n}\Sigma x_i} =$$

$$= \frac{\Sigma x_i y_i - \bar{x}n\frac{\Sigma y_i}{n}}{\Sigma x_i^2 - \bar{x}n\frac{\Sigma x_i}{n}} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}.$$

Így a pontokat legjobban közelítő egyenes egyenlete:

$$y = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}(x-\bar{x}) + \bar{y}. \qquad \text{(II)}$$

26. *Mutassuk ki, hogy a most kapott (II) egyenlet* azonos az (I) képletcsoportban kapott egyenlettel, vagyis az

$$y = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}(x-\bar{x}) + \bar{y} \qquad \text{(I)}$$

egyenlettel!

Látszik, hogy csak az x együtthatóinak az egyenlőségét kell kimutatnunk. Az (I) egyenletből:

$$\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = \frac{\Sigma(x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y})}{\Sigma(x_i^2 - 2x_i\bar{x} + \bar{x}^2)} =$$

$$= \frac{\Sigma x_i y_i - \bar{x}n\frac{\Sigma y_i}{n} - \bar{y}n\frac{\Sigma x_i}{n} + \Sigma\bar{x}\bar{y}}{\Sigma x_i^2 - 2\bar{x}n\frac{\Sigma x_i}{n} + \Sigma\bar{x}^2} =$$

$$= \frac{\Sigma x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\Sigma x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}.$$

(Felhasználtuk, hogy $\sum_{i=1}^{n} \bar{x}\bar{y} = n\bar{x}\bar{y}$, az $\bar{x}\bar{y}$ szorzatot n-szer kell összeadni; $\sum_{i=1}^{n} \bar{x}^2 = n\bar{x}^2$, mert a Σ jel n összeadást jelent.)

Ez már egyezik a (II) egyenletben az iránytangensre (az x együtthatóra) kapott kifejezéssel.

– Ugyanígy igazolható, hogy a korrelációs együtthatóra is kaphatunk egy (II) típusú kifejezést:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. \qquad \text{(II)}$$

Feladatainkban a két egyenlettípus közül mindig azt célszerű használni, amelyikkel az adott feladatban könnyebb számolni.

27. Adjuk meg a lineáris regressziónál kapott normálegyenletek megoldását az ismeretlen regressziós együtthatókra mátrixokkal!

A bizonyítást csak kétváltozós lineáris regresszióra adjuk meg!

A legkisebb négyzetek elvén alapuló levezetésnél az $y = ax+b$ egyenletben álló a és b meghatározására az alábbi egyenletrendszert kaptuk:

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} 1$$

$$\underline{\Sigma x_i y_i = a\Sigma x_i^2 + b\Sigma x_i}$$

Észrevehető, hogy az első egyenlet az egyenes $y = ax+b$ egyenletéből formálisan a Σ jellel mint operátorral való „szorzás" útján kapható, a második egyenlet formálisan az előtte levőből az x_i-vel a Σ jel mögé való beszorzás útján kapható.

Azért érdemes erre felfigyelni, mert ez a módszer az $y = ax^2+bx+c$, $y = ax^3+bx^2+cx+d$ stb. alakú nemlineáris regressziós függvényeknél is érvényes, könnyű megjegyzési lehetőséget ad az ismeretlen együtthatók kiszámítására szolgáló, ún. normálegyenletekhez.

Az ismeretlen a és b együtthatók meghatározása mátrixokkal is történhet.

A kiinduló adatokat az alábbi két mátrix tartalmazza:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix}.$$

Az $y = ax+b$ ismeretlen együtthatóit tartalmazza az alábbi mátrix:

$$\mathbf{b} = \begin{bmatrix} b \\ a \end{bmatrix}.$$

Szorozzuk meg $\mathbf{X}$ transzponáltját jobbról az $\mathbf{X}$ mátrixszal, az eredmény egy $2 \cdot 2$-es négyzetmátrix:

$$\mathbf{X^*X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} \Sigma 1 & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$$

$\left(\sum_{i=1}^{n} 1 = n \text{ természetesen}\right)$. Az $y = b+ax$ egyenletből formálisan könnyen kapható

$$\Sigma y_i = b\Sigma 1 + a\Sigma x_i,$$
$$\Sigma x_i y_i = b\Sigma x_i + a\Sigma x_i^2$$

normálegyenlet-rendszernek a legutóbb kapott mátrix az együttható mátrixa (az ismeretlenek a és b!), a bal oldali állandókat az $\mathbf{X^*Y}$ szorzat állítja elő:

$$\mathbf{X^*Y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

Az egyenletrendszer tömören így írható:

$$\mathbf{X^*Y} = \mathbf{X^*Xb}.$$

Az ismeretlen $\mathbf{b}$ mátrixot úgy kapjuk meg, hogy mindkét oldalt szorozzuk balról az együtthatómátrix inverzével:

$$(\mathbf{X^*X})^{-1}\mathbf{X^*Y} = \mathbf{b} = \begin{bmatrix} b \\ a \end{bmatrix}.$$

A levezetés azért érdekes, mert a többváltozós lineáris regressziónál is hasonló módszerrel számíthatjuk ki az ismeretlen együtthatókat.

Keressük például

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots b_pX_p$$

alakban Y regressziós becslését az $X_1, X_2, \dots, X_p$ függvényeként! A mért adatokat az alábbi két mátrix tartalmazza:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix}.$$

X második oszlopában áll az X_1-re mért n érték, harmadik oszlopában az X_2-re mért n érték stb.

A regressziós együtthatók **b** oszlopvektora az **Y** és **X** mátrixok segítségével ugyanúgy számítható, mint előbb:

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_p \end{bmatrix} = (\mathbf{X^*X})^{-1}\mathbf{XY}.$$

28. Igazoljuk, hogy a Gauss-féle normálegyenletek képzési elve kétváltozós másodfokú, harmadfokú, ..., n-edfokú regresszióra is érvényes!

A levezetést csak másodfokú regresszióra végezzük el!

Ha két sztochasztikus változóra mért $(x_i;\, y_i)$ pontpárok a síkban egy parabola körül csoportosulnak, akkor a két változó között az

$$y = a + bx + cx^2$$

parabolikus regressziót tételezzük fel. Az n darab pontot legjobban közelítő parabola egyenletét annak alapján kívánjuk meghatározni, hogy a mért és számított y értékek különbségének négyzetösszege minimális legyen (legkisebb négyzetek elve), vagyis

$$F(a, b, c) = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2 \to \min.$$

Bizonyítható, hogy ez teljesülni fog akkor, ha az a, b, c ismeretlen együtthatók szerinti parciális deriváltak 0-val egyenlők:

$$F'_a = \Sigma 2(y_i - a - bx_i - cx_i^2)(-1) = 0,$$
$$F'_b = \Sigma 2(y_i - a - bx_i - cx_i^2)(-x_i) = 0,$$
$$F'_c = \Sigma 2(y_i - a - bx_i - cx_i^2)(-x_i^2) = 0.$$

Az egyenleteket rendre (-2)-vel végigosztva, az összegezéseket tagról tagra elvégezve az alábbi ún. normálegyenletekhez jutunk:

$$\Sigma y_i - a\Sigma 1 - b\Sigma x_i - c\Sigma x_i^2 = 0,$$
$$\Sigma x_i y_i - a\Sigma x_i - b\Sigma x_i^2 - c\Sigma x_i^3 = 0,$$
$$\Sigma x_i^2 y_i - a\Sigma x_i^2 - b\Sigma x_i^3 - c\Sigma x_i^4 = 0.$$

Ezekből a-t, b-t és c-t meghatározhatjuk.

Figyeljük meg, hogy a normálegyenleteket formálisan itt is a regressziós egyenlet

$$y - a - bx - cx^2 = 0$$

alakjából kapjuk úgy, hogy formálisan végigszorozzuk az egyenletet Σ jellel (első egyenlet), a Σ jel mögé x-szel szorzunk (második egyenlet), a Σ jel mögé x^2-tel szorzunk (harmadik egyenlet). Természetesen mindegyik Σ i-re vonatkozik, mögötte is a mért x_i, y_i értékeket vagy azok szorzatát, hatványát kell összegeznünk.

Ez könnyű megjegyzési lehetőséget ad a normálegyenletekhez. Bizonyítható, hogy ez az eljárás magasabb fokú $y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$ regressziók esetén is érvényes.

29. Két tanár $n = 10$ diákot rangsorolt a képességük alapján:

Diák	A	B	C	D	E	F	G	H	I	J
1. rangsor	2	1	3	4	6	5	8	7	10	9
2. rangsor	3	2	1	4	6	7	5	9	10	8

Van-e egyezés a két tanár között elbírálásban?

A Spearman-féle rangkorrelációs együtthatóval számolunk. Képezzük először a rangkülönbségeket és azok négyzetét!

1. rangsor	2	1	3	4	6	5	8	7	10	9
2. rangsor	3	2	1	4	6	7	5	9	10	8
Rangkülönbségek, d_i	1	1	-2	0	0	2	-3	2	0	-1
Rangkülönbségek négyzete, d_i^2	1	1	4	0	0	4	9	4	0	1

A Spearman-féle rangkorrelációs együttható mintabeli becslése az $n=10$ elemű mintából:

$$r = 1 - \frac{6\Sigma d_i^2}{n^3-n} = 1 - \frac{6 \cdot 24}{1000-10} = 0{,}85.$$

Ez közeljár 1-hez, ennek alapján jó, pozitív korreláció (egyezés) tételezhető fel a két bíráló közötti véleményadásban.

Felmerülhet azonban, hogy nem csupán a mintában kapunk ilyen jó egyezést. Ha a két tanár összes elbírálási rangsorát összehasonlítjuk, nem lenne-e az elméleti rangkorrelációs együttható, R, nullával egyenlő?

Ha $n \geq 10$, akkor H_0: $R=0$ esetén a

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

próbastatisztika közelítőleg Student-eloszlású, $n-2$ szabadságfokkal. Ezzel teszteljük az $R=0$ feltevést.

$$|t| = 0{,}85\sqrt{\frac{8}{1-0{,}85^2}} = 4{,}55 \geq t_{95} = 2{,}306,$$

95% biztonsági szinten elvetjük az $R=0$ feltevést, a két megfigyelő egyetértése szignifikáns.

A Student-táblázat megfelelő részlete:

$n=f+1$	95%
9	2,306

30. Van tíz, egyre nehezebb súlyunk és valakinek a súlyérzékét akarjuk kipróbálni. Az általa adott rangsorolás a következő:

Valódi rangsor	1	2	3	4	5	6	7	8	9	10
Adott rangszámok	3	2	4	1	7	5	10	6	9	8

A kétféle rangszám közti eltérés és négyzete:

d_i	2	0	1	−3	2	−1	3	−2	0	−2
d_i^2	4	0	1	9	4	1	9	4	0	4

Így

$$r = 1 - \frac{6 \cdot 36}{1000-10} = 0{,}78,$$

talán még elég jó az illető súlyérzéke (a minta alapján). És ha a rangsorolásbeli egyezésre megint „egész sokaságot” vennénk figyelembe, hogy állja meg helyét a H_0: $R=0$ hipotézis?

$$|t| = 0{,}78\sqrt{\frac{8}{1-0{,}78^2}} = 3{,}5 > t_{95} = 2{,}306.$$

95% biztonsági szinten elvetjük, hogy csak véletlenül kaptunk 0,78 korrelációs együtthatót, tehát elvetjük, hogy a valóságban $R=0$ lenne vagyis, hogy rossz volna az illető súlyérzéke.

31. $m=5$ bíráló ítélt meg $n=7$ fagylaltkészítményt. Az ízlési rangsorok:

1. rangsor	2	4	3	7	5	1	6
2. rangsor	4	5	2	3	6	1	7
3. rangsor	1	3	2	4	6	5	7
4. rangsor	3	1	4	2	7	6	5
5. rangsor	1	3	5	7	6	2	4
Rangösszegek:	11	16	16	23	30	15	29

Van-e egyetértés (konkordancia) a bírálók közt?

Ha egyáltalán nincs egyetértés, az ún. várt rangösszeg:

$$Z = \frac{m(n+1)}{2} = \frac{5 \cdot 8}{2} = 20,$$

az egy-egy fagylaltnál megfigyelt rangösszeg és a várt rangösszeg eltéréseinek négyzetösszege, $S = 328$, ugyanis:

Rangösszegek	11	16	16	23	30	15	29	Σ
Eltérések 20-tól	9	4	4	−3	−10	5	−9	
Eltérések négyzete	81	16	16	9	100	25	81	328

Így az egyetértés fokát jellemző konkordancia-együttható:

$$W = \frac{12S}{m^2(n^3-n)} = \frac{12 \cdot 328}{25(343-7)} = 0{,}47.$$

$W=0$ a rangszámok össze-vissza, teljesen véletlenszerű elhelyezését jelenti, $W=1$ a bírálók véleményének tökéletes egyezését. $W=0{,}47$ közepes egyetértést jelent.

Véletlen-e ez az érték, $W=0$ feltehető-e az egész sokaságban? A korrigált W érték:

$$W_{\text{korr}} = \frac{S-1}{\frac{m^2(n^3-n)}{12}+2} = \frac{327}{\frac{25(343-7)}{12}+2} = 0{,}47,$$

az F-próbához szükséges F érték:

$$F = \frac{(m-1)W_{\text{korr}}}{1-W_{\text{korr}}} = \frac{4 \cdot 0{,}47}{0{,}53} = 3{,}5.$$

Az f_1 szabadságfok:

$$f_1 = n-1-\frac{2}{m} = 6-\frac{2}{5} = 5{,}6,$$

az f_2 szabadságfok:

$$f_2 = (m-1)\left(n-1-\frac{2}{m}\right) = 4 \cdot 5{,}6 = 22{,}4.$$

Az F-eloszlás 9. táblázatából a szükséges részlet:

f_2	f_1	
	5	6
22	2,66	2,55
24	2,62	2,51

A számított $F=3{,}5$, a táblázatbeli ennél kisebb. $W=0$ tehát 95%-os biztonsági szinten elvetendő, a kapott mintabeli eltérés (0,47) szignifikáns (statisztikailag jelentős).

32. Végezzük el az alábbi vizsgálatot, amely oktatási kérdőívek rangkorreláció- és konkordancia-értékelését mutatja be (számítógép segítségével)!

Főiskolai hallgatóknak kiosztott kérdőíveken felmérték, hogy a hallgatóság a tanult tárgyakat tetszés szempontjából hogyan rangsorolta, és milyen eredményeket ért el az egyes tárgyakból.

TPA/i számítógép segítségével kiszámították, hogy az átlagok és szórások alapján milyen rangsorba sorolhatók az egyes tárgyak, homogénnek tekinthető-e az évfolyam véleményadás szempontjából, milyen rangkorreláció van a tetszés foka és a tárgyban elért eredmények között. A számítógépprogram folyamatábrája az 52. ábrán látható.

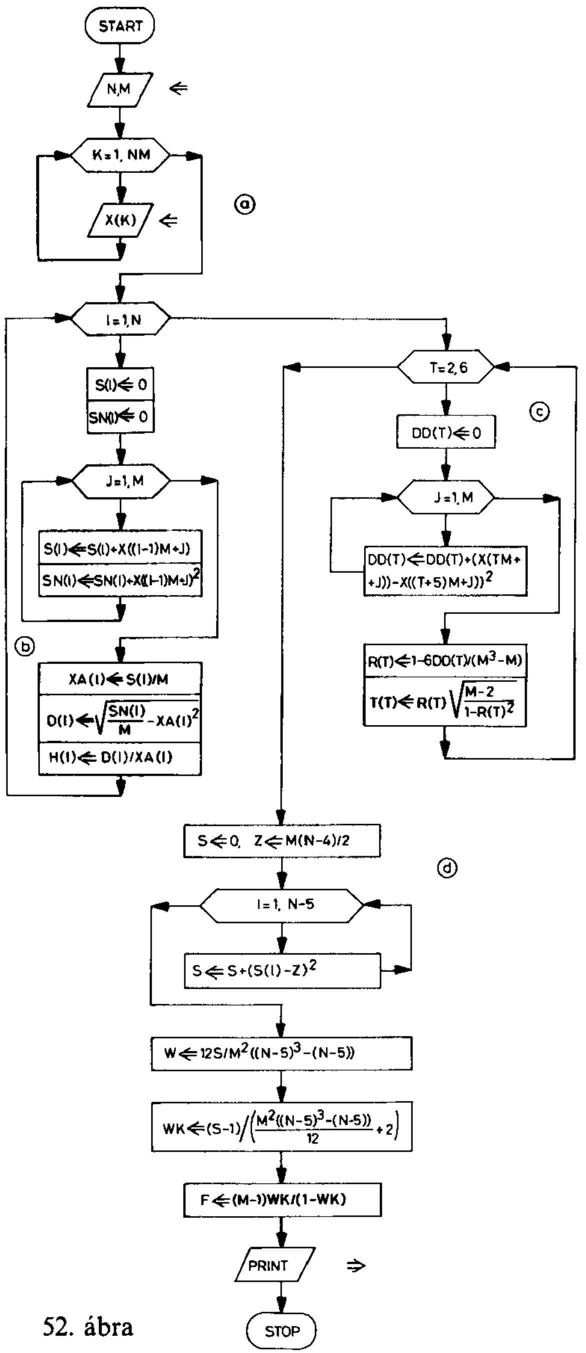
Az évfolyam véleményének homogén voltát értékeli az átlagrangsorral való rangkorrelációk vizsgálata, ezeknek a rangkorrelációs együttható abszolút értékeknek az átlaga.

Az oktatással foglalkozók többségét érdeklik az alábbi kérdések:

– szeretik-e az oktatott tárgyat a hallgatók,
– mennyire befolyásolja a tárgyból elért eredményt a tárgy iránti érdeklődés,
– a tárgy iránti érdeklődésben, a tárgy megítélésében mennyire homogén a hallgatóság.

A második kérdésre mindenki azt mondja: világos, hogy a tárgy szeretete kapcsolatban, korrelációban van a belőle elért eredménnyel. De hogy *milyen erős korrelációban,* azt előre már nem lehet megmondani, csak mérés döntheti el.

A Bánki Donát Gépipari Műszaki Főiskola egy évfolyamán a műszaki tanárszakos hallgatók mindegyike névtelen *adatlapon* beadta



52. ábra

– a tanult 7 tárgy *kedveltségi sorrendjét* (1-es rang a legkedveltebb, 2-es a kevésbé kedvelt, és így tovább, 7-es rang a legkevésbé kedvelt tárgy mellett),

– a tanult tárgyból szerzett 2 jegyét (pl. gyakorlati és vizsgajegyét, vagy pl. 2 vizsgaosztályzatot). Az ezekből képezett átlagokat ranggá alakítottuk: 5-ös átlag = 1-es rang, 4,5-es átlag = 2-es rang, 4-es átlag = 3-as rang, és így tovább, 2-es átlag = 7-es rang, további rang nincs, mert az évismétlésre kényszerült hallgatók már nem voltak az évfolyamon. Megjegyezzük, hogy logikából és magyarból nem volt vizsga.

Az induló adatok az alábbiak:

Kedveltségi fok	Logika	7	7	6	6	4	5	3	6	6	7	7	6	7	3	6	5	7
	Magyar	5	6	3	2	2	3	7	7	4	4	5	7	3	5	3	4	5
	Matematika	1	1	1	7	3	4	1	2	1	2	3	4	2	4	5	1	3
	Számítás-technika	6	3	5	4	6	6	2	5	5	5	2	3	6	6	4	6	6
	Kémia	4	2	7	5	7	7	4	1	3	6	6	2	5	1	7	7	2
	Mechanika	2	5	2	3	1	2	6	3	2	1	4	1	1	7	1	2	1
	Géprajz	3	4	4	1	5	1	5	4	7	3	1	5	4	2	2	3	4
Elért eredmény, mint rang	Matematika	1	2	2	7	7	7	4	7	5	7	7	7	6	6	7	3	3
	Számítás-technika	5	1	5	3	7	6	4	7	3	6	6	4	3	4	5	6	2
	Kémia	3	3	6	3	5	7	1	5	3	5	5	5	5	3	5	5	3
	Mechanika	2	3	5	3	6	6	6	7	4	6	6	5	7	5	3	6	4
	Géprajz	2	5	4	3	5	5	6	7	5	6	5	6	5	5	4	4	2

a) Hogyan alkalmazhatjuk az ismert matematikai statisztikai számításokat az oktatási kérdőívek vizsgálatára?

b) Az *átlagrangsor* a rangkorrelációt egy új mérőszámmal méri. Mindez nagy elemszámú sokaságnál számítógépes munkát igényel.

a) Készítsük el a statisztikához az adatmátrixot! A számítógép az adatokat egy $N=12$ sorból és $M=17$ oszlopból álló mátrixként kezeli. *A főiskolán használt TPA/i számítógépprogram a mátrix elemeit sorfolytonosan indexeli:*

$$\begin{array}{l} \text{1. logika}\\ \text{2. magyar}\\ \text{3. matemat.}\\ \cdots\cdots\cdots\\ \text{7. géprajz}\\ \hline \text{8. matemat.}\\ \cdots\cdots\cdots\\ \text{12. géprajz}\end{array}
\left[\begin{array}{l} x_1\ x_2\ x_3\ \ldots x_{17}\\ x_{18}\ x_{19}\ x_{20}\ \ldots x_{34}\\ x_{35}\ x_{36}\ x_{37}\ \ldots x_{51}\\ \cdots\cdots\cdots\\ x_{103}x_{104}x_{105}\ldots x_{119}\\ x_{120}x_{121}x_{122}\ldots x_{136}\\ \cdots\cdots\cdots\\ x_{187}x_{188}x_{189}\ldots x_{204}\end{array}\right]
\left.\begin{array}{l} \left.\vphantom{\begin{array}{l}a\\a\\a\\a\\a\end{array}}\right\}\text{kedveltségi rangok}\quad\Bigg\}\ N-5\text{ sor}\\ \left.\vphantom{\begin{array}{l}a\\a\\a\end{array}}\right\}\text{elért eredmények}\quad\Big\}\ 5\text{ sor}\end{array}\right\}N\text{ sor}$$

Az adatmátrix elemeinek beolvasását a folyamatábra *a)* része alapján végezzük el (N és M beolvasása, $X_K=X(K)$ értékek beolvasása).

b) Sorátlagok, szórások, relatív szórások. Az első sor elemeinek átlagát így kaphatjuk meg:

$$\bar{X}_1=\frac{\sum_{J=1}^{M} X_J}{M},$$

a második és a harmadik sor elemeinek átlaga:

$$\bar{X}_2=\frac{\sum_{J=1}^{M} X_{M+J}}{M},\quad \bar{X}_3=\frac{\sum_{J=1}^{M} X_{2M+J}}{M},$$

általában az *I-edik sor elemeinek átlaga:*

$$\bar{X}_I=\frac{\sum_{J=1}^{M} X_{(I-1)M+J}}{M}\qquad (I=1,2,3,\ldots,N).$$

Az I-edik sor elemeinek szórása a következő!

$$D_I=\sqrt{\frac{\sum_{J=1}^{M} X^2_{(I-1)M+J}}{M}-\bar{X}_I^2}\qquad (I=1,2,3,\ldots,N).$$

Mint tudjuk, ez a szórás az átlagtól való eltérést (az adatok átlag körüli „szóródását") jellemzi.

A relatív szórás

$$H_I=\frac{D_I}{\bar{X}_I}$$

megadja, hogy hányad része a szórás az átlagnak.

Az egyes sorokban levő adatok átlagának, szórásának és relatív szórásának kiszámítását a program a folyamatábra b) része alapján végzi. Ez egy ciklusba ágyazott másik ciklus. A belső ciklusban a program $J=1$-től 17-ig elvégzi az adatmátrix egy adott, I-edik sorában

- az X_I értékek összegezését, ezt tárolja az $S(I)$ tárolóban,
- az X_I^2 értékek összegezését, ezt tárolja az $SN(I)$ tárolóban,
- kiszámítja az I-edik sor elemeinek átlagát, $XA(I)$-t, szórását, $D(I)$-t, relatív szórását, $H(I)$-t a fenti képletek alapján.

Ezt az eljárást megismételjük a külső ciklus alapján, ahol az I index változik 1-től N-ig (1-től 12-ig).

c) A rangkorrelációs együttható,

$$R=1-\frac{6\sum_{J=1}^{M} D_J^2}{M^3-M}$$

az egyes tárgyakra adott kedveltségi rangok és a tárgyból elért eredmény közötti korrelációt méri,

$$-1\leqq R\leqq 1,\qquad 0\leqq|R|\leqq 1.$$

Minél közelebb van $|R|$ a 0-hoz, annál kevésbé korrelál egymással a két rangsor, míg minél közelebb van $|R|$ az 1-hez, annál erősebb a két rangsor korrelációja.

A képletben M a bírálók (oszlopok) számát jelenti, D_J pedig a szóban forgó tárgy rangjának és a belőle elért eredmény rangjának különbségét. Az adatmátrixban például a matematika rangok és eredmények különbségének négyzetösszege:

$$(X_{35}-X_{120})^2+(X_{36}-X_{121})^2+\ldots+(X_{51}-X_{136})^2=$$
$$=\sum_{J=1}^{M}(X_{2M+J}-X_{7M+J})^2.$$

Itt az indexben a $2M$ azért jön létre, mert logikából és magyarból csak rangsor van, vizsgajegy nincs, a matematika rangok előtt $2M$ mátrixelem van. Ugyanígy a matematikaeredmények előtt $7M$ mátrixelem van, így jön létre a $7M+J$ index ($J = 1, 2, \ldots, M$).

A rangok és elért eredmények különbségének négyzetösszege általában

$$\sum_{J=1}^{M} (X_{TM+J} - X_{(T+5)M+J})^2$$

alakban írható, ahol $T=2, 3, 4, 5, 6$.

A T indexhez tartozó $R(T)$ rangkorrelációs együttható értékének a számítását is kettős, egymásba ágyazott ciklussal kell megoldani, a belső ciklus indexe $J = 1, 2, \ldots, M$, a külső ciklus indexe $T=2, 3, 4, 5, 6$ (l. a folyamatábra *c)* részét). Ugyanebben a részben kiszámíttattunk még egy t-próbához szükséges $T(T)$ értéket is. Ugyanis tekinthetjük úgy, hogy az $M=17$ mérés csak egy kis elemszámú minta az egész sokaságból (a *teljes* hallgatóságból). Bizonyítható, hogy ha a rangkorrelációs együttható az egész sokaságban 0, akkor

$$T(T) = R\sqrt{\frac{M-2}{1-R^2}}$$

$M-2$ szabadságfokú, Student-féle, t-eloszlású változó. Így arra a feltevésre, hogy az egész sokaságban nincs semmiféle korreláció a két rangsorban, Student-próbát is végezhetünk.

d) A konkordancia („egyetértési") együttható:

$$W = \frac{12S}{m^2(n^3-n)}.$$

A változók jelentése: m a bírálók száma (jelen esetben $m=M=17$), n a rangsorolandó tárgyak száma (jelen esetben ez csak $n = N-5 = 12-5 = 7$, hiszen csak a 7 tantárgy rangsorolásában akarjuk vizsgálni a véleményadók válaszainak homogén voltát).

Ha azt várjuk, hogy egyáltalán nincs egyetértés a véleménynyilvánításban, akkor az ún. várt rangösszeg:

$$Z = \frac{m(n+1)}{2},$$

ez jelen esetben $\frac{M(N-4)}{2} = \frac{17 \cdot (12-4)}{2} = 68$. Az S jelenti az egy-egy tárgyban megfigyelt rangösszeg és az említett várt rangösszeg-eltéréseknek a négyzetösszegét (a mátrix egy-egy sorában a rangok összegéből ki kell vonni az $\frac{m(n+1)}{2}$ értéket, ezt az eltérést négyzetre kell emelni, a kapott n db négyzetszámot össze kell adni).

A W konkordancia- („egyetértési") együttható,

$$0 \leq W \leq 1,$$

minél közelebb van W a 0-hoz, annál kevésbé van egymással korrelációban a bírálók véleménye, minél közelebb van W az 1-hez, annál tökéletesebben egyezik a bírálók véleménye (egyeznek a rangszámok).

Itt is tekinthetjük úgy, hogy az $M=17$ hallgató válasza csak egy minta az egész sokaságból (a teljes hallgatóságból). Feltehető-e, hogy a mintából kapott W 0-tól való eltérése ellenére *az egész sokaságban a konkordancia-együttható 0?*

Bizonyítható, hogy ez esetben az

$$F = \frac{(m-1)W_{\text{korr}}}{1-W_{\text{korr}}}$$

F-eloszlású változó, $n-1-\frac{2}{m}$ és $(m-1)\left(n-1-\frac{2}{m}\right)$ szabadságfokkal. (Az első érték a nagyobb szórásnégyzetű becslés szabadságfokának, a második érték a kisebb szórásnégyzetű becslés szabadságfokának felel meg az F értékek táblázatának szokásos használatakor. Szükség esetén lineárisan interpolálni lehet.) Így arra a feltevésre, hogy a mintából kapott W érték ellenére *az egész sokaságban* a konkordancia-együttható 0 volna, F próbát végezhetünk.

Itt a konkordancia-együttható korrigált értéke

$$W_{\text{korr}} = \frac{S-1}{\frac{m^2(n^3-n)}{12}+2}.$$

Nézzük a folyamatábra *d)* részét!

Mivel az adatmátrix egy-egy sorában álló rangösszegeket a program már kiszámította (l. az *a)* részben az $S(I)$ értékeket), ezért az $S(I)$ értékek (a megfigyelt rangösszegek) és a $Z = \frac{M(N-4)}{2}$ (a várt rangösszeg) eltérésének négyzetösszegét egy S tárban összegezhetjük ($I = 1, 2, 3, \ldots, N-5$). Ebből kiszámítjuk W-t, a korrigált W-t (WK-t), F-et.

A kiiratással a program végetér:

TARGYAK	X AATLAG	SZORAAS	REL. SZORAAS
LOG	5.7647	1.3074	0.2268
MAGY	4.4118	1.6110	0.3651
MAT	2.6471	1.6783	0.6340
SZ. T	4.7059	1.4044	0.2984
KEM	4.4706	2.1994	0.4920
MECH	2.5882	1.8169	0.7020
GEPR	3.4118	1.6109	0.4722
MAT. J	5.1765	2.1209	0.4097
SZ. T. J.	4.5294	1.6845	0.3719
KEM. J.	4.2353	1.4361	0.3391
MECH. J.	4.9412	1.4741	0.2983
GEPR. J.	4.6471	1.3258	0.2853

KONKORDANCIA EGYUTTHATO: W=0.2860
F-ERTEK: F= 6.5892

TARGYAK=	RANGKORR.	EGYUTTHATOK	T ERTEKEK
MAT	R=	0.8100	T=5.3505
SZ. T	R=	0.9179	T=8.9584
KEM	R=	0.9265	T=9.5338
MECH	R=	0.7696	T=4.6681
GEPR	R=	0.9007	T=8.0313

A listában található:

– az egyes tárgyak rangsorának, ill. a belőle elért eredménynek az átlaga, szórása, relatív szórása;

– a konkordancia-együttható, W, a korrigált konkordancia együttható, WK, az F-próbához szükséges, adatokból számított F érték,

– az egyes tárgyakra adott rangok és a belőlük elért eredmények közötti rangkorrelációs együtthatók, az R értékek, ill. a hozzájuk tartozó, t-próbához szükséges T értékek (ne feledjük el, hogy a tárgyak sorrendje soronként; matematika, számítástechnika, kémia, mechanika, géprajz).

A futtatási eredmények értékelése:

I. Az *átlagok alapján a tárgyak kedveltségi sorrendje* (a legkedveltebbtől a kevésbé kedvelt felé haladva): mechanika, matematika, géprajz, magyar, kémia, számítástechnika, logika. Ez az *átlagos* kedveltségi sorrend, az egyedi kedveltségi sorrendek ettől jelentős eltérést mutatnak (a relatív szórás például a mechanika esetében 77%, a matematika esetében 70%).

A tantárgyakból elért átlag eredmény alapján a sorrend (a legjobb, azaz legkisebb rangú tárgytól a kevésbé jó, magasabb ranggal rendelkező tárgy felé haladva): kémia, számítástechnika, géprajz, mechanika, matematika. A relatív szórások itt is elég nagyok, de kisebbek, mint a kedveltségi rangoknál.

II. *A konkordancia-együttható,* $W=0{,}2860$, gyenge egyetértésre utal a kedveltségi rangsorolásban.

Az adatokból számított $F=6{,}5892$. Ezt kell összehasonlítani az $n-1-\frac{2}{m}=6-\frac{2}{17}=5{,}88$ és $(m-1)\left(n-1-\frac{2}{m}\right)=16\left(6-\frac{2}{17}\right)=94{,}12$ szabadságfokú, táblázatbeli F értékkel:

f_2	f_1: 5	f_1: 6
60	2,37	2,25
100	2,30	2,19

A számított F érték (6,5892) biztosan nagyobb, mint a táblázatbeli, ezért azt a feltevést, hogy az *egész sokaságban a konkordancia-együttható 0 volna (a bírálók egyáltalán nem értenek egyet), elvetjük* 95%-os biztonsági szinten (a táblázat erre a szintre készült).

III. *A tárgyak rangsora és az elért eredmények* erős korrelációban vannak, erre utalnak a 0,7696-tól 0,9265-ig terjedő *rangkorrelációs együtthatók.* 3 rangkorrelációs együttható is 0,90 fölött van.

Lehet-e, hogy az *egész* (általunk nem vizsgált) *sokaságban* a rangkorrelációs együttható mégis 0 volna (nem korrelálnak a rangsorok és az elért eredmények)? Erre t-próbával adhatunk választ.

A számított t értékek rendre nagyobbak, mint a táblázatbeli, $m-2=15=f$ szabadságfokhoz tartozó t érték:

f	95%	99,9%
15	2,131	4,073

Ezért 99,9% biztonsági szinten *elvetjük azt a feltevést, hogy az egész sokaságban nem korrelál egymással a kedveltségi rangsor és a tárgyból elért eredmények rangsora.*

Rangkorreláció az átlagrangsorral:

Vezessük be a „bírálók" által adott rangsorok egyezésének vizsgálatára a következőt! Számítsuk ki *az egyes bírálók által adott rangsor és az egész évfolyam véleményéből kialakított átlagrangsor közötti rangkorrelációs együtthatókat! Az r_i rangkorrelációs együtthatók abszolút értékének átlaga,*

$$\bar{R} = \frac{\sum_{i=1}^{m} |r_i|}{m}$$

nyilván 0 és 1 közötti szám, hiszen minden $|r_i|$ 0 és 1 közé esik ($i = 1, 2, \ldots, m = 17$). $\bar{R} = 0$ akkor és csak akkor, ha minden $|r_i| = 0$, tehát nincs korreláció a bírálók által adott rangsorok és az átlagrangsor között. $\bar{R} = 1$ akkor és csak akkor, ha minden $|r_i| = 1$, vagyis mindenki az átlagrangsorral egyező vagy tökéletesen ellentétes rangsorolást adott.

Minél közelebb van $\bar{R}$ az 1-hez, annál erősebb az egyezés az átlagrangsorral. Minél közelebb van $\bar{R}$ a 0-hoz, annál gyengébb. $\bar{R}$ tehát tekinthető az átlagrangsorral való egyezés mérőszámának.

A szükséges számításokat tartalmazza a mellékelt táblázat. A rangkorrelációs együtthatók közül csak kettő negatív (két bíráló véleménye enyhén ellenkezik az átlagrangsorral), a többi pozitív. (A vélemények kisebb-nagyobb mértékben egyeznek az átlagrangsorral). A rangkorrelációs együtthatók abszolút értékei 0,2 és 1 között majdnem egyenletesen oszlanak meg, gyenge–közepes–erős korreláció az átlagrangsorral egyaránt előfordul. A rangkorrelációs együtthatók abszolút értékének az átlaga,

$$\bar{R} = 0{,}5966$$

közepes korrelációra utal. Mivel a rangkorrelációs együttható értékek szórása 0,2537, ez $\bar{R}$-hoz viszonyítva 43%-os relatív szórást jelent. Ez is arra utal, hogy az átlag, ($\bar{R}$) körül erősen szóródnak a kapott rangkorrelációs együtthatók abszolút értékei.

Tantárgyak	Átlag rangsor	Rangkülönbségek és négyzeteik																
		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
Logika	7	0 0	0 0	−1 1	−1 1	−3 9	−2 4	−4 16	−1 1	−1 1	0 0	0 0	−1 1	0 0	−4 16	−1 1	−2 4	0 0
Magyar	4	1 1	2 4	−1 1	−2 4	−2 4	−1 1	3 9	3 9	0 0	0 0	1 1	3 9	−1 1	1 1	−1 1	0 0	1 1
Matematika	2	1 1	−1 1	−1 1	5 25	1 1	2 4	−1 1	0 0	−1 1	0 0	1 1	2 4	0 0	2 4	3 9	−1 1	1 1
Számítástechnika	6	0 0	−3 9	−1 1	−2 4	0 0	0 0	−4 16	−1 1	−1 1	−1 1	−4 16	−3 9	0 0	0 0	−2 4	0 0	0 0
Kémia	5	1 1	−3 9	2 4	0 0	2 4	2 4	−1 1	−4 16	−2 4	1 1	1 1	−3 9	0 0	−4 16	2 4	2 4	−3 9
Mechnika	1	1 1	4 16	1 1	2 4	0 0	1 1	5 25	2 4	1 1	0 0	3 9	0 0	0 0	6 36	0 0	1 1	0 0
Géprajz	3	0 0	1 1	1 1	−2 4	2 4	−2 4	2 4	1 1	4 16	0 0	−2 4	2 4	1 1	−1 1	−1 1	0 0	1 1
A rangkülönbségek négyzetösszege		4	40	10	42	22	18	72	32	24	2	32	36	2	74	20	10	12
Rangkorrelációs együttható, r_i		0,9286	0,2857	0,8214	0,2500	0,6071	0,6786	−0,2857	0,4286	0,5714	0,9643	0,4286	0,3571	0,9643	−0,3214	0,6429	0,8214	0,7857

A mellékelt számításokat az előzőekben bemutatott program lefuttatása után, külön végeztük el.

Feladatként marad a folyamatábrát ezzel a kiegészítéssel bővíteni (itt az adatmátrix oszlopait kell kezeltetni a számítógéppel).

33. A következő táblázat az egyes munkások teljesítménybeli rangsorait mutatja a kontrollálandó és az új, kísérleti technológiák alkalmazása esetén. (A teljesítményt az egy műszak alatt gyártott munkadarabok számával mérték.)

Munkás	Technológia				
	kontroll	I.	II.	III.	IV.
1	1	2	5	4	3
2	5	3	4	2	1
3	1	5	3	4	2
4	2	4	5	1	3
5	4	5	1	2	3
Rangösszegek:	13	19	18	13	12

Feltehető-e, hogy az új technológiák nem hoznak javulást vagy romlást a teljesítményben?

Nullhipotézisünk ellenőrzésére χ^2-próbát végzünk.

$$\chi^2 = \frac{12}{np(p+1)}\sum_{i=1}^{p} S^2p - 3n(p+1) = \frac{12}{5\cdot 5\cdot 6}(13^2+19^2+18^2+$$

$$+13^2+12^2)-3\cdot 5\cdot 6 = 3{,}36<\chi^2_{95} = 9{,}5.$$

A nullhipotézist elfogadjuk, az új technológiáknak nincs kellő hatása a teljesítményre (sem nem javítanak, sem nem rontanak). A szabadságfokok száma $p-1 = 4$ volt.

Megjegyzés

a) Ha feltehető volna az új technológiák szignifikáns hatása, akkor érdemes lenne az egyes technológiákat külön-külön összehasonlítani egymással vagy a kontrollálandó technológiával.

b) Ha a rangösszegeket nem „függőleges”, oszloponként történő összeadással készítjük el, hanem „vízszintesen”, soronként összegezünk, n-t és p-t felcseréljük, akkor arra kapnánk választ, hogy van-e szignifikáns eltérés az egyes munkások termelékenysége közt.

4. NÉHÁNY GYAKORLATI ALKALMAZÁS

1. *Selejtarány-ellenőrző kártya.* Egy gyártási folyamat során a selejtarány a $p=0{,}02$ selejtvalószínűség körül ingadozik. Ezt a selejtvalószínűséget elfogadjuk. Ennek ellenőrzésére többször $n=50$-elemű mintát veszünk visszatevéssel. Maximálisan hány selejtes esetén fogadjuk el a $p=0{,}02$ feltevést?

Annak a valószínűsége, hogy legfeljebb k_{max} selejtes van a mintában,

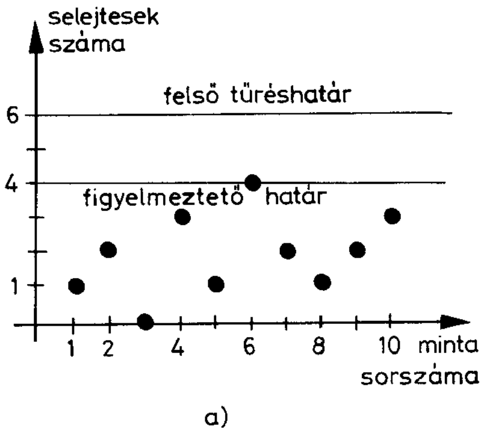
$$P(X\leq k_{max}) = \sum_{k=0}^{k_{max}}\binom{50}{k}0{,}02^k\cdot 0{,}98^{50-k} \approx \sum_{k=0}^{k_{max}}\frac{1^k e^{-1}}{k!}$$

(felhasználtuk, hogy $\lambda=np = 50\cdot 0{,}02 = 1$). Ennek a valószínűségnek az értékét a 4. táblázatból tudjuk kiolvasni:

$\lambda=np$	k_{max}							
	0	1	2	3	4	5	6	7
1,00	368	736	920	981	996	999	1000	

Közel $\frac{1000}{1000} = 1$ $\left(\text{pontosabban: } \frac{999{,}9}{1000}\right)$ annak a valószínűsége, hogy $p=0{,}02$ esetén maximum 6 selejtes lesz a gyártmányok között $n=50$-elemű mintában. Kb. 0,996 annak a valószínűsége, hogy maximum 4 selejtes lesz az $n=50$-elemű mintában. Ha 6-nál több selejtes van, igen nagy, 0,9999 a valószínűsége, hogy a selejtarány megnőtt. Ha ezt állítjuk, 10 000 eset közül átlagosan 1-szer tévedünk, 9999-szer jól döntünk.

Ennek segítségével selejtarány-ellenőrző kártyát tervezhetünk a $p=0{,}02$ selejtvalószínűség ellenőrzésére. Ha 6-nál több a selejtesek száma az $n=50$-elemű mintában, elvetjük a $p=0{,}02$ feltevést, úgy tekintjük, hogy a selejtarány megnőtt. A gépet újból beállítjuk. 4 és 6 között fokozottabban figyeljük a gyártást. 4 alatt elfogadjuk a $p=0{,}02$ feltevést. A 6 a felső tűréshatár, a 4 a figyelmeztető felső határ (53*a* ábra). Ha a pontok a megfelelő tartományba esnek, a folyamat statisztikailag kontrollált.



53*a* ábra

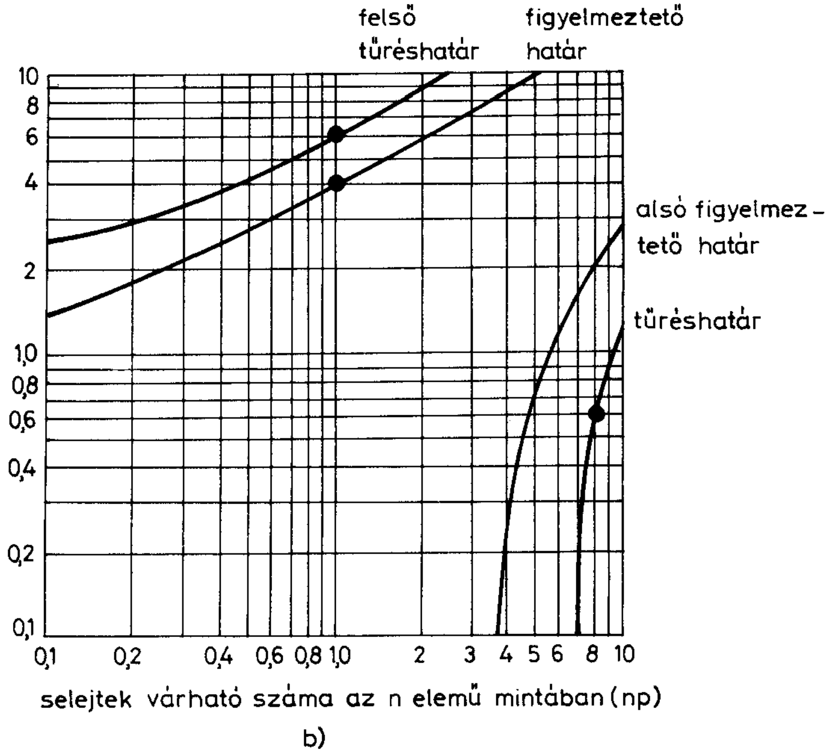
Néha az is előfordul, hogy nemcsak felső, hanem alsó tűréshatárt is szerkesztenek, hiszen az a gyártás lényeges javulására utalhat.

A különböző *np* értékekhez különböző tűréshatár-értékek tartoznak. Ezek a tűréshatárértékek (azonos biztonsági szint esetén) egy görbén vannak. A 0,999, ill. 0,975-ös valószínűségi szinten megszerkesztettük a felső tűréshatár- és felső figyelmeztetőhatár-görbéket. (A biztonsági szint apró eltéréseit a Poisson-eloszlás közelítő volta okozza.) Alsó figyelmeztető-, ill. tűréshatárgörbét is berajzoltunk (53*b* ábra).

Megemlítjük, hogy nemcsak a selejtarányra, hanem a mintaterjedelemre, átlagra, névleges értéktől való eltérésre stb. is szerkesztenek ellenőrzőkártyákat.

Az eljárás fogyatékossága: csak az elfogadott selejtvalószínűséget figyeli, a megtűrt selejtvalószínűséget nem (lásd a mintavételi tervek készítését).

2. Egy gyártási folyamatnál a megengedett selejtvalószínűség $p=0{,}08$. Többször $n=100$-elemű mintát véve visszatevéssel, maximum 1 selejteset találunk. Állapítsuk meg az 53*b* ábra segítségével, elfogadható-e, hogy a selejtarány maximum 8% körül ingadozik!



53*b* ábra

Elfogadható, az 1 selejt $np=8$ esetén az alsó tűréshatár és az alsó figyelmeztetőhatár közé esik. Tehát még az is lehet, hogy $p<0{,}08$.

3. Tekintsük selejtszámnak a kedvenc csapatunk által kapott gólok számát! A csapat mérkőzésenként várható gólszáma 0,2. Olvassuk le az 53*b* ábráról, hogyan minősítse a vezető, ha a csapat egy mérkőzésen *a)* 1, *b)* 2, *c)* 3 gólt kap!

$np=0{,}2$. *a)* Elfogadható, *b)* fokozottan figyelni kell, *c)* nem fogadható el, romlott csapatunk kitűnő felkészültsége.

4. *Egyszeres mintavételi terv. a)* Az átadó fél azt állítja, hogy a szállítmányban a selejt valószínűsége, $p_1 = 0,05$, a legfeljebb 5% selejtet tartalmazó szállítmányt az átvevő is elfogadja. $n = 20$-elemű (visszatevéses) mintát választva mennyi annak a valószínűsége, hogy abban a selejtes darabok száma, X legfeljebb 4 lesz?

Számítsuk ki először annak a valószínűségét, hogy k selejtes lesz a szállítmányban!

$$P(X=k) = \binom{20}{k} 0{,}05^k \cdot 0{,}95^{20-k}.$$

A „legfeljebb 4 selejtes" valószínűsége az egyes esetek valószínűségének az összege:

$$P(X \leqq 4) = \sum_{k=0}^{4} \binom{20}{k} 0{,}05^k \cdot 0{,}95^{20-k}.$$

Ennek értékét kiolvashatjuk a 2. táblázatból.

p	k_{max}	$n=20$
0,05	3	0,9841
	4	0,9974 = = P($X \leqq 4$)

Igen nagy valószínűségű tehát, hogy $p_1 = 0,05$ selejtvalószínűség esetén az $n = 20$-elemű visszatevéses mintában a selejtesek száma legfeljebb $k_{max} = 4$ lesz.

Ha tehát legfeljebb 4 (azaz 0 vagy 1, 2, 3 vagy 4) selejtes darab esetén átveszem a tételt, ennél többet találva visszautasítom, akkor több, mint 0,99 valószínűséggel (99%-os biztonsággal) feltehetem, hogy valóban csak a megengedett (5%) selejtarányú tételt veszem át, a többit elutasítom.

Ha a selejtvalószínűség még kisebb, mint 0,05, akkor a fentinél még nagyobb valószínűségű, hogy 20-ból legfeljebb 4 selejtes lesz (l. a 2. táblázatot).

A 99,75% a biztonsági szint – tehát átlagosan az esetek 0,25%-ában téved az átadó. Előfordulhat tehát, hogy megfelelő tételt visszadob a mintavételi eljárás (elsőfajú hiba), de csak nagyon kicsi, 25 : 10 000 valószínűséggel.

Ez így nagyszerű mintavételi tervnek látszik. Rögtön belátjuk, hogy sajnos nem az.

b) Az átvevő fél azt szeretné, hogy $p_2 = 0,10$ megtűrt selejtvalószínűségű (vagy annál nagyobb selejtvalószínűségű) tételt a mintavételi terv alapján csak nagyon kicsi valószínűséggel vegyen át.

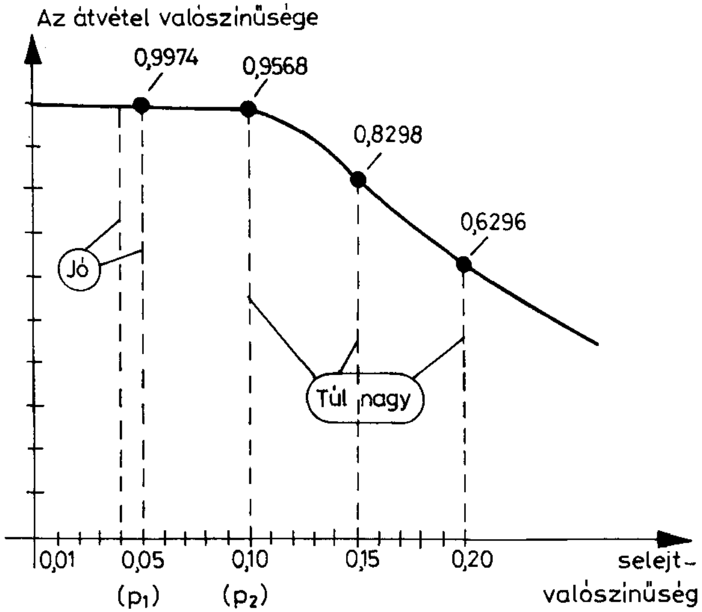
Nézzük meg mindkét táblázatot!

I. táblázat				II. táblázat		
n	k	$p=0{,}10$	összesen	p	k_{max}	$n=20$
20	0	0,1216		0,08	4	0,9817
	1	0,2702		0,09	4	0,9710
	2	0,2852		0,10	4	0,9568 = = $P(X \leqq 4)$
	3	0,1901	0,8671	0,15	4	0,8298
	4	0,0898	0,9569 = = $P(X \leqq 4)$	0,20	4	0,6296

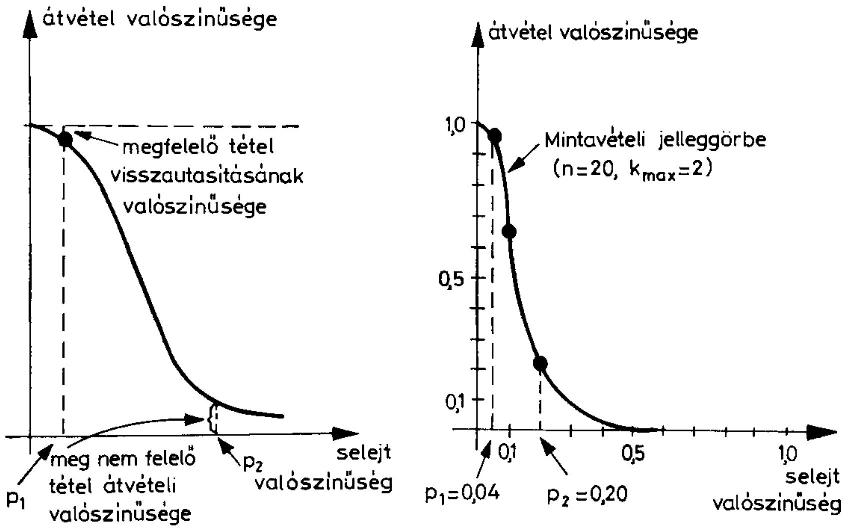
Az esetek átlagosan 95,7%-ában tehát átvesszük a 10% selejtet tartalmazó tételt, a meg nem felelő tétel átvétele (az ún. másodfajú hiba) túl nagy valószínűségű. (A 15%-ot tartalmazó tételt is kb. 83% valószínűséggel átveszi a terv.) Az 54. ábra illusztrálja, hogy miben jó és miben rossz a mintavételi terv (ún. mintavételi jelleggörbét szerkesztettünk).

c) Milyen legyen a jó mintavételi terv (és a mintavételi jelleggörbe)?

Nagy valószínűséggel engedje át a jó (a megengedett selejtvalószínűségű) tételt, kis valószínűséggel engedje át a rossz (a megtűrt selejtvalószínűségű) tételt.



54. ábra



a) b)

55*ab* ábra

Az 55*a* ábra mutatja a jó mintavételi tervnek megfelelő jó mintavételi jelleggörbe alakját. Az 55*b* ábrán felrajzoltuk az $n=20$, $k_{max}=2$ mintavételi tervnek megfelelő jelleggörbét. Ez a mintavételi terv azt jelenti, hogy $n=20$-elemű (visszatevéses) mintát veszünk, 0, 1 vagy 2 selejtes esetén átvesszük a tételt, 2-nél több esetén visszautasítjuk. Az alábbi táblázat mutatja, hogy mekkora kockázatokat jelent ez $p_1=0{,}04$ megengedett és $p_2=0{,}20$ megtűrt selejtarány esetén:

p selejt-valószínűség	Átvételi valószínűség ($n=20$, $k_{max}=2$ esetén	Visszautasítás valószínűsége
0	1,000	
0,02	0,993	
$p_1=0{,}04$	0,956	0,044 = átadó, k_1 kockázata (elsőfajú hiba)
0,06	0,885	
0,08	0,788	
0,10	0,677	
0,15	0,405	
$p_2=0{,}20$	0,206 = átvevő, k_2 kockázata (másodfajú hiba)	
0,25	0,091	
0,30	0,036	
1,00	0,000	

A mintavétel képletekkel leírva a következőt jelenti:

– Nagy valószínűséggel átvesszük a tételt, ha a selejtes darabok száma $X \leqq k_{max}$; megfelelő tétel átvételi valószínűsége:

$$P(X \leqq k_{max}) = \sum_{k=0}^{k_{max}} \binom{n}{k} p_1^k (1-p_1)^{n-k} = \text{1-hez közeli érték} = 1-k_1;$$

– kis valószínűséggel vesszük át a tételt, ha az meg nem felelő (selejtvalószínűség $\geqq p_2$); meg nem felelő tétel átvételi valószínűsége:

$$\sum_{k=0}^{k_{max}} \binom{n}{k} p_2^k (1-p_2)^{n-k} = \text{0-hoz közeli érték} = k_2.$$

d) Az $n=20$, $k_{max}=2$ mintavételi terv tehát kis valószínűséggel engedi át a meg nem felelő tételt ($1-0{,}956=0{,}044$ valószínűséggel), 0,206 valószínűséggel engedi át a megtűrt ($p_2=0{,}20$ selejtvalószínűségű) tételt. Ha ez utóbbi nem felel meg az átvevőnek, akkor tovább kell javítani a mintavételi terven.

Néhány javított mintavételi terv:

Sorszám	A minta elemszáma, n	A selejtek maximális száma, k_{max}	Elfogadott selejtvalószínűség, p_1	Jó tétel átvételének a valószínűsége	Megtűrt selejtvalószínűség, p_2	Rossz tétel átvételének a valószínűsége	A mintavételi terv értékelése
1.	25	1	0,02	0,9114	0,15	0,0931	Elfogadható terv, ha $p_2=0{,}15$ megengedhető
2.	100	10	0,05	0,986	0,10	0,583	Lejjebb szorítottuk p_2-t, de ennek árán n megnőtt, a rossz tétel átvételi valószínűsége túl nagy. Ha ez baj, a terv nem fogadható el
3.	200	10	0,03	0,957	0,10	0,011	n további növelése árán leszorítottuk a rossz tétel átvételi valószínűségét. Ha a mintavétel nem nagyon költséges, a terv jó
4.	200	10	0,03	0,957	0,08	0,077	Az előbbi terv módosítása: p_2 fokozatos leszorításával a rossz tétel átvételi valószínűsége (sajnos) nő
5.	200	10	0,03	0,957	0,07	0,176	

e) Érdekesebb úgy kialakítani mintavételi tervet, hogy az átadó és az átvevő fél megegyezik:

- az elfogadott selejtvalószínűségben, $p_1=0{,}01$;
- a megtűrt selejtvalószínűségben, $p_2=0{,}04$;
- a megfelelő tétel átvételi valószínűségében, legyen ez pl. 95%;
- a rossz tétel átvételi valószínűségében, legyen ez pl. 5%.

A fentiek biztosítására milyen elemszámú mintát vegyünk; maximálisan hány selejtes termék esetén vegyük át a szállítmányt?

$$\sum_{k=0}^{k_{max}} \binom{n}{k} 0{,}01^k \cdot 0{,}99^{n-k} = 0{,}95,$$

$$\sum_{k=0}^{k_{max}} \binom{n}{k} 0{,}04^k \cdot 0{,}96^{n-k} = 0{,}05.$$

Az egyenletrendszerben n és k_{max} az ismeretlenek. Táblázatok segítségével vagy számítógép-program segítségével egy megoldás: $n=261$, $k_{max}=5$.

A terv nagy biztonságot nyújtó, szigorú mintavételi terv. Ennek fejében nagy (261) elemszámú minta megvizsgálását követeli meg.

f) Nemcsak a hibás darabok, hanem a hibás méretek, kilogrammok száma stb. is lehet minősítő határ.

5. *Többszörös mintavételi terv. a)* Az egyszeres mintavételi tervet már ismertettük. Ott egyetlen próbacsoport alapján kell dönteni, hogy a tétel megfelelő-e vagy sem. Láttuk, hogy a megvizsgálandó minta elemszáma túl nagy lehet.

b) Ezen segít a kétszeres mintavételi terv. Ott kisebb első csoportot kell venni, mint az egyszereshez, és második próbacsoportot nem kell minden esetben kivenni. A döntéshez szükséges mintaelemszám általában kisebb, mint az egyszeres próbánál.

Bemutatunk egy ilyen kétlépcsős próbát.

Visszatevéssel $n_1=20$-elemű mintát veszünk. Legyen az első mintában a selejtes darabok száma k_1.

Ha $k_1 \leqq c_{10}=1$, a tételt átvesszük.

Ha $c_{10}<k_1<c_{11}=5$, újabb mintát veszünk.

Ha $k_1 \geqq c_{11}=5$, a tételt elutasítjuk.

Ha újabb mintát kell vennünk, akkor $n_2=25$ db-ot kiválasztva vizsgáljuk a selejtes darabok számát, k_2-t.

Ha $k_1+k_2 \leqq c_{20} = 6$, a tételt átvesszük.

Ha $k_1+k_2 \geqq c_{20}+1 = c_{21} = 7$, a tételt elutasítjuk.

Az első mintánál tehát vagy döntünk, vagy elodázzuk a döntést. A második minta vétele mindenképpen döntésre vezet.

Nézzük meg, hogy mennyi egy tétel átvételének a valószínűsége:

$$P[k_1 \leqq 1 \text{ vagy } (1 < k_1 < 5 \text{ és } k_1 + k_2 \leqq 6)] =$$

$$= P(A+BC) = P(A) + P(BC) =$$

$$= \sum_{k_1=0}^{1} \binom{n_1}{k_1} p^{k_1}(1-p)^{n_1-k_1} +$$

$$+ \sum_{k_1=2}^{4} \left[\binom{n_1}{k_1} p^{k_1}(1-p)^{n_1-k_1} \underbrace{\sum_{k_2=0}^{6-k_1}}_{k_1+k_2 \leqq 6} \binom{n_2}{k_2} p^{k_2}(1-p)^{n_2-k_2} \right].$$

A második szummához gondoljuk meg, hogy $k_1 = 2$ vagy 3, vagy 4, adott k_1 értékhez k_2-t úgy kell megválasztani, hogy $k_1 + k_2 \leqq 6$ legyen, ha pl.

$k_1 = 2$, akkor $k_2 = 0$ vagy 1, vagy 2, vagy 3, vagy 4;
$k_1 = 3$, akkor $k_2 = 0$ vagy 1, vagy 2, vagy 3;
$k_1 = 4$, akkor $k_2 = 0$ vagy 1, vagy 2.

Az egymást kizáró események összegének valószínűségei viszont az egyes valószínűségek összegéből számíthatók ki, a B és C együttes bekövetkezésének valószínűsége a függetlenség miatt az egyes valószínűségek szorzata.

	I. táblázat $n_1=20$	II. táblázat $n_2=25$	II. táblázat $n_1=20$
$p_2=0{,}20$ k_1 $6-k_1$	$\binom{n_1}{k_1} p^{k_1}(1-p)^{n_1-k_1}$	$\sum_{k_2=0}^{6-k_1} \binom{n_2}{k_2} p^{k_2} \times$ $\times (1-p)^{n_2-k_2}$	$P(A)=0{,}0692$ + szorzatok
2 | 4	0,1369	0,4207	0,0576
3 | 3	0,2054	0,2340	0,0481
4 | 2	0,2182	0,0982	0,0214
		összesen:	0,1963
$p_1=0{,}05$			$P(A)=0{,}7358$ + szorzatok
2 | 4	0,1887	0,9928	0,1873
3 | 3	0,0596	0,9659	0,0576
4 | 2	0,0133	0,8729	0,0116
		összesen:	0,9923

A keresett valószínűségeket az I. és a II. táblázatból lehet megkapni.

Legyen az elfogadott selejtvalószínűség: $p_1 = 0{,}05$, a megtűrt: $p_2 = 0{,}20$. Nézzük meg, hogy a kétlépcsős terv mekkora valószínűséggel engedi át a jó és a rossz tételeket!

$n_1 = 20$, $c_{10} = 1$, $c_{11} = 5$,
$n_2 = 25$, $c_{20} = 6$, $c_{21} = c_{20} + 1 = 7$,
$p_1 = 0{,}05$ $p_2 = 0{,}20$

A kétlépcsős terv 0,1963 valószínűséggel bocsátja át a megtűrt selejtvalószínűségű rossz tételt, 0,9923 valószínűséggel a megengedett selejtvalószínűségű jó tételt. Ha ezek az adatok elfogadhatóak, dolgozzunk ezzel a tervvel!

c) Ugyanúgy alkalmazhatunk többszörös mintavételt is. Ahányszoros a mintavételi terv, annyiadik vizsgálatnál feltétlen döntéshez kell jutnunk. A számolás helyett azonban nézze meg az Olvasó az MSZ 247/1–75 „Tömegcikkek matematikai statisztikai minősítése, Próbavételi tervek" c. szabványt. A szabvány használatáról csak ízelítőt adunk. A szabvány tartalmaz egyszeres, kétszeres és többszörös mintavételi terveket.

A feladatot normális próbavételi tervvel kell kezdeni, a 10. oldalon meg van adva, hogy mikor kell áttérni szigorított tervre vagy enyhítettre.

A szabványban az M1 táblázatból megállapítandó a kulcsjel, ez a vizsgálatnak még egy szigorúságát szabályozza. Az M1 táblázat II. fokozatú általános vizsgálati szintjéhez tartozó táblázatrészlet itt látható:

Tételnagyság	Kulcsjel II. fokozatú általános vizsgálati szint esetén
– 25	C
26– 50	D
51– 90	E
91– 150	F
151– 280	G
281– 500	H
501– 1 200	J
1 201– 3 200	K
3 201– 10 000	L
10 001– 35 000	M
35 001–150 000	N
150 001–500 000	P
500 000 felett	Q

Mellékelten közöljük a szabvány M6 táblázatát, amely kétszeres tervtípusokat tartalmaz normál vizsgálatra. (A névleges hibaszázalék legnagyobb értéke 10, míg a száz egységre eső hiba 0,010-től 1000-ig terjed.)

Kulcsjel	Próbamennyiség kiveendő	Próbamennyiség együttesen	Névleges hibaszázalék 0,010	0,015	0,025	0,040	0,065	0,10	0,15	0,25	0,40	0,65	1,0
	db	db	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n
A			↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
B	2 2	2 4	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
C	3 3	3 6	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
D	5 5	5 10	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
E	8 8	8 16	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	*
F	13 13	13 26	↓	↓	↓	↓	↓	↓	↓	↓	↓	*	△
G	20 20	20 40	↓	↓	↓	↓	↓	↓	↓	↓	*	△	▽
H	32 32	32 64	↓	↓	↓	↓	↓	↓	↓	*	△	▽	0 2 1 2
J	50 50	50 100	↓	↓	↓	↓	↓	↓	*	△	▽	0 2 1 2	0 3 3 4
K	80 80	80 160	↓	↓	↓	↓	↓	*	△	▽	0 2 1 2	0 3 3 4	1 4 4 5
L	125 125	125 250	↓	↓	↓	↓	*	△	▽	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7
M	200 200	200 400	↓	↓	↓	*	△	▽	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9
N	315 315	315 630	↓	↓	*	△	▽	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13
P	500 500	500 1000	↓	*	△	▽	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19
Q	800 800	800 1600	*	↑	▽	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27
R	1250 1250	1250 2500	△	↑	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	△

▽ a nyíl alatt legelőször található tervet kell alkalmazni. Ha a próbamennyiség meghaladja a tétel nagyságát, akkor százalékos vizsgálatra van szükség. △ a nyíl felett legelőször található tervet kell alkalmazni. *m* a megfelelő darabok száma. *n* a nem megfelelő darabok száma. * egyszeres tervtípust kell alkalmazni (vagy alább levő kétszeres tervtípust, ahol az rendelkezésre áll).

Kulcsjel	(normális vizsgálatnál) 1,5	2,5	4,0	6,5	10	15	25	40	65	100	150	250	400	650	1000
	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n
A	↓	↓	▽	*	↓	▽	*	*	*	*	*	*	*	*	*
B	↓	↓	*	△	↓	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	17 22 37 38	25 31 56 57
C	↓	*	△	▽	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	17 22 37 38	25 31 56 57	↑
D	*	△	▽	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	17 22 37 38	25 31 56 57	↑	↑
E	△	▽	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	17 22 37 38	25 31 56 57	↑	↑	↑
F	▽	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	↑	↑	↑	↑	↑	↑
G	0 2 1 2	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	↑	↑	↑	↑	↑	↑	↑
H	0 3 3 4	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	↑	↑	↑	↑	↑	↑	↑	↑
J	1 4 4 5	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑
K	2 5 6 7	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
L	3 7 8 9	5 9 12 13	7 11 18 19	11 16 26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
M	5 9 12 13	7 11 18 19	11 16 26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
N	7 11 18 19	11 16 26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
P	11 16 26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
Q	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
R	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑

6. Egy árucikk ellenőrzésénél általános vizsgálati szint II. fokozatú szigorúsággal akarunk eljárni. Egy tétel 200 db-ot tartalmaz. Elfogadott selejtvalószínűség: $p_1 = 10\%$ $(=0{,}10)$. Készítsünk kétszeres mintavételi tervet!

A II. fokozatnak, a 151–280 közti tételnagyságnak az M1 táblázatban G kulcsjel felel meg. Az M6 táblázat megfelelő részlete:

Kulcsjel	Kiveendő	Névleges hibaszázalék $p_1=0{,}10$	
		m	n
G	20	3	7
	20	8	9

Ebből látható, hogy a kiveendő próbamennyiség az 1. mintánál $n_1=20$, a másodiknál $n_2=20$. Ha az első mintában maximálisan 3 selejteset találunk, a tételt átvesszük, ha 4, 5, 6 darabot, új mintát veszünk, ha 7 vagy több a selejt, a tételt visszautasítjuk. Ha a második mintában és az elsőben együtt talált selejtes darabok száma legfeljebb 8, a tételt átvesszük, 9 vagy több selejt esetén visszautasítjuk.

A szabványban megtalálhatók az adott kulcsjelhez tartozó jelleggörbék, alattuk a jelleggörbék néhány pontjának a táblázata. Például a G kulcsjelhez, 10%-os selejtvalószínűséghez tartozó görbe néhány pontja:

Átvétel valószínűsége, %	Selejtvalószínűség, %
99,0	9,75
95,0	13,1
–	–
50,0	23,7
25,0	29,0
10,0	34,1

Ebből kiolvasható, hogy a 0,0975 selejtvalószínűségű, jó tételeket a fenti kétlépcsős terv 0,99 valószínűséggel, a 0,29 selejtvalószínűségű, rossz tételeket 0,25 valószínűséggel veszi át a mintavételi terv. (Ugyanezek az ott talált *OC* görbéről – *O*perating *C*haracteristic görbéről – is nagyjából kiolvashatóak.)

7. Mit jelent a G kulcsjelű, normális vizsgálatnál az $n_1=20$, $c_{10}=2$, $c_{11}=5$, $n_2=20$, $c_{20}=6$, $c_{21}=7$ kétszeres (kétfokozatú) mintavételi terv? $p_1=10\%$, $p_2=30\%$ selejtvalószínűségű terméket milyen valószínűséggel enged át a terv? Mennyi az átadó és átvevő kockázata, vagyis az elsőfajú és másodfajú hiba valószínűsége?

Először 20-elemű mintát veszünk, maximum 2 selejtes esetén a tételt átvesszük, minimum 5 selejt esetén visszautasítjuk, 3 vagy 4 esetén második mintát veszünk (újból 20 elemmel). Ha az első és második mintában maximum 6 a selejtes, a tételt elfogadjuk, 7 vagy több selejt esetén visszautasítjuk. Az átadó kockázata (az elsőfajú hiba valószínűsége) az elfogadott 10% vagy kisebb selejtvalószínűségű tételek visszautasításának a valószínűsége, ez 1%. Az átvevő kockázata (másodfajú hiba valószínűsége) a megtűrt 30% vagy annál nagyobb selejtvalószínűségű tételek elfogadásának a valószínűsége, ez 22%.

8. Mit jelent a következő, G kulcsjelhez tartozó kétfokozatú mintavételi terv?

Kiveendő	Megfelelő	Nem megfelelő
	darabok száma	
8	0	4
8	1	6
8	3	8
8	5	10
8	7	11
8	10	12
8	13	14

Először 8-elemű mintát veszünk, 0 selejt esetén elfogadunk, 4 vagy több selejt esetén visszautasítunk. 1–2–3 selejt esetén másodszor is 8-elemű mintát veszünk. Ha az első és második mintában együttvéve maximum 1 selejtes van, elfogadunk, ha minimum 6, visszautasítunk. Együttvéve 2–3–4–5 selejt esetén harmadszor is 8-elemű mintát veszünk. Ha az első, második, harmadik mintában együttvéve maximum 3 selejtes van, elfogadunk, minimum 8 selejt esetén visszautasítunk. Együttvéve 4–5–6–7 selejt esetén negyedszer is 8-elemű mintát veszünk, és így tovább. Megjegyezzük, hogy az átadó–átvevő kockázata $p_1=0{,}10$, $p_2=0{,}30$ esetén nagyjából ugyanannyi, mint az előző feladatban. (Az egyszeres, kétszeres, többszörös tervtípusok jelleggörbéi azonos kulcsjel esetén nagyjából egybeesnek.)

9. Golyósorsó–golyósanya fárasztási vizsgálatának eredményei (anyag: C60 orsó, csiszoltbekezdésű anyával fárasztva):

Sorszám, i	Törési valószínűség, p_i	Igénybevételi szám százezredrésze, $N_i/10^5$
1	1/16 = 0,0625	0,256 60
2	1/16 = 0,0625	0,406 56
3	1/16 = 0,0625	0,472 89
4	1/16 = 0,0625	0,506 16
5	1/16 = 0,0625	0,507 02
6	1/16 = 0,0625	0,508 14
7	1/16 = 0,0625	0,524 62
8	1/16 = 0,0625	0,566 12
9	1/16 = 0,0625	0,611 76
10	1/16 = 0,0625	0,620 41
11	1/16 = 0,0625	0,659 75
12	1/16 = 0,0625	0,688 94
13	1/16 = 0,0625	0,910 77
14	1/16 = 0,0625	0,973 04
15	1/16 = 0,0625	1,028 04
16	1/16 = 0,0625	1,073 31

(Az igénybevételi szám megmutatja, hogy hányadik fordulatnál következett be a törés.)

Szerkesszük meg az N_i igénybevételi szám eloszlásfüggvényét! Mekkora a 10%, 15,9%, 50% összegezett törési valószínűséghez tartozó N_{10}, $N_{15,9}$, N_{50} igénybevételi szám (élettartam)? Mekkora $S(N) = N_{50} - N_{15,9}$?

Mekkora az igénybevételi szám várható értéke, korrigált szórása?

$\bar{x} = 64\,463{,}313$, $s^* = 23\,485{,}567$. A keresett értékek körülbelül a következők: $N_{10} = 40\,000$, $N_{15,9} = 45\,000$, $N_{50} = 60\,000$, $S(N) = 15\,000$.

10. Egy iparvállalat 14 havi villamosenergia-felhasználása (W) és a termelt mennyiség (T) közötti összefüggést (az ún. *energiafelhasználási függvényt*) keressük lineáris regressziós függvény formájában! Az adatokat l. a táblázatban.

W, kWh	42 010	31 694	33 858	37 263	43 289	51 627	49 671
T, kg	36 445	30 911	33 883	31 394	49 652	48 418	52 722

W, kWh	16 523	54 673	47 500	50 000	66 245	68 696	62 753
T, kg	19 348	40 687	48 898	55 574	45 692	44 989	34 693

Keressük meg W-t mint T függvényét a feltételezett lineáris regresszió alapján! Becsüljük meg, hogy $T = 60\,000$ kg eléréséhez várhatóan hány kWh energia szükséges!

$W = 0{,}8698T + 11\,240$, $T = 60\,000$ esetén $W = 63\,428$.

11. Gördülőcsapágyak élettartamát vizsgálták. A mérési adatok:

Élettartam osztályok (10^5 fordulat)	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8–9	9–10
Gyakoriság	1	2	5	7	10	7	5	3	1

Sűrűségfüggvénnyel szemléltethető, hogy az élettartam jó közelítésben normál eloszlású.

95% biztonsággal milyen megbízhatósági intervallum tartalmazza az élettartam várható értékét? Becsülje meg annak a valószínűségét, hogy a csapágy $5 \cdot 10^5$ fordulat alatt elkopik!

$\bar{x}=5{,}5732, \quad s^*=1{,}808\,18 \quad (10^5 \text{ fordulat}),$

$$\frac{s^*}{\bar{x}} = 32\%,$$

$$5{,}0025 \leqq m \leqq 6{,}1439,$$

$$P(X<5)=F(5)=\Phi(-0{,}317)=0{,}3735,$$

azaz várhatóan az esetek 37%-ában lesz az élettartam ötszázezer fordulatnál kisebb.

Megjegyzés

A gördülőcsapágyak minőségére a csapágygyárak igen nagy hangsúlyt fektetnek, hiszen ez az autók, motorok meghibásodásával szorosan összefüggő tényező. Legtöbbször Weibull-eloszlással közelíthető, néha normálissal, néha log-normálissal.

12. Egy revolver-esztergapad által gyártott rudak átmérőjének eloszlása előzetes tapasztalatok alapján jól közelíthető normál eloszlással. A mérési adatok az átmérőre (mm-ben): 14,9 – 15,0 – 15,0 – 15,1 – 15,1 – 15,1 – 15,2 – 15,2 – 15,2 – 15,2 – 15,3 – 15,3 – 15,3 – 15,4 – 15,4 – 15,5.
Megfelelnek-e a rudak az előírásnak, amely szerint az egész sokaságban a rudak várható értéke 15,4 mm?

$\bar{x}=15{,}2, \quad s^*=0{,}1633,$

$$|t| = \left|\sqrt{16}\,\frac{15{,}2-15{,}4}{0{,}1633}\right| = 4{,}8989 > t_{95} = 2{,}131,$$

$$> t_{99,9} = 4{,}073,$$

az $m_0=15{,}4$ feltevést nem fogadjuk el!

13. Bizonyos izzók élettartama normál eloszlással jól közelíthető. A kétféle technológiával előállított izzók élettartamára az alábbi adatokat kapták (órában):

I. technológia	II. technológia
$\bar{x}_1=1000,$	$\bar{x}_2=1050,$
$s_1^*=10,$	$s_2^*=12,$
$n_1=12,$	$n_2=13.$

Feltehető-e, hogy a két technológia egyformán jó izzókat gyárt (pontosabban: azonos normál eloszlású-e a két változó)?

$$F = \frac{12^2}{10^2} = 1{,}44 < F_{95} = 2{,}79,$$

a szórások egyezése elfogadható (a szabadságfokok: $n_2-1 = 12$ és $n_1-1 = 11$).

$$|t| = \frac{|1000-1050|}{\sqrt{\frac{11\cdot 10^2+12\cdot 12^2}{12+13-2}}\sqrt{\frac{1}{12}+\frac{1}{13}}} = 11{,}2638 > t_{99} = 2{,}807,$$

a várható értékek egyezését elvetjük 99% biztonsági szinten. A Student-táblázat megfelelő részlete:

$n=f+1$	90%	99%
24	1,714	2,807

A két mintából feltehető, hogy a II. technológia jobb, nagyobb várható élettartamú izzókat eredményez.

14. A metróban csúcsforgalmi időben az elvárt átlagos menetidő (a követési idő várható értéke) két állomás közt $m_0=3$ perc. A mért értékeket egy óra mutatja, a megfigyelt értékek percben:

| 2,0 | 2,1 | 2,2 | 2,6 | 2,6 | 2,6 | 2,7 | 2,7 | 2,7 | 2,8 | 2,8 | 2,8 | 2,9 | 3,0 | 3,1 | 3,1 | 3,1 | 3,2 | 3,2 | 3,2 | 3,3 | 3,3 | 3,4 | 3,6 | 3,7 | 3,8 | 4,4 perc.

a) Szemléltessük, hogy a menetidő közelíthető normál eloszlással! *b)* Számítsuk ki, mennyi a valószínűsége annak, hogy a menetidő 3 perc fölé esik! *c)* 95% biztonsági szinten döntsük el, feltehető-e hogy a várható menetidő valóban $m_0=3$ perc!

---

a) Az adatokat

2,0–2,5–3,0–3,5–4,0–4,5

határokkal rendelkező osztályokba sorolva, a gyakoriságok:

3–10–10–3–1 $\qquad$ ($n=27$ adat),

ezek normalitásra utalnak. Ha

2,0–2,4–2,8–3,2–3,6–4,0–4,4

határoknak megfelelő intervallumokba sorolunk, akkor is jól látszik a normalitás, mert a gyakoriságok:

3–6–8–6–3–1.

A sűrűséghisztogram felrajzolását az Olvasó az eddigiek alapján könnyen elvégzi.

b) $\bar{x}=2{,}9963$ perc, $s^*=0{,}2588$,

$$P(X\geq 3) = 1-P(X<3) = 1-\Phi\left(\frac{3-2{,}9963}{0{,}5288}\right) = 0{,}4972.$$

A minta alapján közel 50%-ra becsülhetjük annak a valószínűségét, hogy a szerelvények követési ideje nagyobb az előírt 3 percnél (legalábbis a megfigyelt időszakban).

c) Feltételezhető-e, hogy a teljes üzemidő alatt (az egész sokaságban) teljesül az $m_0=3$ perc előírás? Egymintás t-próbával döntünk:

$$|t| = \left|\sqrt{27}\,\frac{2{,}9963-3}{0{,}5288}\right| = 0{,}0364 < t_{95} = 2{,}056,$$
$$< t_{80} = 1{,}315,$$

az egész üzemidőben már feltehetően teljesül a 3-perces előírás.

15. Egy manipulátor pályakövetési pontosságának vizsgálatánál a manipulátor megfogójának egy előírt egyenestől való távolságát mérték, különböző időpillanatokban. Az előírt távolság és minőségi előírása 12,8 ± 1 mm. A mért távolságokat osztályokba sorolva, a gyakoriságokkal együtt az alábbi táblázat tartalmazza:

Sorszám	Osztályhatárok, mm	Gyakoriság
1.	10,6 – 11,0	4
2.	11,0 – 11,4	13
3.	11,4 – 11,8	55
4.	11,8 – 12,2	118
5.	12,2 – 12,6	202
6.	12,6 – 13,0	241
7.	13,0 – 13,4	212
8.	13,4 – 13,8	121
9.	13,8 – 14,2	29
10.	14,2 – 14,6	12
11.	14,6 – 15,0	5

a) Határozzuk meg az eltérések átlagát, korrigált szórását és relatív szórását!

b) A sűrűséghisztogram megszerkesztésével mutassuk ki, hogy a mért adatok eloszlása jó közelítéssel normál eloszlás!

c) 95%-os biztonsági szinten milyen megbízhatósági intervallumba esik az egész sokaság várható értéke?

d) Számítsuk ki, hogy milyen százalékban nem teljesítette a manipulátor a minőségi előírásokat!

---

a) Helyettesítsük az adatokat az x_i osztályközepekkel ($i = 1, 2, \ldots$, $n = \Sigma f_i = 1012$), ekkor:

$$\bar{x} = \frac{1}{n}\Sigma f_i x_i = 12{,}7759 \text{ mm},$$

$$s^* = \sqrt{\frac{1}{n-1}\Sigma f_i (x_i-\bar{x})^2} = 0{,}6594 \text{ mm},$$

$$\frac{s^*}{\bar{x}} = 0{,}0516,$$

az 5%-os relatív szórás elég kicsi.

b) A sűrűséghisztogram a normalitást jól mutatja (gyakoriságokból már látható). Megszerkesztését az Olvasóra bízzuk.

c) A megbízhatósági intervallum fél hossza:

$$t_{95}\frac{s^*}{\sqrt{n}} = 1{,}96\,\frac{0{,}6594}{\sqrt{1012}} = 0{,}0406\ \text{mm},$$

így az m várható érték megbízhatósági intervalluma 95%-os szinten:

$$(12{,}7353;\ 12{,}8165).$$

d) $P(11{,}8 \leqq X \leqq 13{,}8) = F(13{,}8) - F(11{,}8)$,

$$F(13{,}8) = \varnothing(1{,}55) = 0{,}9332,$$
$$F(11{,}8) = \varnothing(-1{,}48) = 0{,}0694,$$

így a robotkéz (manipulátor) beállítása 86,38%-ban az előírásnak megfelelő pontosságot eredményez, 13,62%-ban nem.

16. Egy gépgyárban egy d méter hosszú tengelyre alkatrészeket kell felhelyezni (szorosan egymás mellé). Ha az előírásnak megfelelően készül el az n darab alkatrész, akkor a tengelyre felfér. Az egyes alkatrészek hossza (tengelyirányban) $X_1, X_2, \ldots, X_n$ külön-külön független, normál eloszlású véletlen változó, $m_1, m_2, \ldots, m_n$ várható értékkel, $\sigma_1, \sigma_2, \ldots, \sigma_n$ szórással.

Várhatóan az esetek hány százalékában férnek fel az alkatrészek? (Ekkor ugyanis nem kell pótlólagos javítás.)

Beszéltünk már arról, hogy független, normál eloszlású változók X összege ($X = \Sigma X_i$) maga is normál eloszlású,

$$m = \Sigma m_i$$

várható értékkel,

$$\sigma^2 = \Sigma\sigma_i^2$$

szórásnégyzettel. (A szummázások mindenütt $i = 1$-től n-ig mennek).

Így annak a valószínűsége, hogy az alkatrészek összhossza nem lesz nagyobb, mint a tengely teljes hossza, a következő:

$$P(X \leqq d) = F(d) = \varnothing\left(\frac{d - \Sigma m_i}{\sqrt{\Sigma\sigma_i^2}}\right).$$

Ezzel a keresett valószínűséget megkaptuk. A kapott általános képlettel több tengelyméretre, több alkatrésztípusra ki lehet számítani a ráférés valószínűségét.

17. Egy kis műhely x_i m^3 vizet fogyasztott és a fogyasztásból y_i %-ot használt fel közvetlenül a termelés céljaira. A következő adatokat mérték:

x_i, m^3	597,	627,	710,	784,	850,	914,
y_i, %	44,8,	45,6,	49,6,	50,7,	52,5	53,8.

Keressünk lineáris regressziót a két változó között!

A pontokat a legkisebb négyzetek elve alapján legjobban megközelítő függvény:

$$y = 14{,}4657 + 0{,}0469x.$$

Ebből látható, hogy 1 m^3 vízfelhasználás-növekedés a közvetlen termelésre fordított vízfelhasználást 4,69%-kal növeli.

18. Egy üzem y_i GJ hőt használt fel és ebből x_i%-ot gépcsarnokai fűtésére fordította. Határozzuk meg a mért adatokat legjobban megközelítő

$$y = a + bx + cx^2$$

másodfokú polinom egyenletét! Az adatokat a következő táblázat mutatja:

y_i	x_i	x_i^2	x_i^3	x_i^4	x_iy_i	$x_i^2y_i$
175	5,1	26,01	132,651	676,5201	892,5	4 551,75
271	6,4	40,96	262,144	1 677,7216	1 734,5	11 100,16
449	7,5	56,25	421,875	3 164,0625	3 367,5	25 256,25
483	8,7	75,69	658,503	5 728,9761	4 202,1	36 558,27
614	9,8	96,04	941,192	9 223,6816	6 017,2	58 968,56
823	11,0	121,00	1 331,000	14 641,0000	9 053,0	99 583,00
1 049	12,2	148,84	1 815,848	24 594,3456	12 797,8	15C 133,16
1 321	13,4	179,56	2 406,104	32 341,7936	17 701,4	237 198,76
1 475	14,9	222,01	3 307,949	49 290,4401	21 927,5	327 464,75
2 260	17,3	290,29	5 177,717	84 258,2841	38 872,0	656 055,40
2 910	23,3	542,89	12 649,337	294 810,5521	67 803,0	1 575 809,90
Összesen: 11 830	129,6	1799,54	29 104,320	520 407,3774	184 368,4	3 188 679,96

A normálegyenletek:

$$11\,830 = 11a + 129{,}6b + 1799{,}54c,$$
$$184\,368{,}4 = 129{,}6a + 1799{,}54b + 29\,104{,}320c,$$
$$3\,188\,679{,}96 = 1799{,}54a + 29\,104{,}320b + 520\,407{,}3774c,$$

ahonnan $a = 2115{,}312$, $b = -321{,}200$, $c = 16{,}776$, vagyis $y = 2115{,}312 - 321{,}200x + 16{,}776x^2$.

19. Határozzuk meg a hazai gépipar energiafelhasználási függvényét lineáris és parabolikus trenddel (a pontokat a legkisebb négyzetek elve alapján legjobban közelítő egyenessel, ill. parabolával)!

Hogy kisebb számokkal kelljen dolgozni, a t év helyett vezessük be az

$$x = t - 1957$$

változót (minden t értékből levonjuk az átlagát)! Ezáltal azt is elérjük, hogy a normálegyenletekben

$$\Sigma x_i = 0, \quad \Sigma x_i^3 = 0$$

lesz, így a normálegyenletek egyszerűsödnek.

Az energiafelhasználás adatait és a számításokat a következő táblázat tartalmazza:

Év, t	$x = t - 1957$	Energiafelhasználás, W (10^9 kcal/év)	x^2	x^4	xW	x^2W	Trendérték: lineáris trend, W_1'	Trendérték: parabolikus trend, W_2'
1953	−4	3 584	16	256	−14 336	57 344	3233	3664
1954	−3	3 422	9	81	−10 266	30 798	3272	3379
1955	−2	3 388	4	16	− 6 776	13 552	3311	3187
1956	−1	2 976	1	1	− 2 976	2 976	3350	3087
1957	0	2 978	0	0	0	0	3389	3080
1958	+1	3 167	1	1	+ 3 167	3 167	3428	3165
1959	+2	3 337	4	16	+ 6 674	13 348	3467	3343
1960	+3	3 725	9	81	+11 175	33 525	3506	3613
1961	+4	3 920	16	256	+15 680	62 720	3545	3976
Összesen:	0	30 497	60	708	+ 2 342	217 430		

Lineáris esetben a normálegyenletek:

$$30\,497 = 9a,$$
$$2\,342 = 60b.$$

Innen $a = 3389$, $b \approx 39$.

A lineáris trend:

$$W_1 = 3389 + 39x = 3389 + 39(t - 1957).$$

Parabolikus közelítéskor a normálegyenletek:

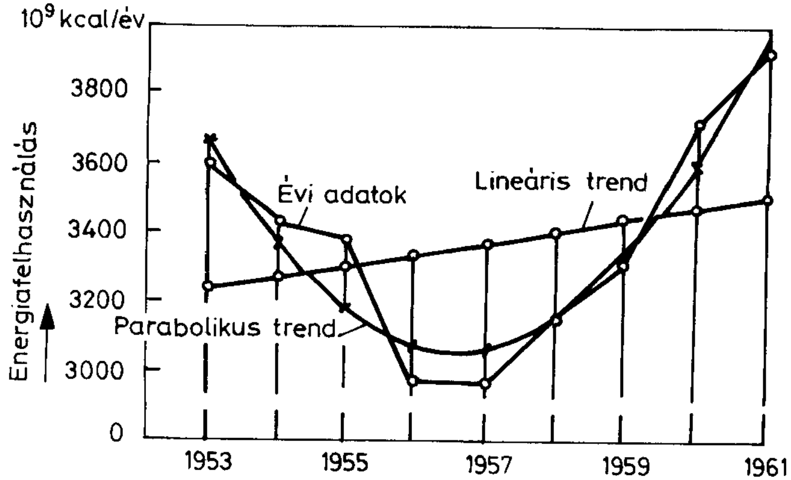
$$30\,497 = 9a + 60c,$$
$$2\,342 = 60b,$$
$$217\,430 = 60a + 708c.$$

Innen $a = 3080$, $b \approx 39$, $c = 46{,}3$.

A trendfüggvény:

$$W_2 = 3080 + 39(t - 1957) + 46{,}3(t - 1957)^2.$$

A mérési pontokra a parabolikus trendfüggvény jól illeszkedik, a lineáris kevésbé (56. ábra).



56. ábra

20. A *bőrkonfekció-iparban* Magyarországon jelenleg még háromféle termelési módszer létezik: *a)* egyedi termelés (kisipar, egy munkadarabot egy ember készít el), *b)* csoportos termelés (több ember közötti szélesebb munkamegosztással), *c)* futószalagos termelés (a futószalag mentén történő még szélesebb munkamegosztással).

Az alábbiakban vizsgáljuk az EI. 41-es táskánál *a futószalagos és csoportos termelés fajlagos költségeinek alakulását.* Az egyik változó az előállított darabszám, a másik változó a táska fajlagos anyagköltsége (Ft/db).

A mérési adatok alapján adódó pontokat ábrázolva látható, hogy mindkét esetben jól közelíthetők a pontok egyenessel.

A regressziós egyenesek metszéspontjának kiszámításával adjuk meg az ún. szériahatárpontot (vagyis *azt a darabszámot, amely alatt a csoportos termelés gazdaságosabb, utána pedig a futószalagos).*

A számítógépes futtatást a már közölt lineáris regressziós programmal végeztük. A bemenő adatokat és a számítás eredményét mutatjuk be az alábbiakban.

Az egymás után következő (X, Y) adatpárok az összetartozó darab és Ft/db fajlagos költségek.

1. *Futószalagos termelés:* a mérési pontok száma 13.

x	500	625	750	875	1000	1125	1250	1375	1500	1625	1750	1875	2000
y	19,8	16	15,8	13,8	13,8	12,9	12,4	12	11,6	11,2	11	10,7	10,5

```
X ATLAG= 1250
Y ATLAG= 13.1923077
MEREDEKSEG M=-5.03736261E-03
KORR.EGYH.=-.916066274
Y ERT. SZORASA= 2.57187793
MARADEKOK SZORASA= 1.03139226
T ERTEK=-7.57617021
```


2. *Csoportos termelés:* a mérési pontok száma 13.

x	500	625	750	875	1000	1125	1250	1375	1500	1625	1750	1875	2000
y	17,6	16,2	14,9	14	13,5	13	12,8	11,8	12,4	12	11,9	11,8	11,6

```
X ATLAG= 1250
Y ATLAG= 13.3461539
MEREDEKSEG M=-3.49010099E-03
KORR.EGYH.=-.90651323
Y ERT. SZORASA= 1.80069015
MARADEKOK SZORASA= .760210701
T ERTEK=-7.12156594
```

A két regressziós egyenes egyenlete:

(Futószalagos termelés) $y = -0{,}005\,037(x-1250)+13{,}192\,308,$

(Csoportos termelés) $y = -0{,}003\,490(x-1250)+13{,}346\,154,$

azaz

$$y = -0{,}005\,037x+19{,}489\,01,$$
$$y = -0{,}003\,490x+17{,}708\,79.$$

Számítsuk ki a két egyenes metszéspontjának abszcisszáját:

$$x = \frac{1{,}780\,22}{1{,}5472527\cdot 10^{-3}} = 1150{,}5684.$$

1150 darab vagy kevesebb táska esetén érdemesebb a manuális, csoportos termelést folytatni, 1151 db vagy több táska esetén rá kell térni a futószalagos termelésre.

21. Egy rendszer három *független* komponensből áll. A komponensek mindegyike p valószínűséggel működik. Mennyi a valószínűsége annak, hogy a rendszer működik, ha a komponensek *a)* sorosan, *b)* párhuzamosan kapcsoltak?

a) p^3; *b)* $1-(1-p)^3$.

a-hoz: A sorbakapcsolt komponensekből álló rendszer csak akkor működik, ha mindegyik komponens működik. Ennek valószínűsége – a komponensek működésének függetlensége folytán – a három működési valószínűség szorzata.

b-hez: A párhuzamos rendszer működik, ha legalább az egyik komponens működik. Célszerű először a komplementer esemény valószínűségét meghatározni. Annak a valószínűsége, hogy mindhárom komponens hibás: $(1-p)^3$. Itt megint a függetlenséget használtuk ki.

Megjegyzés

Gyakorlásul és ellenőrzésként határozzuk meg a „legalább az egyik komponens működik" esemény valószínűségét direkt módon is!

$$(3p(1-p)^2+3p^2(1-p)+p^3.)$$

22. Hány héten keresztül kell játszani heti egy lottószelvénnyel, hogy annak a valószínűsége, hogy legalább egyszer legalább négyes találatot érjünk el, z-nél nagyobb legyen? $(0<z<1)$.

Az $1-(1-p)^n>z$ egyenlőtlenséget kell n-re megoldani, ahol

$$p=\frac{\binom{5}{4}\cdot\binom{85}{1}+1}{\binom{90}{5}}.$$

A legalább négy – 4 vagy 5 – találat valószínűségét jelöljük p-vel.

A legfeljebb három találat – az előbbi esemény komplementerének – valószínűsége: $1-p$.

Annak a valószínűsége, hogy n héten keresztül mindig legfeljebb csak három találatot érünk el az egyes heti húzások függetlensége miatt $(1-p)^n$.

Ezen $(1-p)^n$ valószínűségű esemény komplementer eseménye bekövetkezési valószínűségének kell z-nél nagyobbnak lennie. Ebből n kifejezhető.

23. Csapágygyűrűket készítenek olyan automata gépsoron, amely helyes beállítás esetén hibátlan terméket állít elő. A beállítás azonban minden darab gyártása közben 0,02 valószínűséggel megváltozhat, s így selejtes terméket kapunk. Az ellenőrzés automatikus, és amint selejtes gyártmány készül, a gépsor azonnal leáll, hogy a beállítást korrigálhassák.

Határozzuk meg két beállítás között készülő csapágygyűrűk számának várható értékét!

$n=50$.

Jelölje az X valószínűségi változó a két beállítás között elkészülő munkadarabok számát.

$$P(X=k)=0{,}98^{k-1}\cdot 0{,}02,\qquad (k=1,2,\ldots).$$
$$M(X)=\sum_{k=1}^{\infty}k\cdot 0{,}98^{k-1}\cdot 0{,}02=0{,}02\cdot\sum_{k=1}^{\infty}k\cdot 0{,}98^{k-1}.$$

Mivel a $\sum_{k=1}^{\infty}x^k$ hatványsor (geometriai sor) $|x|<1$ esetén konvergens, sőt egyenletesen konvergens, a tagonkénti deriválással nyert sorra fennáll, hogy

$$\sum_{k=1}^{\infty}kx^{k-1}=\frac{d}{dx}\sum_{k=1}^{\infty}x^k=\frac{d}{dx}\left(\frac{1}{1-x}-1\right)=\frac{1}{(1-x)^2}.$$

Ezt az összefüggést figyelembe véve:

$$M(X)=0{,}02\cdot\frac{1}{(1-0{,}98)^2}=\frac{1}{0{,}02}.$$

24. Gázmolekulák sebességét olyan X valószínűségi változónak tekintjük, amelynek sűrűségfüggvénye a Maxwell-eloszlás szerint:

$$f(x)=\begin{cases}0, & \text{ha } x\leqq 0,\\ \dfrac{4h^3}{\sqrt{\pi}}x^2e^{-h^2x^2}, & \text{ha } x>0,\ (h=\text{állandó}).\end{cases}$$

Számítsuk ki a gázmolekulák sebességének a várható értékét!

$$M(X) = \frac{2}{h\sqrt{\pi}}.$$

A várható érték parciális integrálással adódik. (Használjuk ki, hogy

$$\lim_{x\to\infty} x^2 e^{-h^2x^2} = 0.)$$

25. Bizonyos munkadarabok elkészítése során egy-egy tengelyre 10 db egyforma méretű alkatrészt helyeznek, szorosan egymás mellé. Az alkatrészek hossza egymástól független ingadozást mutat 10 cm körül, 0,2 szórással.

Mekkora az összeszerelt munkadarabok várható hossza és mekkora a szórása?

$$M(X) = l = 1 \text{ m}; \qquad D(X) = \sqrt{0{,}4} \text{ cm}.$$

26. Egy üzem termékeit I., II., és III. osztályú minősítéssel látják el. A készáruraktárban az üzem A és B jelű termékei az alábbi megoszlásban szerepelnek:

Termék	I.	II.	III.
A	500	200	100
B	650	300	50

Ebből a halmazból egy terméket véletlenszerűen kiemelnek. $X=0$ jelentse, hogy az A, $X=1$ pedig, hogy B a termék! Az $Y=1$, $Y=2$, $Y=3$ érték jelentse a kivett termék minősítését!

a) Számítsuk ki $X+Y$ várható értékét és szórását!

b) Határozzuk meg X és Y korrelációs együtthatóját!

---

a) $M(X+Y) = 2$; $D(X+Y) = \dfrac{\sqrt{22}}{6}$.

b) $R(X, Y) = -\sqrt{\dfrac{2}{335}}$.

27. Egy szövőgép 500 szállal dolgozik. A szálak munka közben egymástól függetlenül elszakadhatnak. Az egy műszak alatti szálszakadás valószínűsége minden szálra 0,012.

Mennyi annak a valószínűsége, hogy egy műszak alatt: *a)* legalább három szál, *b)* legfeljebb három szál, *c)* pontosan három szál szakad el?

---

a) 0,9745; *b)* 0,0719; *c)* 0,0464.

Az egy műszak alatt elszakadó szálak száma binomiális eloszlást alkot. Jelöljük p_i-vel annak a valószínűségét, hogy egy műszak alatt *pontosan i* darab szál szakad el. Ezzel a jelöléssel élve, rendre a

$$\sum_{i=3}^{600} p_i; \quad \sum_{i=0}^{3} p_i;$$

ill. a p_3 valószínűségek meghatározása a feladatunk.

Mivel $n=600$ igen nagy szám, és $p=0{,}012$ elég kicsi, a binomiális eloszlás tagjai jól közelíthetők a $\lambda = np = 600 \cdot 0{,}012 = 7{,}2$ paraméterű Poisson-eloszlás megfelelő tagjaival. Így

$$\sum_{i=3}^{600} p_i = 1 - \sum_{i=0}^{2} p_i \approx 1 - \sum_{i=0}^{2} \frac{7{,}2^i}{i!} e^{-7{,}2} =$$

$$= 1-(0{,}0007+0{,}0054+0{,}0194) = 0{,}9745.$$

Továbbá

$$\sum_{i=0}^{3} p_i \approx \sum_{i=0}^{3} \frac{7{,}2^i}{i!} e^{-7{,}2} = 0{,}0007+0{,}0054+0{,}0194+0{,}0464 = 0{,}0719.$$

Végül

$$p_3 \approx \frac{7{,}2^3}{3!} e^{-7{,}2} = 0{,}0464.$$

28. Egy biztosítótársaság 10 000 hasonló szociális helyzetű és korú biztosítottja január 1-én 10-10 Ft-ot fizet a biztosító pénztárába. A biztosított elhalálozása esetén hozzátartozója 4000 Ft-ot kap a biztosítótól.

Annak a valószínűsége, hogy egy biztosított személy az év folyamán meghal, 0,002. Mennyi a valószínűsége annak, hogy a társaságnak *a)* sem nyeresége, sem vesztesége nem lesz; *b)* nem lesz nyeresége?

---

a) 0,0446; *b)* 0,1667.

a)-hoz: Tulajdonképpen az a kérdés, hogy mi a valószínűsége annak, hogy a biztosító a befolyt $10\,000 \cdot 10 = 10^5$ Ft összeget maradéktalanul kifizeti (és nem kell a tőkéjéből pótolnia sem), vagyis, hogy pontosan $\frac{10^5}{4000} = 25$ haláleset következik be az év folyamán (egymástól függetlenül).

Az $n = 10\,000$, $p = 0{,}002$ paraméterű binomiális eloszlásról van szó. Ezt a $\lambda = 20$ paraméterű Poisson-eloszlással közelítjük. Most $k = 25$, így táblázatunkból a $P(X = 25) = 0{,}0446$ értéket olvassuk ki.

b)-hez: Az előző feladathoz képest az a változás, hogy a 25 vagy annál több haláleset bekövetkezési valószínűségét kell meghatározni.

Ismét a táblázatból:

$P(X \geqq 25) = 0{,}1667$ adódik.

29. Valamely áruszállítással kapcsolatban 0,2 a valószínűsége annak, hogy egy gépkocsinak 1 óránál többet kell várnia.

a) Mi a valószínűsége, hogy egy érkező gépkocsi fél órán belül sorrakerül, ha a várakozási idő exponenciális eloszlású valószínűségi változó?

b) Mennyi várakozási időt kell a gépkocsiknak betervezniük?

---

a) $\approx 0{,}5528$; *b)* $\approx 0{,}62$ óra.

a)-hoz: $P\left(X < \frac{1}{2}\right) = 1 - \sqrt{0{,}2}$,

b)-hez: $M(X) = \frac{1}{\lambda}$; $e^{-\lambda} = 0{,}2$,

$$M(X) = \frac{1}{\ln 5}.$$

30. Egy automata daraboló 10 cm várható értékkel dolgozik. A szórás a gép beállításával szabályozható. Határozzuk meg, hogy mekkora szórás esetén teljesül a

$$P(|X - M(X)| \leqq 0{,}1) = 0{,}95$$

feltétel! (Az automatákon gyártott termékek mérete normál eloszlásúnak tekinthető.)

$\sigma \approx 0{,}05$ cm.

A feltételi egyenlet ugyanis ekvivalens a

$$2 \cdot \varphi\left(\frac{0{,}1}{\sigma}\right) - 1 = 0{,}95$$

egyenlettel.

31. Görgőscsapágyak készítéséhez hengereket (görgőket) gyártanak. A görgők hossza (X) és átmérője (Y) olyan normál eloszlású valószínűségi változónak tekinthető, amelyek ingadozása egymástól független.

Statisztikai megfigyelések alapján a 12 mm hosszúságúra tervezett hosszméret szórása 0,055 mm, a 6 mm-nek tervezett átmérő szórása pedig 0,028 mm.

Egy görgő akkor tekinthető selejtesnek, ha hosszmérete a tervezett értéktől 0,1 mm-nél, átmérőjének eltérése pedig 0,05 mm-nél nagyobb.

Mennyi a valószínű selejtszázalék?

Kb. 13,8%.

Jelentse A a $11{,}9 \leqq X \leqq 12{,}1$, B az $5{,}95 \leqq Y \leqq 6{,}05$ eseményt.

Feladatunk a $P(\bar{A} + \bar{B})$ valószínűség meghatározása!

A De Morgan-féle azonosságot és az események függetlenségét felhasználva, a kérdéses valószínűségre adódik:

$$P(\bar{A} + \bar{B}) = P(\overline{AB}) = 1 - P(AB) = 1 - P(A)P(B).$$

(Ugyanerre az eredményre jutunk, ha a valószínűségszámításban megismert tételeket alkalmazzuk:

$$P(\bar{A} + \bar{B}) = P(\bar{A}) + P(\bar{B}) - P(\bar{A} \cdot \bar{B}) = P(\bar{A}) + P(\bar{B}) - P(\bar{A})P(\bar{B}) \quad \text{sít.})$$

32. Olyan ellenállásokkal dolgozunk, amelyek egymástól független ingadozásokat mutatnak 550 Ω várható értékkel és 30 Ω szórással, egyenletes eloszlásban. A centrális határeloszlás segítségével adjunk becslést annak a valószínűségére, hogy 10, sorbakapcsolt ellenállás eredője 5600 Ω-nál nagyobb!

Az eredő ellenállás olyan valószínűségi változó, amelynek várható értéke 5500 Ω és szórása $\sqrt{10}\cdot 30\,\Omega$. Így standardizáltja: $Y = \dfrac{X-5500}{\sqrt{10}\cdot 30}$.

Az $X > 5600\,\Omega$ esemény ekvivalens az $Y > \dfrac{\sqrt{10}}{3}$ eseménnyel. Ennek ellentettje az $Y \leqq \dfrac{\sqrt{10}}{3}$ esemény, amelynek bekövetkezési valószínűségére a centrális határeloszlástétel segítségével a $P\left(Y \leqq \dfrac{\sqrt{10}}{3}\right) \approx \Phi\left(\dfrac{\sqrt{10}}{3}\right)$ becslés írható fel.

Így a kérdezett valószínűség becslése:

$$P(X > 5600) = P\left(Y > \frac{\sqrt{10}}{3}\right) \approx 1-\Phi\left(\frac{\sqrt{10}}{3}\right) \approx 14{,}7\%.$$

33. Egy üzemben csavarokat csomagolnak. Egy-egy dobozba átlagosan 5000 csavar kerül. A csavarok számának szórása a tapasztalat szerint 20 darab. Mit mondhatunk annak valószínűségéről, hogy egy dobozban a csavarok száma 4900 és 5100 közé esik, ha az eloszlást nem ismerjük?

Legalább 96% a valószínűsége annak, hogy a csavarok száma 4900 és 5100 között van.

34. Automata vizsgálót használva 10^5 számú vizsgálat után milyen biztonsággal állíthatjuk, hogy a selejt előfordulásának relatív gyakorisága és a tényleges selejtarány legfeljebb 0,01-dal tér el egymástól?

A feladat a nagy számok Bernoulli-törvénye alapján oldható meg. Adataink: $n = 10^5$, $\varepsilon = 10^{-2}$; p-t nem ismerjük, és ezért δ-ra felső becslést adunk.

$$\delta = \frac{pq}{n\varepsilon^2} = \frac{1}{4n\varepsilon^2} = 0{,}025.$$

Így $1-\delta \geqq 0{,}975$. Vagyis az $\left|\dfrac{k}{10^5} - p\right| < 0{,}01$ esemény bekövetkezési valószínűsége legalább 97,5%.

35. Egy ládában kétféle méretű csavar van összekeverve, igen nagy mennyiségben. A számunkra megfelelő csavarok aránya 70%. A halmazból véletlenszerűen kiemelünk 10 000 darabot.

a) Mennyi lesz ezek között a céljainknak megfelelő csavarok várható száma?

b) Mennyi a valószínűsége, hogy a megfelelő csavarok számának valódi értéke a várható értéktől annak legfeljebb 5%-ával tér el?

---

a) 7000; *b)* legalább 0,983.

b)-hez: 7000 csavar 5%-a 350.

Az $|k-7000| < 350$ esemény ekvivalens az $\left|\dfrac{k}{10^4} - 0{,}7\right| < 0{,}035$ eseménnyel.

A $p = 0{,}7$; $n = 10^4$; $\varepsilon = 0{,}035$ értékekből δ és így az $1-\delta$ valószínűség (alsó becslés) közvetlen számítással adódik.

36. Egy célpontra 200 lövést adnak le. A találat valószínűsége minden lövésnél 0,4. Milyen határok közé fog esni 90% valószínűséggel a találatok száma?

A találatok száma 90% valószínűséggel 58 és 102 közé fog esni. (Ennél pontosabb becslés is adható a centrális határeloszlás-tétel alapján.)

Első megoldás: A nagy számok Bernoulli-féle törvénye alapján ε-t úgy kell meghatározni, hogy az

$$\left|\frac{k}{200}-0,4\right|<\varepsilon$$

esemény legalább 0,9 valószínűséggel következzen be.

$$\varepsilon^2=\frac{pq}{n\delta}\text{-ból}\quad \varepsilon=0,11 \text{ adódik.}$$

Így $0,29<\frac{k}{200}<0,51$.

Második megoldás: A Moivre–Laplace-formula felhasználásával. A találatok várható száma: $np=80$. A szórás: $\sqrt{npq}=\sqrt{48}=4\sqrt{3}$. A találatok számának a várható értéktől való eltérése legyen $\lambda\sqrt{npq}$.

Így $\Sigma\binom{200}{k}0,4^k0,6^{200-k}\approx 2\Phi(\lambda)-1=0,9.$

$80-\lambda\cdot 4\sqrt{3}\leqq k\leqq 80+\lambda 4\sqrt{3}.$

Innen $\Phi(\lambda)=0,95$; $\lambda=1,645$.

Tehát $80-1,645\cdot 4\cdot\sqrt{3}\leqq k\leqq 80+1,645\cdot 4\cdot\sqrt{3}$, vagyis $68<k<92$.

37. Egy műszer r párhuzamosan kapcsolt komponensből áll. Jelölje p annak a valószínűségét, hogy egy komponens egy munkaperiódus alatt nem romlik el! p meghatározására „n” független munkaperiódust figyeltek meg. Ezalatt a műszer k-szor romlott el. Határozzuk meg p maximum-likelihood becslését, $\hat{p}$-t! (Az egyes komponensek meghibásodása legyen egymástól független!)

A műszer mindaddig működik, míg legalább egy komponense funkcionál. Tehát csak akkor romlik el, ha minden egyes komponense tönkremegy. Ezen esemény bekövetkezésének valószínűsége: $z=(1-p)^r$, az egyes komponensek meghibásodásának egymástól való függetlensége miatt.

Az n munkaperiódus alatti műszermeghibásodások száma a feladat értelmében k.

A szóban forgó esemény bekövetkezési valószínűsége tehát:

$$z^k(1-z)^{n-k}$$

(z paraméterű karakterisztikus valószínűségi változóhoz tartozó likelihood-függvény).

Feladatunk ezen valószínűség (mint z függvénye) maximumának meghatározása.

$$f(z)=z^k(1-z)^{n-k},$$
$$f'(z)=[kz^{k-1}(1-z)^{n-k}+z^k(n-k)(1-z)^{n-k-1}\cdot(-1)]=$$
$$=z^{k-1}(1-z)^{n-k-1}\cdot[k(1-z)-z(n-k)],$$

és ez zérus, ha $z=\frac{k}{n}$.

Így $z=(1-p)^r$-ből p maximum-likelihood becslése:

$$\hat{p}=1-\sqrt[r]{z}=1-\sqrt[r]{\frac{k}{n}}.$$

38. Oldjuk meg az előző feladatot r sorosan kapcsolt komponensekből álló műszert feltételezve!

$$\hat{p}=\sqrt[r]{1-\frac{k}{n}}.$$

39. Valamely fonál szakítószilárdságát vizsgálva, 12 mérésből az $\bar{x}=3,25$ kp értéket kapták.

A szórás ismert: $\sigma=0,30$ kp.

a) Állapítsuk meg a mintaátlag megbízhatósági intervallumát 95%-os szinten!

b) Legalább hány mérést kellene végezni ahhoz, hogy 95%-os megbízhatósági szint esetén a konfidencia-intervallum hossza 0,2 kp legyen?

---

a) [3,08; 3,42]; *b)* 35.

a)-hoz: Az empirikus várható érték (mintaátlag) olyan valószínűségi változó, amelynek várható értéke az elméleti várható érték: m;

szórása pedig az elméleti szórás $\sqrt{n}$-edrésze, $\frac{\sigma}{\sqrt{n}}$. Így standardizáltja: $\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}$ a centrális határeloszlás-tétel értelmében aszimptotikusan $N(0, 1)$-eloszlású valószínűségi változó.

A feladatban közölt adatokat felhasználva, tehát a

$$P\left(\left|\frac{\bar{X}-3{,}25}{\frac{\sigma}{\sqrt{n}}}\right| < x\right) \approx \underline{2\cdot \Phi(x)-1 = 0{,}95}$$

összefüggés aláhúzott egyenlőségéből először x-et határozzuk meg, majd x ismeretében az $|\bar{X}-m|$ konfidenciasugarat.

Az 5. táblázatból: $x=1{,}96$. Így

$$|\bar{X}-m| < 1{,}96\frac{0{,}3}{2\sqrt{3}} \approx 0{,}17.$$

Tehát a várható értéknek a megadott mintából számított 95%-os megbízhatósági intervalluma:

(3,08; 3,42).

b)-hez: Adott valószínűségi szinten a konfidenciasugár akkor csökkenthető, ha több információ áll rendelkezésünkre.

$$1{,}96\frac{0{,}3}{\sqrt{n}} \leqq 0{,}1 \text{ teljesül, ha } n \geqq 35.$$

Tehát legalább 35 mérést kell végezni ahhoz, hogy 95%-os megbízhatósági szinten 0,1 kp sugarú konfidencia-tartományt adhassunk meg.

40. *a)* Hogyan módosulnak a 95%-os megbízhatósági intervallum határai az előző, 39. feladatban, ha az alapsokaság szórása nem ismert, hanem azt is a 12 mérésből állapították meg. Legyen tehát a korrigált empirikus szórás: $s_n^* = 0{,}30$ kp.

b) Legalább hány mérés szükséges ebben az esetben ugyanolyan – 0,2 kp – hosszúságú konfidencia-intervallum biztosításához, mint az előző feladatban?

a) (3,06; 3,44); *b)* 44.

a)-hoz: Mivel az alapsokaság szórását is a mintából kellett megbecsülni, a bizonytalanság megnő. Változatlan megbízhatósági szinten – érthetően – szélesebb konfidencia-intervallumot tudunk csak megadni. Az $\frac{\bar{X}-m}{\frac{s}{\sqrt{n}}}$ statisztika most $(n-1)$-szabadságfokú Student-eloszlást követ.

Eredményünk a

$$P\left(\left|\frac{\bar{X}-3{,}25}{\frac{0{,}30}{2\sqrt{3}}}\right| < t\right) \approx \underline{S_{11}(t) = 0{,}05}$$

összefüggésláncból adódik, az előző feladatnál alkalmazott gondolatmenet analógiájára, a 8. táblázat adatainak felhasználásával.

b)-hez: 12-nél nyilván több mérésre van szükség. A 11 szabadságfokhoz tartozó $t=2{,}201$ értékkel dolgozva, a szükséges mérések számát biztosan majoráljuk. (L. a táblázatot!)

$2{,}201\cdot\frac{0{,}3}{\sqrt{n}} \leqq 0{,}1$ összefüggésből $n \geqq 44$ adódik.

Megjegyzés

Ha az előző feladat eredményét ($n \geqq 35$) felhasználjuk, a becslés finomítható: $n \geqq 37$.

41. Árammérőket úgy igazítanak be, hogy a mérőket szinkron működtetik egy standard árammérővel.

Beigazítás után 10 árammérőből álló mintát vesznek, és ellenőrzik, hogy konstansaik csak véletlenszerűen térnek-e el a standard árammérő konstansától. Ebből a célból precíziós wattmérővel és stopperrel pontosan meghatározzák a szóban forgó árammérők konstansait, majd t-próbával ellenőrzik azt a feltevést, hogy az eltérések véletlenszerűek. A 10 mérő konstansai:

0,985	0,987
1,003	0,993
0,996	0,991
0,994	1,004
1,002	0,985.

A standard mérő konstansa 1. Szignifikáns-e az eltérés 99%-os szinten?

Az eltérés ezen a szinten nem szignifikáns. t-statisztikánk helyettesítési értéke 2,626. Ez kisebb, mint a kilenc szabadságfokhoz tartozó kritikus t érték: 3,250. (Az eredmény mégsem megnyugtató, mert kicsi elemszámú mintából adódik, és a szignifikanciaszint igen magas.)

42. Egy városban, adott időpontban 240 parkolóhelyen végeztek felmérést. A parkolóhelyeken talált gépkocsik számára nézve az alábbi táblázatot készítették el:

Gépkocsik száma, g_i	Parkolóhelyek száma, ν_i
$0 \leqq g_1 < 10$	11
$10 \leqq g_2 < 20$	22
$20 \leqq g_3 < 30$	46
$30 \leqq g_4 < 40$	85
$40 \leqq g_5 < 50$	43
$50 \leqq g_6 < 60$	20
$60 \leqq g_7 < 70$	13
Összesen:	240

A parkolóhelyeken állomásozó járművek száma normális eloszlású-e 95%-os szignifikanciaszinten?

Nem.

A szignifikanciapróba elvégzéséhez célszerű táblázatot készíteni:

Sorszám	Osztályok	Osztályközepek, x_i	Mintabeli gyakoriság, ν_i	$z_i = \dfrac{x_i-34,5}{10}$	$\nu_i z_i$	$\nu_i z_i^2$	p_i	Várt gyakoriságok, $240p_i$	$\dfrac{(\nu_i-240p_i)^2}{240p_i}$
1.	[0,10)	4,5	11	−3	−33	99	0,0409	9,82	0,14
2.	[10,20)	14,5	22	−2	−44	88	0,1106	26,54	0,78
3.	[20,30)	24,5	46	−1	−46	46	0,2241	53,78	1,13
4.	[30,40)	34,5	85	0	0	0	0,2787	66,89	4,90
5.	[40,50)	44,5	43	1	43	43	0,2110	50,64	1,15
6.	[50,60)	54,5	20	2	40	80	0,1000	24,00	0,67
7.	[60,70)	64,5	13	3	39	117	0,0347	8,33	2,62
Összesen:			240		−1	473	1,000		11,59

Az osztályok száma: 7.

$$\bar{x} \approx 34,5 + \frac{10}{240}\cdot(-1) \approx 34,46.$$

$$s_n^2 \approx \frac{100}{239}\left(473 - \frac{1}{240}\right), \qquad s_n \approx 14,07.$$

Feladatunk tehát, hogy 95%-os szignifikanciaszinten döntsük el: származhat-e a minta $N(34,46;\ 14,07)$ normál eloszlásból.

A próbát $7-1-2 = 4$ szabadságfokú, aszimptotikusan χ^2-eloszlást követő statisztikára alapozzuk. (Az osztályok számát a mintából becsült két paraméter számával is csökkentettük a szabadságfok megállapításakor!) A próba megbízhatóságához szükséges $np_i \geqq 10$ feltétel lényegében teljesül.

A p_i valószínűségeket az 5. táblázat adatainak felhasználásával számítjuk ki. Például:

$$\begin{aligned} p_1 &= P(X<10) = F(10) = \Phi\left(\frac{10-34,46}{14,07}\right) = \Phi(-1,74) = \\ &= 1-\Phi(1,74) = 1-0,9591 = 0,0409. \\ p_2 &= P(10 \leqq X < 20) = F(20)-F(10) = \Phi(-1,03)-\Phi(-1,74) = \\ &= \Phi(1,74)-\Phi(1,03) = 0,9591-0,8485 = 0,1106. \end{aligned}$$

χ^2-statisztikánk mintából számított helyettesítési értéke (táblázatunk utolsó oszlopának összege): 11,59. Ez az érték nagyobb, mint a kritikus 9,488 érték. Ezen a szinten tehát szignifikáns az eltérés a H_0 hipotézishez képest: nem tekinthető normál eloszlású valószínűségi változónak a parkolóhelyeken előforduló gépkocsik száma. (Bár „ránézésre” az eloszlás szimmetrikus: könnyen normálisnak tűnhet.)

43. Egy gyár munkavédelmi osztályán azt a kérdést vizsgálják, hogy 1 év alatt az 1 munkásra jutó balesetek száma Poisson-eloszlást mutat-e. A vizsgálatokhoz 400 munkást választottak ki, véletlenszerűen. Ez a „400 elemű minta” az alábbi adatokat szolgáltatta:

Az egy főre eső balesetek száma	0	1	2	3	4	>4	Összesen
Annak gyakorisága	141	150	83	18	8	0	400

Ellenőrizzük a hipotézist χ^2-próbával, 95%-os szignifikancia szinten! (A H_0 hipotézis Poisson-eloszlásának ismeretlen paraméterét annak maximum-hkelihood becslésével, $\bar{x}$-sal – a mintaátlaggal – pótoljuk!)

A hipotézis 95%-os szinten elfogadható,

$$\hat{\lambda} = \bar{x} = 1{,}005.$$

A

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$$

statisztikánk helyettesítési értéke:

$$\frac{1}{400}\left[\frac{(141-400\cdot 0{,}3661)^2}{0{,}3661} + \frac{(150-400\cdot 0{,}3678)^2}{0{,}3678} + \right.$$

$$+ \frac{(83-400\cdot 0{,}1848)^2}{0{,}1848} + \frac{(18-400\cdot 0{,}062)^2}{0{,}062} +$$

$$\left. + \frac{(8-400\cdot 0{,}0155)^2}{0{,}0155} + \frac{(0-400\cdot 0{,}0038)^2}{0{,}0038}\right] = 5{,}28.$$

(A p_i valószínűségeket $\lambda = 1{,}005$ miatt interpolációval kaptuk a 3. táblázatból.)

Ezt a helyettesítési értéket kell összehasonlítani az $f=4$ szabadságfokú kritikus χ^2 értékkel.

A szabadságfok a csoportok számánál 2-vel kisebb, hiszen az eloszlás paraméterét is a mintából becsültük.

Mivel 5,28 lényegesen kisebb a 7. táblázatból kiolvasható 9,488 kritikus értéknél, azért elfogadhatjuk azt a feltételezést, hogy az egy főre jutó üzemi balesetek száma Poisson-eloszlást mutat.

44. Az 1974. és az 1975. évi baleseti statisztikák közelítő adatait tartalmazza az alábbi táblázat:

Év	Összes sérültek száma	48 órán belül meghalt	30 napon belül meghalt	A többi
1974	26 500	1350	1800	23 350
1975	26 000	1400	1900	22 700

Vizsgáljuk meg a fenti minták alapján, hogy a különböző súlyosságú sérülések bekövetkezése az egyes években azonos eloszlásúnak tekinthető-e, azaz el lehet-e fogadni 99%-os szinten a

$$H_0\colon\ p_1 = q_1;\ p_2 = q_2;\ p_3 = q_3$$

nullhipotézist, ahol p_i, ill. q_i ($i = 1, 2, 3$) a különböző típusú sérülések bekövetkezésének valószínűségét jelöli az 1974., illetőleg az 1975. évben!

A nullhipotézis (a magas) 99%-os szignifikanciaszinten elfogadható.

A mintából számított

$$\chi^2 = \frac{1}{mn}\sum_{i=1}^{r} \frac{(n\mu_i - m\nu_i)^2}{\mu_i + \nu_i} = 8{,}03,$$

ez kisebb, mint az $f=2$ szabadságfokú, 99%-os szinthez tartozó kritikus χ^2 érték: 9,21.

45. Egy egyorsós forgácsolóautomatán gyártott 394 darab golyósolajozótörzs átmérő és hossz szerinti eloszlását az alábbi táblázat tartalmazza:

Átmérő, mm 7+	Hossz, mm 10+					
	0,05-ig	0,06-tól 0,10-ig	0,11-től 0,15-ig	0,16-tól 0,20-ig	0,21-től	Összesen, db
–0,07-ig	12	14	13	20	17	76
–0,06-tól –0,05-ig	7	18	34	18	8	85
–0,03-tól –0,01-ig	13	43	41	24	13	134
–0,00-tól	11	35	34	17	2	99
Összesen, db	43	110	122	79	40	394

Kérdés, hogy az átmérő és a hossz ingadozása 99,9%-os szinten függetlennek tekinthető-e egymástól?

Nem függetlenek.

Az adatokból számított χ^2-statisztikára 37,03 adódik, míg a 12 szabadságfok mellett $\chi^2_{0,999} = 32{,}9$. Ez azt jelenti, hogy e nagy biztonságot szolgáltató szinten sem egyeztethető össze a kapott χ^2 érték a számára – függetlenség esetén – fennálló eloszlással.

46. Egy vezetőn ismeretlen erősségű áram halad. Az áramerősséget 1, 2, 3, 5, 7, 10 amperrel megnövelve, a felmelegedés rendre 11, 12, 15, 22, 34, 61 °C-nak adódott. (Ezek mért adatok, a tényleges áramerősség és a tényleges felmelegedés ettől eltérhet.)

Készítsünk korrelációs diagramot, s ennek alapján végezzünk kvadratikus regressziót!

$$y = 0{,}56x^2 - 0{,}63x + 11{,}25.$$

Δ_i [A]	1	2	3	5	7	10
Δ_T [C°]	11	12	15	22	34	61

$$\sum_{i=1}^{6} x_i = 28; \quad \sum_{i=1}^{6} x_i^2 = 188; \quad \sum_{i=1}^{6} x_i^3 = 1504;$$

$$\sum_{i=1}^{6} x_i^4 = 13\,124; \quad \sum_{i=1}^{6} y_i = 155; \quad \sum_{i=1}^{6} x_i y_i = 1038;$$

$$\sum_{i=1}^{6} x_i^2 y_i = 8510.$$

A megoldandó egyenletrendszer:

$$\begin{aligned} 13\,124a + 1504b + 188c &= 8510 \\ 1\,504a + 188b + 28c &= 1038 \\ 188a + 28b + 6c &= 155 \end{aligned}$$

$$a = 0{,}56; \quad b = -0{,}63; \quad c = 11{,}25.$$

47. Legyen (X, Y) kétdimenziós, normál eloszlású változó, X és Y különkülön standard normál eloszlásúak. Mit tudunk mondani $Z = X + Y$ sűrűségfüggvényéről,

a) ha X és Y függenek egymástól, korrelációs együtthatójuk $r \neq 0$;

b) X és Y függetlenek, korrelációs együtthatójuk $r = 0$?

c) ha X és Y standard normál eloszlásúak, akkor $\sigma_1 X + m_1$ $N(m_1, \sigma_1)$-eloszlású, $\sigma_2 Y + m_2$ $N(m_2, \sigma_2)$-eloszlású. Mit tudunk mondani a két, általános normál eloszlású változó összegének, $Z = \sigma_1 X + \sigma_2 Y + m_1 + m_2$-nek a sűrűségfüggvényéről?

---

a) Mivel az együttes sűrűségfüggvény:

$$f(x, y) = \frac{1}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}(x^2 - 2rxy + y^2)},$$

ezért X és Y összegének sűrűségfüggvénye:

$$g_1(z) = \int_{-\infty}^{\infty} f(t; z-t)\, dt =$$

$$= \frac{1}{2\pi \cdot \sqrt{1-r^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-r^2)}[t^2 - 2rt(z-t) + (z-t)^2]}\, dt.$$

Az e kitevőjében a szögletes zárójelben álló kifejezésben kiválasztunk egy teljes négyzetet:

$$t^2 - 2rtz + 2rt^2 + z^2 - 2tz + t^2 = (2r+2)t^2 - (2r+2)tz + z^2 =$$

$$= (2r+2)\left(t^2 - tz + \frac{z^2}{4}\right) + z^2 - \frac{z^2}{4}(2r+2) =$$

$$= (2r+2)\left(t - \frac{z}{2}\right)^2 + z^2\left(\frac{1}{2} - \frac{r}{2}\right).$$

Az összeg sűrűségfüggvényében a csak z változót tartalmazó tényező független az integrációs változótól, t-től, így az integráljel elé kivihető:

$$g_1(z) = \frac{1}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} z^2 \cdot \frac{1-r}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-r^2)} \cdot 2(1+r)\left(t - \frac{z}{2}\right)^2}\, dt =$$

$$= \frac{1}{2\pi\sqrt{1-r^2}} e^{-\frac{z^2}{4(1+r)}} \int_{-\infty}^{\infty} e^{-\frac{1}{1-r}\left(t-\frac{z}{2}\right)^2} dt.$$

Fel akarjuk használni, hogy $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = 1$. Az

$$\frac{1}{\sqrt{1-r}}\left(t-\frac{z}{2}\right) = \frac{u}{\sqrt{2}}, \qquad dt = \sqrt{\frac{1-r}{2}}\, du$$

helyettesítést kell végrehajtanunk. A helyettesítéskor a határok nem változnak. Így a keresett sűrűségfüggvény:

$$g_1(z) = \frac{1}{2\pi\sqrt{1-r^2}} \frac{\sqrt{1-r}}{\sqrt{2}} e^{-\frac{z^2}{2(\sqrt{2(1+r)})^2}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du,$$

$$g_1(z) = \frac{1}{\sqrt{2\pi}\sqrt{2(1+r)}} e^{-\frac{z^2}{2(\sqrt{2(1+r)})^2}},$$

ez azt jelenti, hogy *két olyan standard normál eloszlású változó összege, amelyek korrelációs együtthatója $r \neq 0$, szintén normál eloszlású, $m=0$ várható értékkel, $\sigma = \sqrt{2(1+r)}$ szórással.*

b) Ha a két változó független, akkor $r=0$, tehát *két független, standard normál eloszlású változó összege szintén normál eloszlású, $m=0$ várható értékkel, $\sigma=\sqrt{2}$ szórással.*

c) $Z = \sigma_1 X + \sigma_2 Y + m$ (ahol $m = m_1 + m_2$) a

$$z = \sigma_1 x + \sigma_2 y + m, \qquad \text{azaz} \qquad x = \frac{1}{\sigma_1}(z - \sigma_2 t - m),$$
$$t = y, \qquad\qquad y = t$$

transzformációval és inverz transzformációval kapható meg ($\sigma_1 \neq 0$ miatt az inverz transzformáció mindig létezik). A Jacobi-determináns nem 0:

$$y = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial t} \end{vmatrix} = \begin{vmatrix} \frac{1}{\sigma_1} & -\frac{\sigma_2}{\sigma_1} \\ 0 & 1 \end{vmatrix} = \frac{1}{\sigma_1} \neq 0.$$

Z és T együttes sűrűségfüggvényét X és Y

$$f(x, y) = \frac{1}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}(x^2 - 2rxy + y^2)}$$

együttes sűrűségfüggvényéből kaphatjuk meg, x és y helyébe az inverz transzformációnál szereplő kifejezéseket helyettesítve és a Jacobi-determináns abszolút értékével szorozva:

$$g(z, t) = \frac{1}{2\pi\sigma_1\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left[\frac{1}{\sigma_1^2}(z-\sigma_2 t - m)^2 - \frac{2r}{\sigma_1}(z-\sigma_2 t - m)t + t^2\right]}.$$

$Z = \sigma_1 X + \sigma_2 Y + m$ keresett sűrűségfüggvénye (ún. peremsűrűségfüggvény) ennek t szerinti integrálásával így kapható:

$$g_1(z) = \frac{1}{2\pi\sigma_1\sqrt{1-r^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-r^2)}\left[\frac{1}{\sigma_1^2}(z-\sigma_2 t - m)^2 - \frac{2r}{\sigma_1}(z-\sigma_2 t - m)t + t^2\right]} dt.$$

Két normál eloszlású változó összegének ez a sűrűségfüggvénye. Ha a két változó, $\sigma_1 X + m_1$ és $\sigma_2 X + m_2$ függetlenek, akkor X és Y és független, korrelációs együtthatójuk 0. Ekkor a sűrűségfüggvényük egyszerűbb alakra hozható. Az $r=0$ érték behelyettesítésével:

$$g_1(z) = \frac{1}{2\pi\sigma_1} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{1}{\sigma_1^2}(z-m-\sigma_2 t)^2 + t^2\right]} dt.$$

Emeljük ki az integrál elé a t-től független $e^{-\frac{(z-m)^2}{2(\sigma_1^2+\sigma_2^2)}}$-et, így az integráljelen belül e kitevőjében egy teljes négyzet alakítható ki:

$$g_1(z) = \frac{1}{2\pi\sigma_1} e^{-\frac{(z-m)^2}{2(\sigma_1^2+\sigma_2^2)}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{\sigma_1^2+\sigma_2^2-\sigma_1^2}{\sigma_1^2(\sigma_1^2+\sigma_2^2)}(z-m)^2 - \frac{2\sigma_2}{\sigma_1}(z-m)t + \frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2}t^2\right]} dt =$$

$$= \frac{1}{2\pi\sigma_1} e^{-\frac{(z-m)^2}{2(\sigma_1^2+\sigma_2^2)}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{\sigma_2}{\sigma_1\sqrt{\sigma_1^2+\sigma_2^2}}(z-m) - \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1}t\right]^2} dt.$$

Az $u = \frac{\sigma_2}{\sigma_1\sqrt{\sigma_1^2+\sigma_2^2}}(z-m) - \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1}t$ helyettesítés esetén $du =$

$$= -\frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1}dt, \text{ így } \left(\text{felhasználva, hogy } \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{u^2}{2}}du = 1\right):$$

$$g_1(z) = \frac{1}{2\pi\sigma_1} e^{-\frac{(z-m)^2}{2(\sigma_1^2+\sigma_2^2)}} \frac{\sigma_1}{\sqrt{\sigma_1^2+\sigma_2^2}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du =$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}} e^{-\frac{(z-m)^2}{2(\sigma_1^2+\sigma_2^2)}}.$$

Ez az eredmény azt jelenti, hogy *független, $N(m_1, \sigma_1)$- és $N(m_2, \sigma_2)$-eloszlású változók összege maga is normál eloszlású, $m = m_1+m_2$ várható értékkel, $\sigma^2 = \sigma_1^2+\sigma_2^2$ szórásnégyzettel.*

Megjegyzés

A most kapott ismert tételt az ún. karakterisztikus függvénnyel szokták bizonyítani. Ebből az előző, *b)* alatti állítás is megkapható.

TÁBLÁZATOK

Binomiális eloszlás $P(X = k)$

$$\binom{n}{k} p^k (1-p)^{n-k}$$

1. táblázat

	p									
n k	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1 0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
1	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2 0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
2	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3 0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4 0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5 0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0312
1	0,2036	0,3280	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1562
2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
4	0,0000	0,0004	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1562
5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0312
6 0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
2	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
3	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7 0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
2	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734

n k	p: 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8 0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
1	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0312
2	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0312
8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039
9 0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10 0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
11 0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
1	0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054
2	0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0513	0,0269
3	0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806
4	0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611

n k	p: 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
5	0,0001	0,0025	0,0132	0,0388	0,0803	0,1321	0,1830	0,2207	0,2360	0,2256
6	0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256
7	0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611
8	0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806
9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269
10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005
12 0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
1	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
2	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
3	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
4	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
5	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
6	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
7	0,0000	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934
8	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
9	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0068	0,0161
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
13 0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
1	0,3512	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016
2	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095
3	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349
4	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873
5	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571
6	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095
7	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095
8	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571
9	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873
10	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095
12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016
13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
14 0	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
1	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009
2	0,1229	0,2570	0,2912	0,2501	0,1802	0,1134	0,0634	0,0317	0,0141	0,0056
3	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222
4	0,0037	0,0349	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611

n	k	p = 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	5	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222
	6	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833
	7	0,0000	0,0002	0,0019	0,0092	0,0280	0,0618	0,1082	0,1574	0,1952	0,2095
	8	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833
	9	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222
	10	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
	2	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
	3	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
	4	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1268	0,0780	0,0417
	5	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916
	6	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
	7	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1964
	8	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
	9	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527
	10	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
16	0	0,4401	0,1853	0,0743	0,0281	0,0100	0,0033	0,0010	0,0003	0,0001	0,0000
	1	0,3706	0,3294	0,2097	0,1126	0,0535	0,0228	0,0087	0,0030	0,0009	0,0002
	2	0,1463	0,2745	0,2775	0,2111	0,1336	0,0732	0,0353	0,0150	0,0056	0,0018
	3	0,0359	0,1423	0,2285	0,2463	0,2079	0,1465	0,0888	0,0468	0,0215	0,0085
	4	0,0061	0,0514	0,1311	0,2001	0,2252	0,2040	0,1553	0,1014	0,0572	0,0278
	5	0,0008	0,0137	0,0555	0,1201	0,1802	0,2099	0,2008	0,1623	0,1123	0,0667
	6	0,0001	0,0028	0,0180	0,0550	0,1101	0,1649	0,1982	0,1983	0,1684	0,1222
	7	0,0000	0,0004	0,0045	0,0197	0,0524	0,1010	0,1524	0,1889	0,1969	0,1746
	8	0,0000	0,0001	0,0009	0,0055	0,0197	0,0487	0,0923	0,1417	0,1812	0,1964
	9	0,0000	0,0000	0,0001	0,0012	0,0058	0,0185	0,0442	0,0840	0,1318	0,1746
	10	0,0000	0,0000	0,0000	0,0002	0,0014	0,0056	0,0167	0,0392	0,0755	0,1222
	11	0,0000	0,0000	0,0000	0,0000	0,0002	0,0013	0,0049	0,0142	0,0337	0,0667
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0040	0,0115	0,0278
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0029	0,0085
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018

n	k	p = 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
17	0	0,4181	0,1668	0,0631	0,0225	0,0075	0,0023	0,0007	0,0002	0,0000	0,0000
	1	0,3741	0,3150	0,1893	0,0957	0,0426	0,0169	0,0060	0,0019	0,0005	0,0001
	2	0,1575	0,2800	0,2673	0,1914	0,1136	0,0581	0,0260	0,0102	0,0035	0,0010
	3	0,0415	0,1556	0,2359	0,2393	0,1893	0,1245	0,0701	0,0341	0,0144	0,0052
	4	0,0076	0,0605	0,1457	0,2093	0,2209	0,1868	0,1320	0,0796	0,0411	0,0182
	5	0,0010	0,0175	0,0668	0,1361	0,1914	0,2081	0,1849	0,1379	0,0875	0,0472
	6	0,0001	0,0039	0,0236	0,0680	0,1276	0,1784	0,1991	0,1839	0,1432	0,0944
	7	0,0000	0,0007	0,0065	0,0267	0,0668	0,1201	0,1685	0,1927	0,1841	0,1484
	8	0,0000	0,0001	0,0014	0,0084	0,0279	0,0644	0,1134	0,1606	0,1883	0,1855
	9	0,0000	0,0000	0,0003	0,0021	0,0093	0,0276	0,0611	0,1070	0,1540	0,1855
	10	0,0000	0,0000	0,0000	0,0004	0,0025	0,0095	0,0263	0,0571	0,1008	0,1484
	11	0,0000	0,0000	0,0000	0,0001	0,0005	0,0026	0,0090	0,0242	0,0525	0,0944
	12	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0081	0,0215	0,0472
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0021	0,0068	0,0182
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
18	0	0,3972	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
	1	0,3763	0,3002	0,1704	0,0811	0,0338	0,0126	0,0042	0,0012	0,0003	0,0001
	2	0,1683	0,2835	0,2556	0,1723	0,0958	0,0458	0,0190	0,0069	0,0022	0,0006
	3	0,0473	0,1680	0,2406	0,2297	0,1704	0,1046	0,0547	0,0246	0,0095	0,0031
	4	0,0093	0,0700	0,1592	0,2153	0,2130	0,1681	0,1104	0,0614	0,0291	0,0117
	5	0,0014	0,0218	0,0787	0,1507	0,1988	0,2017	0,1664	0,1146	0,0666	0,0327
	6	0,0002	0,0052	0,0301	0,0816	0,1436	0,1873	0,1941	0,1655	0,1181	0,0708
	7	0,0000	0,0010	0,0091	0,0350	0,0820	0,1376	0,1792	0,1892	0,1657	0,1214
	8	0,0000	0,0002	0,0022	0,0120	0,0376	0,0811	0,1327	0,1734	0,1864	0,1669
	9	0,0000	0,0000	0,0004	0,0033	0,0139	0,0386	0,0794	0,1284	0,1694	0,1855
	10	0,0000	0,0000	0,0001	0,0008	0,0042	0,0149	0,0385	0,0771	0,1248	0,1669
	11	0,0000	0,0000	0,0000	0,0001	0,0010	0,0046	0,0151	0,0374	0,0742	0,1214
	12	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0047	0,0145	0,0354	0,0708
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0045	0,0134	0,0327
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0039	0,0117
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0031
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

n k	p: 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
19 0	0,3774	0,1351	0,0456	0,0144	0,0042	0,0011	0,0003	0,0001	0,0000	0,0000
1	0,3774	0,2852	0,1529	0,0685	0,0268	0,0093	0,0029	0,0008	0,0002	0,0000
2	0,1787	0,2852	0,2428	0,1540	0,0803	0,0358	0,0138	0,0046	0,0013	0,0003
3	0,0533	0,1796	0,2428	0,2182	0,1517	0,0869	0,0422	0,0175	0,0062	0,0018
4	0,0112	0,0798	0,1714	0,2182	0,2023	0,1491	0,0909	0,0467	0,0203	0,0074
5	0,0018	0,0266	0,0907	0,1636	0,2023	0,1916	0,1468	0,0933	0,0497	0,0222
6	0,0002	0,0069	0,0374	0,0955	0,1574	0,1916	0,1844	0,1451	0,0949	0,0518
7	0,0000	0,0014	0,0122	0,0443	0,0974	0,1525	0,1844	0,1797	0,1443	0,0961
8	0,0000	0,0002	0,0032	0,0166	0,0487	0,0981	0,1489	0,1797	0,1771	0,1442
9	0,0000	0,0000	0,0007	0,0051	0,0198	0,0514	0,0980	0,1464	0,1771	0,1762
10	0,0000	0,0000	0,0001	0,0013	0,0066	0,0220	0,0528	0,0976	0,1449	0,1762
11	0,0000	0,0000	0,0000	0,0003	0,0018	0,0077	0,0233	0,0532	0,0970	0,1442
12	0,0000	0,0000	0,0000	0,0000	0,0004	0,0022	0,0083	0,0237	0,0529	0,0961
13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0024	0,0085	0,0233	0,0518
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0082	0,0222
15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0022	0,0074
16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003
18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20 0	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
1	0,3774	0,2702	0,1368	0,0576	0,0211	0,0068	0,0020	0,0005	0,0001	0,0000
2	0,1887	0,2852	0,2293	0,1369	0,0669	0,0278	0,0100	0,0031	0,0008	0,0002
3	0,0596	0,1901	0,2428	0,2054	0,1339	0,0716	0,0323	0,0123	0,0040	0,0011
4	0,0133	0,0898	0,1821	0,2182	0,1897	0,1304	0,0738	0,0350	0,0139	0,0046
5	0,0022	0,0319	0,1028	0,1746	0,2023	0,1789	0,1272	0,0746	0,0365	0,0148
6	0,0003	0,0089	0,0454	0,1091	0,1686	0,1916	0,1712	0,1244	0,0746	0,0370
7	0,0000	0,0020	0,0160	0,0545	0,1124	0,1643	0,1844	0,1659	0,1221	0,0739
8	0,0000	0,0004	0,0046	0,0222	0,0609	0,1144	0,1614	0,1797	0,1623	0,1201
9	0,0000	0,0001	0,0011	0,0074	0,0271	0,0654	0,1158	0,1597	0,1771	0,1602
10	0,0000	0,0000	0,0002	0,0020	0,0099	0,0308	0,0686	0,1171	0,1593	0,1762
11	0,0000	0,0000	0,0000	0,0005	0,0030	0,0120	0,0336	0,0710	0,1185	0,1602
12	0,0000	0,0000	0,0000	0,0001	0,0008	0,0039	0,0136	0,0355	0,0727	0,1201
13	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0045	0,0146	0,0366	0,0739
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0049	0,0150	0,0370
15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0049	0,0148
16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0046
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011
18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Binomiális eloszlás $\Sigma\ P(X=k)$

A binomiális eloszlásnál szereplő valószínűségek összegének,
a $\sum_{k=0}^{k_{max}} \binom{n}{k} p^k(1-p)^{n-k}$ összegnek az értékei
(pl. az n darabból álló mintában k_{max} vagy kevesebb hibás darab előfordulásának a valószínűségei)

2. táblázat

A tételben levő hibás darabok részaránya, p	A tételben levő hibás db száma max, k_{max}	A mintában levő összes darab száma, n: 5	6	7	8	9	10	15	20	25
0,005	0	0,9752	0,9704	0,9655	0,9607	0,9559	0,9511	0,9276	0,9046	0,8882
	1	0,9998	0,9996	0,9995	0,9993	0,9991	0,9989	0,9975	0,9955	0,9931
	2	–	–	–	–	–	–	0,9999	0,9999	0,9997
0,01	0	0,9510	0,9415	0,9321	0,9227	0,9135	0,9044	0,8601	0,8179	0.7778
	1	0,9990	0,9985	0,9980	0,9973	0,9966	0,9957	0,9904	0,9831	0,9742
	2	–	–	–	0,9999	0,9999	0,9999	0,9996	0,9990	0,9980
	3	–	–	–	–	–	–	–	–	0,9999
0,02	0	0,9039	0,8858	0,8681	0,8508	0,8337	0,8171	0,7386	0,6676	0.6035
	1	0,9962	0,9943	0,9921	0,9897	0,9869	0,9838	0,9647	0,9401	0,9114
	2	0,9999	0,9998	0,9997	0,9996	0,9994	0,9991	0,9970	0,9929	0,9868
	3	–	–	–	–	–	–	0,9998	0,9994	0,9986
	4	–	–	–	–	–	–	–	–	0,9999
0,03	0	0,8587	0,8330	0,8080	0,7837	0,7602	0,7374	0,6333	0,5438	0.4670
	1	0,9915	0,9875	0,9827	0,9777	0,9718	0,9655	0,9270	0,8802	0,8280
	2	0,9997	0,9995	0,9991	0,9986	0,9980	0,9972	0,9906	0,9790	0,9620
	3	–	–	–	0,9999	0,9999	0,9999	0,9992	0,9973	0,9938
	4	–	–	–	–	–	–	0,9999	0,9997	0,9992
	5	–	–	–	–	–	–	–	–	0,9999
0,04	0	0,8154	0,7827	0,7514	0,7214	0,6925	0,6648	0,5421	0,4420	0,3604
	1	0,9852	0,9784	0,9706	0,9619	0,9522	0,9418	0,8809	0,8103	0,7358
	2	0,9994	0,9988	0,9980	0,9969	0,9955	0,9938	0,9797	0,9561	0,9235
	3	–	–	0,9999	0,9998	0,9997	0,9996	0,9976	0,9926	0,9835
	4	–	–	–	–	–	–	0,9998	0,9990	0,9972
	5	–	–	–	–	–	–	–	0,9999	0,9996

A tételben levő hibás darabok részaránya, p	A tételben levő hibás db száma max, k_{max}	A mintában levő összes darab száma, n								
		5	6	7	8	9	10	15	20	25
0,05	0	0,7738	0,7351	0,6983	0,6634	0,6302	0,5987	0,4633	0,3585	0,2774
	1	0,9774	0,9672	0,9556	0,9428	0,9288	0,9139	0,8290	0,7358	0,6424
	2	0,9988	0,9978	0,9962	0,9942	0,9916	0,9885	0,9638	0,9245	0,8729
	3	–	0,9999	0,9998	0,9996	0,9994	0,9990	0,9945	0,9841	0,9659
	4	–	–	–	–	–	0,9999	0,9994	0,9974	0,9928
	5			–	–	–	–	0,9999	0,9997	0,9988
	6	–	–	–	–	–	–	–	–	0,9998
0,06	0	0,7339	0,6899	0,6485	0,6096	0,5730	0,5386	0,3953	0,2901	0,2129
	1	0,9681	0,9541	0,9382	0,9208	0,9022	0,8824	0,7738	0,6605	0,5527
	2	0,9980	0,9962	0,9937	0,9904	0,9862	0,9812	0,9429	0,8850	0,8129
	3	0,9999	0,9998	0,9996	0,9993	0,9987	0,9980	0,9896	0,9710	0,9402
	4	–	–	–	–	0,9999	0,9998	0,9986	0,9944	0,9850
	5	–	–	–	–		–	0,9998	0,9991	0,9969
	6	–	–	–	–	–	–	–	0,9999	0,9995
	7	–	–	–	–	–	–	–	–	0,9999
0,07	0	0,6957	0,6470	0,6017	0,5596	0,5204	0,4840	0,3367	0,2342	0,1630
	1	0,9575	0,9392	0,9187	0,8965	0,8729	0,8483	0,7168	0,5869	0,4696
	2	0,9969	0,9942	0,9903	0,9853	0,9791	0,9717	0,9171	0,8390	0,7466
	3	0,9999	0,9997	0,9993	0,9987	0,9977	0,9964	0,9825	0,9529	0,9064
	4	–	–	–	0,9999	0,9998	0,9997	0,9972	0,9893	0,9726
	5	–	–	–	–	–	–	0,9997	0,9981	0,9935
	6	–	–	–	–	–	–	–	0,9997	0,9987
	7	–	–	–	–	–	–	–	–	0,9998
0,08	0	0,6591	0,6064	0,5578	0,5132	0,4722	0,4344	0,2863	0,1887	0,1244
	1	0,9456	0,9227	0,8974	0,8702	0,8417	0,8121	0,6597	0,5169	0,3947
	2	0,9955	0,9915	0,9860	0,9789	0,9702	0,9599	0,8870	0,7879	0,6768
	3	0,9998	0,9995	0,9988	0,9978	0,9963	0,9942	0,9927	0,9294	0,8649
	4	–	–	0,9999	0,9999	0,9997	0,9994	0,9950	0,9817	0,9549
	5	–	–	–	–	–	–	0,9993	0,9962	0,9877
	6	–	–	–	–	–	–	0,9999	0,9994	0,9972
	7	–	–	–	–	–	–	–	0,9999	0,9995
	8	–	–	–	–	–	–	–	–	0,9999

A tételben levő hibás darabok részaránya, p	A tételben levő hibás db száma max, k_{max}	A mintában levő összes darab száma, n								
		5	6	7	8	9	10	15	20	25
0,09	0	0,6240	0,5679	0,5168	0,4703	0,4279	0,3984	0,2430	0,1516	0,0946
	1	0,9326	0,9048	0,8745	0,8423	0,8088	0,7746	0,6035	0,4516	0,3286
	2	0,9937	0,9882	0,9807	0,9711	0,9595	0,9460	0,8531	0,7334	0,6063
	3	0,9997	0,9992	0,9982	0,9966	0,9942	0,9912	0,9601	0,9007	0,8169
	4	–	–	0,9999	0,9997	0,9995	0,9990	0,9918	0,9710	0,9314
	5	–	–	–	–	–	0,9999	0,9987	0,9932	0,9790
	6	–	–	–	–	–	–	0,9998	0,9987	0,9846
	7	–	–	–	–	–	–	–	0,9998	0,9989
	8	–	–	–	–	–	–	–	–	0,9998
0,10	0	0,5905	0,5314	0,4783	0,4305	0,3874	0,3487	0,2059	0,1216	0,0718
	1	0,9185	0,8857	0,8503	0,8131	0,7748	0,7361	0,5490	0,3917	0,2712
	2	0,9914	0,9842	0,9743	0,9619	0,9470	0,9298	0,8159	0,6769	0,5371
	3	0,9995	0,9987	0,9973	0,9950	0,9917	0,9872	0,9444	0,8670	0,7636
	4	–	0,9999	0,9998	0,9996	0,9991	0,9984	0,9973	0,9568	0,9020
	5	–	–	–	–	0,9999	0,9999	0,9978	0,9887	0,9666
	6	–	–	–	–	–	–	0,9997	0,9976	0,9905
	7	–	–	–	–	–	–	–	0,9996	0,9977
	8	–	–	–	–	–	–	–	0,9999	0,9995
	9	–	–	–	–	–	–	–	–	0,9999
0,15	0	0,4437	0,3771	0,3206	0,2725	0,2317	0,1969	0,0874	0,0388	0,0172
	1	0,8352	0,7765	0,7166	0,6572	0,5995	0,5443	0,3186	0,1756	0,0931
	2	0,9734	0,9527	0,9262	0,8948	0,8591	0,8202	0,6042	0,4049	0,2537
	3	0,9978	0,9941	0,9879	0,9786	0,9661	0,9500	0,8227	0,6477	0,4711
	4	0,9999	0,9996	0,9988	0,9971	0,9944	0,9901	0,9383	0,8298	0,6821
	5	–	–	0,9999	0,9998	0,9994	0,9986	0,9832	0,9327	0,8385
	6	–	–	–	–	–	0,9999	0,9964	0,9781	0,9305
	7	–	–	–	–	–	–	0,9994	0,9941	0,9745
	8	–	–	–	–	–	–	0,9999	0,9987	0,9920
	9	–	–	–	–	–	–	–	0,9998	0,9979
	10	–	–	–	–	–	–	–	–	0,9995
	11	–	–	–	–	–	–	–	–	0,9999

A tételben levő hibás darabok részaránya, p	A tételben levő hibás db száma max, k_{max}	A mintában levő összes darab száma, n								
		5	6	7	8	9	10	15	20	25
	0	0,3277	0,2621	0,2097	0,1673	0,1342	0,1074	0,0352	0,0115	0,0038
	1	0,7373	0,6554	0,5767	0,5033	0,4362	0,3758	0,1671	0,0692	0,0274
	2	0,9421	0,9011	0,8520	0,7969	0,7382	0,6778	0,3980	0,2061	0,0982
	3	0,9933	0,9830	0,9667	0,9437	0,9144	0,8791	0,6482	0,4114	0,2340
	4	0,9997	0,9984	0,9953	0,9896	0,9804	0,9672	0,8358	0,6296	0,4207
	5	–	0,9999	0,9996	0,9988	0,9969	0,9936	0,9389	0,8042	0,6167
0,20	6	–	–	–	0,9999	0,9997	0,9991	0,9819	0,9133	0,7800
	7	–	–	–	–		0,9999	0,9958	0,9679	0,8909
	8	–	–	–	–	–	–	0,9992	0,9900	0,9532
	9	–	–	–	–	–	–	0,9999	0,9974	0,9827
	10	–	–	–	–	–	–	–	0,9994	0,9944
	11	–	–	–	–	–	–	–	0,9999	0,9985
	12	–	–	–	–	–	–	–	–	0,9996
	13	–	–	–	–	–	–	–	–	0,9999

Poisson-eloszlás $P(X=k)$

$$e^{-\lambda}\frac{\lambda^k}{k!} \qquad (\lambda=np)$$

3. táblázat

k	λ = 0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0000	0,0001	0,0002	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020	0,0031
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

k	λ = 1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496	0,1353
1	0,3662	0,3614	0,3543	0,3452	0,3347	0,3230	0,3106	0,2975	0,2842	0,2707
2	0,2014	0,2169	0,2303	0,2417	0,2510	0,2584	0,2640	0,2678	0,2700	0,2707
3	0,0738	0,0867	0,0998	0,1128	0,1255	0,1378	0,1496	0,1607	0,1710	0,1804
4	0,0203	0,0260	0,0324	0,0395	0,0471	0,0551	0,0636	0,0723	0,0812	0,0902
5	0,0045	0,0062	0,0084	0,0111	0,0141	0,0176	0,0216	0,0260	0,0309	0,0361
6	0,0008	0,0012	0,0018	0,0026	0,0035	0,0047	0,0061	0,0078	0,0098	0,0120
7	0,0001	0,0002	0,0003	0,0005	0,0008	0,0011	0,0015	0,0020	0,0027	0,0034
8	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002	0,0003	0,0005	0,0006	0,0009
9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002

k	λ = 2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608	0,0550	0,0498
1	0,2572	0,2438	0,2306	0,2177	0,2052	0,1931	0,1815	0,1703	0,1596	0,1494
2	0,2700	0,2681	0,2652	0,2613	0,2565	0,2510	0,2450	0,2384	0,2314	0,2240
3	0,1890	0,1966	0,2033	0,2090	0,2138	0,2176	0,2205	0,2225	0,2237	0,2240
4	0,0938	0,1082	0,1169	0,1254	0,1336	0,1414	0,1488	0,1557	0,1622	0,1680

k	λ = 2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
5	0,0417	0,0476	0,0538	0,0602	0,0668	0,0735	0,0804	0,0872	0,0940	0,1008
6	0,0146	0,0174	0,0206	0,0241	0,0278	0,0319	0,0362	0,0407	0,0455	0,0504
7	0,0044	0,0055	0,0068	0,0083	0,0099	0,0118	0,0139	0,0163	0,0188	0,0216
8	0,0011	0,0015	0,0019	0,0025	0,0031	0,0038	0,0047	0,0057	0,0068	0,0081
9	0,0003	0,0004	0,0005	0,0007	0,0009	0,0011	0,0014	0,0018	0,0022	0,0027
10	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0004	0,0005	0,0006	0,0008
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002	0,0002
12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

k	λ = 3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
0	0,0450	0,0408	0,0369	0,0334	0,0302	0,0273	0,0247	0,0224	0,0202	0,0183
1	0,1397	0,1304	0,1217	0,1135	0,1057	0,0984	0,0915	0,0850	0,0789	0,0733
2	0,2165	0,2087	0,2008	0,1929	0,1850	0,1771	0,1692	0,1615	0,1539	0,1465
3	0,2237	0,2226	0,2209	0,2186	0,2158	0,2125	0,2087	0,2046	0,2001	0,1954
4	0,1734	0,1781	0,1823	0,1858	0,1888	0,1912	0,1931	0,1944	0,1951	0,1954
5	0,1075	0,1140	0,1203	0,1264	0,1322	0,1377	0,1429	0,1477	0,1522	0,1563
6	0,0555	0,0608	0,0662	0,0716	0,0771	0,0826	0,0881	0,0936	0,0989	0,1042
7	0,0246	0,0278	0,0312	0,0348	0,0385	0,0425	0,0466	0,0508	0,0551	0,0595
8	0,0095	0,0111	0,0129	0,0148	0,0169	0,0191	0,0215	0,0241	0,0269	0,0298
9	0,0033	0,0040	0,0047	0,0056	0,0066	0,0076	0,0089	0,0102	0,0116	0,0132
10	0,0010	0,0013	0,0016	0,0019	0,0023	0,0028	0,0033	0,0039	0,0045	0,0053
11	0,0003	0,0004	0,0005	0,0006	0,0007	0,0009	0,0011	0,0013	0,0016	0,0019
12	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006
13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

k	λ = 4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
0	0,0166	0,0150	0,0136	0,0123	0,0111	0,0101	0,0091	0,0082	0,0074	0,0067
1	0,0679	0,0630	0,0583	0,0540	0,0500	0,0462	0,0427	0,0395	0,0365	0,0337
2	0,1393	0,1323	0,1254	0,1188	0,1125	0,1063	0,1005	0,0948	0,0894	0,0842
3	0,1904	0,1852	0,1798	0,1743	0,1687	0,1631	0,1574	0,1517	0,1460	0,1404
4	0,1951	0,1944	0,1933	0,1917	0,1898	0,1875	0,1849	0,1820	0,1789	0,1755

k	λ = 4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
5	0,1600	0,1633	0,1662	0,1687	0,1708	0,1725	0,1738	0,1747	0,1753	0,1755
6	0,1093	0,1143	0,1191	0,1237	0,1281	0,1323	0,1362	0,1398	0,1432	0,1462
7	0,0640	0,0686	0,0732	0,0778	0,0824	0,0869	0,0914	0,0959	0,1002	0,1044
8	0,0328	0,0360	0,0393	0,0428	0,0463	0,0500	0,0537	0,0575	0,0614	0,0653
9	0,0150	0,0168	0,0188	0,0209	0,0232	0,0255	0,0280	0,0307	0,0334	0,0363
10	0,0061	0,0071	0,0081	0,0092	0,0104	0,0118	0,0132	0,0147	0,0164	0,0181
11	0,0023	0,0027	0,0032	0,0037	0,0043	0,0049	0,0056	0,0064	0,0073	0,0082
12	0,0008	0,0009	0,0011	0,0014	0,0016	0,0019	0,0022	0,0026	0,0030	0,0034
13	0,0002	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013
14	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005
15	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002

k	λ = 5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	6,0
0	0,0061	0,0055	0,0050	0,0045	0,0041	0,0037	0,0033	0,0030	0,0027	0,0025
1	0,0311	0,0287	0,0265	0,0244	0,0225	0,0207	0,0191	0,0176	0,0162	0,0149
2	0,0793	0,0746	0,0701	0,0659	0,0618	0,0580	0,0544	0,0509	0,0477	0,0446
3	0,1348	0,1293	0,1239	0,1185	0,1133	0,1082	0,1033	0,0985	0,0938	0,0892
4	0,1719	0,1681	0,1641	0,1600	0,1558	0,1515	0,1472	0,1428	0,1383	0,1399
5	0,1752	0,1748	0,1740	0,1728	0,1714	0,1697	0,1678	0,1656	0,1632	0,1606
6	0,1490	0,1515	0,1537	0,1555	0,1571	0,1584	0,1594	0,1601	0,1605	0,1606
7	0,1086	0,1125	0,1163	0,1200	0,1234	0,1267	0,1298	0,1326	0,1353	0,1377
8	0,0692	0,0731	0,0771	0,0810	0,0849	0,0887	0,0925	0,0962	0,0998	0,1033
9	0,0392	0,0423	0,0454	0,0486	0,0519	0,0552	0,0586	0,0620	0,0654	0,0688
10	0,0200	0,0220	0,0241	0,0262	0,0285	0,0309	0,0334	0,0359	0,0386	0,0413
11	0,0093	0,0104	0,0116	0,0129	0,0143	0,0157	0,0173	0,0190	0,0207	0,0225
12	0,0039	0,0045	0,0051	0,0058	0,0065	0,0073	0,0082	0,0092	0,0102	0,0113
13	0,0015	0,0018	0,0021	0,0024	0,0028	0,0032	0,0036	0,0041	0,0046	0,0052
14	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013	0,0015	0,0017	0,0019	0,0022
15	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009
16	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001

	λ									
k	6,1	6,2	6,3	6,4	6,5	6,6	6,7	6,8	6,9	7,0
0	0,0022	0,0020	0,0018	0,0017	0,0015	0,0014	0,0012	0,0011	0,0010	0,0009
1	0,0137	0,0126	0,0116	0,0106	0,0098	0,0090	0,0082	0,0076	0,0070	0,0064
2	0,0417	0,0390	0,0364	0,0340	0,0318	0,0296	0,0276	0,0258	0,0240	0,0223
3	0,0848	0,0806	0,0765	0,0726	0,0688	0,0652	0,0617	0,0584	0,0552	0,0521
4	0,1294	0,1249	0,1205	0,1162	0,1118	0,1076	0,1034	0,0992	0,0952	0,0912
5	0,1579	0,1549	0,1519	0,1487	0,1454	0,1420	0,1385	0,1349	0,1314	0,1277
6	0,1605	0,1601	0,1595	0,1586	0,1575	0,1562	0,1546	0,1529	0,1511	0,1490
7	0,1399	0,1418	0,1435	0,1450	0,1462	0,1472	0,1480	0,1486	0,1489	0,1490
8	0,1066	0,1099	0,1130	0,1160	0,1188	0,1215	0,1240	0,1263	0,1284	0,1304
9	0,0723	0,0757	0,0791	0,0825	0,0858	0,0891	0,0923	0,0954	0,0985	0,1014
10	0,0441	0,0469	0,0498	0,0528	0,0558	0,0588	0,0618	0,0649	0,0679	0,0710
11	0,0245	0,0265	0,0285	0,0307	0,0330	0,0353	0,0377	0,0401	0,0426	0,0452
12	0,0124	0,0137	0,0150	0,0164	0,0179	0,0194	0,0210	0,0227	0,0245	0,0264
13	0,0058	0,0065	0,0073	0,0081	0,0089	0,0098	0,0108	0,0119	0,0130	0,0142
14	0,0025	0,0029	0,0033	0,0037	0,0041	0,0046	0,0052	0,0058	0,0064	0,0071
15	0,0010	0,0012	0,0014	0,0016	0,0018	0,0020	0,0023	0,0026	0,0029	0,0033
16	0,0004	0,0005	0,0005	0,0006	0,0007	0,0008	0,0010	0,0011	0,0013	0,0014
17	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006
18	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001

	λ									
k	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
0	0,0008	0,0007	0,0007	0,0006	0,0006	0,0005	0,0005	0,0004	0,0004	0,0003
1	0,0059	0,0054	0,0049	0,0045	0,0041	0,0038	0,0035	0,0032	0,0029	0,0027
2	0,0208	0,0194	0,0180	0,0167	0,0156	0,0145	0,0134	0,0125	0,0116	0,0107
3	0,0492	0,0464	0,0438	0,0413	0,0389	0,0366	0,0345	0,0324	0,0305	0,0286
4	0,0874	0,0836	0,0799	0,0764	0,0729	0,0696	0,0663	0,0632	0,0602	0,0573
5	0,1241	0,1204	0,1167	0,1130	0,1094	0,1057	0,1021	0,0986	0,0951	0,0916
6	0,1468	0,1445	0,1420	0,1394	0,1367	0,1339	0,1311	0,1282	0,1252	0,1221
7	0,1489	0,1486	0,1481	0,1474	0,1465	0,1454	0,1442	0,1428	0,1413	0,1396
8	0,1321	0,1337	0,1351	0,1363	0,1373	0,1382	0,1388	0,1392	0,1395	0,1396
9	0,1042	0,1070	0,1096	0,1121	0,1144	0,1167	0,1187	0,1207	0,1224	0,1241

	λ									
k	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
10	0,0740	0,0770	0,0800	0,0829	0,0858	0,0887	0,0914	0,0941	0,0967	0,0993
11	0,0478	0,0504	0,0531	0,0558	0,0585	0,0613	0,0640	0,0667	0,0695	0,0722
12	0,0283	0,0303	0,0323	0,0344	0,0366	0,0388	0,0411	0,0434	0,0457	0,0481
13	0,0154	0,0168	0,0181	0,0196	0,0211	0,0227	0,0243	0,0260	0,0278	0,0296
14	0,0078	0,0086	0,0095	0,0104	0,0113	0,0123	0,0134	0,0145	0,0157	0,0169
15	0,0037	0,0041	0,0046	0,0051	0,0057	0,0062	0,0069	0,0075	0,0083	0,0090
16	0,0016	0,0019	0,0021	0,0024	0,0026	0,0030	0,0033	0,0037	0,0041	0,0045
17	0,0007	0,0008	0,0009	0,0010	0,0012	0,0013	0,0015	0,0017	0,0019	0,0021
18	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
19	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0003	0,0004
20	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002
21	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001

	λ									
k	8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,9	9,0
0	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001
1	0,0025	0,0023	0,0021	0,0019	0,0017	0,0016	0,0014	0,0013	0,0012	0,0011
2	0,0100	0,0092	0,0086	0,0079	0,0074	0,0068	0,0063	0,0058	0,0054	0,0050
3	0,0269	0,0252	0,0237	0,0222	0,0208	0,0195	0,0183	0,0171	0,0160	0,0150
4	0,0544	0,0517	0,0491	0,0466	0,0443	0,0420	0,0398	0,0377	0,0357	0,0337
5	0,0882	0,0849	0,0816	0,0784	0,0752	0,0722	0,0692	0,0663	0,0635	0,0607
6	0,1191	0,1160	0,1128	0,1097	0,1066	0,1034	0,1003	0,0972	0,0941	0,0911
7	0,1378	0,1358	0,1338	0,1317	0,1294	0,1271	0,1247	0,1222	0,1197	0,1171
8	0,1395	0,1392	0,1388	0,1382	0,1375	0,1366	0,1356	0,1344	0,1332	0,1318
9	0,1256	0,1269	0,1280	0,1290	0,1299	0,1306	0,1311	0,1315	0,1317	0,1318
10	0,1017	0,1040	0,1063	0,1084	0,1104	0,1123	0,1140	0,1157	0,1172	0,1186
11	0,0749	0,0776	0,0802	0,0828	0,0853	0,0878	0,0902	0,0925	0,0948	0,0970
12	0,0505	0,0530	0,0555	0,0579	0,0604	0,0629	0,0654	0,0679	0,0703	0,0728
13	0,0315	0,0334	0,0354	0,0374	0,0395	0,0416	0,0438	0,0459	0,0481	0,0504
14	0,0182	0,0196	0,0210	0,0225	0,0240	0,0256	0,0272	0,0289	0,0306	0,0324
15	0,0098	0,0107	0,0116	0,0126	0,0136	0,0147	0,0158	0,0169	0,0182	0,0194
16	0,0050	0,0055	0,0060	0,0066	0,0072	0,0079	0,0086	0,0093	0,0101	0,0109
17	0,0024	0,0026	0,0029	0,0033	0,0036	0,0040	0,0044	0,0048	0,0053	0,0058
18	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019	0,0021	0,0024	0,0026	0,0029
19	0,0005	0,0005	0,0006	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014
20	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0005	0,0006
21	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003
22	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

k	λ = 9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10
0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0000
1	0,0010	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0005	0,0005	0,0005
2	0,0046	0,0043	0,0040	0,0037	0,0034	0,0031	0,0029	0,0027	0,0025	0,0023
3	0,0140	0,0131	0,0123	0,0115	0,0107	0,0100	0,0093	0,0087	0,0081	0,0076
4	0,0319	0,0302	0,0285	0,0269	0,0254	0,0240	0,0226	0,0213	0,0201	0,0189
5	0,0581	0,0555	0,0530	0,0506	0,0483	0,0460	0,0439	0,0418	0,0398	0,0378
6	0,0881	0,0851	0,0822	0,0793	0,0764	0,0736	0,0709	0,0682	0,0656	0,0631
7	0,1145	0,1118	0,1091	0,1064	0,1037	0,1010	0,0982	0,0955	0,0928	0,0901
8	0,1302	0,1286	0,1269	0,1251	0,1232	0,1212	0,1191	0,1170	0,1148	0,1126
9	0,1317	0,1315	0,1311	0,1306	0,1300	0,1293	0,1284	0,1274	0,1263	0,1251
10	0,1198	0,1210	0,1219	0,1228	0,1235	0,1241	0,1245	0,1249	0,1250	0,1251
11	0,0991	0,1012	0,1031	0,1049	0,1067	0,1083	0,1098	0,1112	0,1125	0,1137
12	0,0752	0,0776	0,0799	0,0822	0,0844	0,0866	0,0888	0,0908	0,0928	0,0948
13	0,0526	0,0549	0,0572	0,0594	0,0617	0,0640	0,0662	0,0685	0,0707	0,0729
14	0,0342	0,0361	0,0380	0,0399	0,0419	0,0439	0,0459	0,0479	0,0500	0,0521
15	0,0208	0,0221	0,0235	0,0250	0,0265	0,0281	0,0297	0,0313	0,0330	0,0347
16	0,0118	0,0127	0,0137	0,0147	0,0157	0,0168	0,0180	0,0192	0,0204	0,0217
17	0,0063	0,0069	0,0075	0,0081	0,0088	0,0095	0,0103	0,0111	0,0119	0,0128
18	0,0032	0,0035	0,0039	0,0042	0,0046	0,0051	0,0055	0,0060	0,0065	0,0071
19	0,0015	0,0017	0,0019	0,0021	0,0023	0,0026	0,0028	0,0031	0,0034	0,0037
20	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019
21	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
22	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004
23	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
24	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001

k	λ = 11	12	13	14	15	16	17	18	19	20
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0010	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3	0,0037	0,0018	0,0008	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000
4	0,0102	0,0053	0,0027	0,0013	0,0006	0,0003	0,0001	0,0001	0,0000	0,0000
5	0,0224	0,0127	0,0070	0,0037	0,0019	0,0010	0,0005	0,0002	0,0001	0,0001
6	0,0411	0,0255	0,0152	0,0087	0,0048	0,0026	0,0014	0,0007	0,0004	0,0002
7	0,0646	0,0437	0,0281	0,0174	0,0104	0,0060	0,0034	0,0018	0,0010	0,0005
8	0,0888	0,0655	0,0457	0,0304	0,0194	0,0120	0,0072	0,0042	0,0024	0,0013
9	0,1085	0,0874	0,0661	0,0473	0,0324	0,0213	0,0135	0,0083	0,0050	0,0029

k	λ = 11	12	13	14	15	16	17	18	19	20
10	0,1194	0,1048	0,0859	0,0663	0,0486	0,0341	0,0230	0,0150	0,0095	0,0058
11	0,1194	0,1144	0,1015	0,0844	0,0663	0,0496	0,0355	0,0245	0,0164	0,0106
12	0,1094	0,1144	0,1099	0,0984	0,0829	0,0661	0,0504	0,0368	0,0259	0,0176
13	0,0926	0,1056	0,1099	0,1060	0,0956	0,0814	0,0658	0,0509	0,0378	0,0271
14	0,0728	0,0905	0,1021	0,1060	0,1024	0,0930	0,0800	0,0655	0,0514	0,0387
15	0,0534	0,0724	0,0885	0,0989	0,1024	0,0992	0,0906	0,0786	0,0650	0,0516
16	0,0367	0,0543	0,0719	0,0866	0,0960	0,0992	0,0963	0,0884	0,0772	0,0646
17	0,0237	0,0383	0,0550	0,0713	0,0847	0,0934	0,0963	0,0935	0,0863	0,0760
18	0,0145	0,0256	0,0397	0,0554	0,0706	0,0830	0,0909	0,0936	0,0911	0,0844
19	0,0084	0,0161	0,0272	0,0409	0,0557	0,0699	0,0814	0,0887	0,0911	0,0888
20	0,0046	0,0097	0,0177	0,0286	0,0418	0,0559	0,0692	0,0798	0,0866	0,0888
21	0,0024	0,0055	0,0109	0,0191	0,0299	0,0426	0,0560	0,0684	0,0783	0,0846
22	0,0012	0,0030	0,0065	0,0121	0,0204	0,0310	0,0433	0,0560	0,0676	0,0769
23	0,0006	0,0016	0,0037	0,0074	0,0133	0,0216	0,0320	0,0438	0,0559	0,0669
24	0,0003	0,0008	0,0020	0,0043	0,0083	0,0144	0,0226	0,0328	0,0442	0,0557
25	0,0001	0,0004	0,0010	0,0024	0,0050	0,0092	0,0154	0,0237	0,0336	0,0446
26	0,0000	0,0002	0,0005	0,0013	0,0029	0,0057	0,0101	0,0164	0,0246	0,0343
27	0,0000	0,0001	0,0002	0,0007	0,0016	0,0034	0,0063	0,0109	0,0173	0,0254
28	0,0000	0,0000	0,0001	0,0003	0,0009	0,0019	0,0038	0,0070	0,0117	0,0181
29	0,0000	0,0000	0,0001	0,0002	0,0004	0,0011	0,0023	0,0044	0,0077	0,0125
30	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0013	0,0026	0,0049	0,0083
31	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0007	0,0015	0,0030	0,0054
32	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0004	0,0009	0,0018	0,0034
33	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0010	0,0020
34	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0012
35	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0007
36	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004
37	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
38	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
39	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

Poisson-eloszlás $\Sigma\, P(X=k)$

A Poisson-eloszlásnál szereplő valószínűségek összegének 1000-szerese, $1000 \sum_{k=0}^{k_{max}} \frac{e^{-\lambda}\lambda^k}{k!}$

(ahol $\lambda = np$)

4. táblázat

$\lambda=np$	k_{max}									
	0	1	2	3	4	5	6	7	8	9
0,02	980	1000								
0,04	961	999	1000							
0,06	942	998	1000							
0,08	923	997	1000							
0,10	905	995	1000							
0,15	861	990	999	1000						
0,20	819	982	999	1000						
0,25	779	974	998	1000						
0,30	741	963	996	1000						
0,35	705	951	994	1000						
0,40	670	938	992	999	1000					
0,45	638	925	989	999	1000					
0,50	607	910	986	998	1000					
0,55	577	894	982	998	1000					
0,60	549	878	977	997	1000					
0,65	522	861	972	996	999	1000				
0,70	497	844	966	994	999	1000				
0,75	472	827	959	993	999	1000				
0,80	449	809	953	991	999	1000				
0,85	427	791	945	989	998	1000				
0,90	407	772	937	987	998	1000				
0,95	387	754	929	984	997	1000				
1,00	368	736	920	981	996	999	1000			
1,1	333	699	900	974	995	999	1000			
1,2	301	663	879	966	992	998	1000			
1,3	273	627	857	957	989	998	1000			
1,4	247	592	833	946	986	997	999	1000		
1,5	223	558	809	934	981	996	999	1000		
1,6	202	525	783	921	976	994	999	1000		
1,7	183	493	757	907	970	992	998	1000		
1,8	165	463	731	891	964	990	997	999	1000	
1,9	150	434	704	875	956	987	997	999	1000	
2,0	135	406	677	857	947	983	995	999	1000	

4. táblázat 1. folytatása

$\lambda=np$	k_{max}									
	0	1	2	3	4	5	6	7	8	9
2,2	111	355	623	819	928	975	993	998	1000	
2,4	091	308	570	779	904	964	988	997	999	1000
2,6	074	267	518	736	877	951	983	995	999	1000
2,8	061	231	469	692	848	935	976	992	998	999
3,0	050	199	423	647	815	916	966	988	996	999
3,2	041	171	380	603	781	895	955	983	994	998
3,4	033	147	340	558	744	871	942	977	992	997
3,6	027	126	303	515	706	844	927	969	988	996
3,8	022	107	269	473	668	816	909	960	984	994
4,0	018	092	238	433	629	785	889	949	979	992
4,2	015	078	210	395	590	753	867	936	972	989
4,4	012	066	185	359	551	720	844	921	964	985
4,6	010	056	163	326	513	686	818	905	955	980
4,8	008	048	143	294	476	651	791	887	944	975
5,0	007	040	125	265	440	616	762	867	932	968
5,2	006	034	109	238	406	581	732	845	918	960
5,4	005	029	095	213	373	546	702	822	903	951
5,6	004	024	082	191	342	512	670	797	886	941
5,8	003	021	072	170	313	478	638	771	867	929
6,0	002	017	062	151	285	446	606	744	847	916
6,2	002	015	054	134	259	414	574	716	826	902
6,4	002	012	046	119	235	384	542	687	803	886
6,6	001	010	040	105	213	355	511	658	780	869
6,8	001	009	034	093	192	327	480	628	755	850
7,0	001	007	030	082	173	301	450	599	729	830
7,2	001	006	025	072	156	276	420	569	703	810
7,4	001	005	022	063	140	253	392	539	676	788
7,6	001	004	019	055	125	231	365	510	648	765
7,8	000	004	016	048	112	210	338	481	620	741
8,0	000	003	014	042	100	191	313	453	593	717
8,5	000	002	009	030	074	150	256	386	523	653
9,0	000	001	006	021	055	116	207	324	456	587
9,5	000	001	004	015	040	089	165	269	392	522
10,0	000	000	003	010	029	067	130	220	333	458

$\lambda = np$	k_{max}									
	0	1	2	3	4	5	6	7	8	9
10,5	000	000	002	007	021	050	102	179	279	397
11,0	000	000	001	005	015	038	079	143	232	341
11,5	000	000	001	003	011	028	060	114	191	289
12,0	000	000	001	002	008	020	046	090	155	242
12,5	000	000	000	002	005	015	035	070	125	201
13,0	000	000	000	001	004	011	026	054	100	166
13,5	000	000	000	001	003	008	019	041	079	135
14,0	000	000	000	000	002	006	014	032	062	109
14,5	000	000	000	000	001	004	010	024	048	088
15,0	000	000	000	000	001	003	008	018	037	070

$\lambda = np$	k_{max}									
	10	11	12	13	14	15	16	17	18	19
2,8	1000									
3,0	1000									
3,2	1000									
3,4	999	1000								
3,6	999	1000								
3,8	998	999	1000							
4,0	997	999	1000							
4,2	996	999	1000							
4,4	994	998	999	1000						
4,6	992	997	999	1000						
4,8	990	996	999	1000						
5,0	986	995	998	999	1000					
5,2	982	993	997	999	1000					
5,4	977	990	996	999	1000					
5,6	972	988	995	998	999	1000				
5,8	965	984	993	997	999	1000				
6,0	957	980	991	996	999	999	1000			
6,2	949	975	989	995	998	999	1000			
6,4	939	969	986	994	997	999	1000			
6,6	927	963	982	992	997	999	999	1000		
6,8	915	955	978	990	996	998	999	1000		
7,0	901	947	973	987	994	998	999	1000		
7,2	887	937	967	984	993	997	999	999	1000	
7,4	871	926	961	980	991	996	998	999	1000	
7,6	854	915	954	976	989	995	998	999	1000	
7,8	835	902	945	971	986	993	997	999	1000	

$\lambda = np$	k_{max}									
	10	11	12	13	14	15	16	17	18	19
8,0	816	888	936	966	983	992	996	998	999	
8,5	763	849	909	949	973	986	993	997	999	1000
9,0	706	803	876	926	959	978	989	995	998	999
9,5	645	752	836	898	940	967	982	991	996	999
10,0	583	697	792	864	917	951	973	986	993	998
10,5	521	639	742	825	888	932	960	978	988	994
11,0	460	579	689	781	854	907	944	968	982	991
11,5	402	520	633	733	815	878	924	954	974	986
12,0	347	462	576	682	772	844	899	937	963	979
12,5	297	406	519	628	725	806	869	916	948	969
13,0	252	353	463	573	675	764	835	890	930	957
13,5	211	304	409	518	623	718	798	861	908	942
14,0	176	260	358	464	570	669	756	827	883	923
14,5	145	220	311	413	518	619	711	790	853	901
15,0	118	185	268	363	466	568	664	749	819	875

$\lambda = np$	k_{max}									
	20	21	22	23	24	25	26	27	28	29
8,5	1000									
9,0	1000									
9,5	999	1000								
10,0	998	999	1000							
10,5	997	999	999	1000						
11,0	995	998	999	1000						
11,5	992	996	998	999	1000					
12,0	988	994	997	999	999	1000				
12,5	983	991	995	998	999	999	1000			
13,0	975	986	992	996	998	999	1000			
13,5	965	980	989	994	997	998	999	1000		
14,0	952	971	983	991	995	997	999	999	1000	
14,5	936	960	976	986	992	996	998	999	999	1000
15,0	917	947	967	981	989	994	997	998	999	1000

$\lambda = np$	k_{max}									
	4	5	6	7	8	9	10	11	12	13
16	000	001	004	010	022	043	077	127	193	275
17	000	001	002	005	013	026	049	085	135	201
18	000	000	001	003	007	015	030	055	092	143
19	000	000	001	002	004	009	018	035	061	098
20	000	000	000	001	002	005	011	021	039	066

$\lambda = np$	k_{max}									
	4	5	6	7	8	9	10	11	12	13
21	000	000	000	000	001	003	006	013	025	043
22	000	000	000	000	001	002	004	008	015	028
23	000	000	000	000	000	001	002	004	009	017
24	000	000	000	000	000	000	001	003	005	011
25	000	000	000	000	000	000	001	001	003	006

$\lambda = np$	k_{max}									
	14	15	16	17	18	19	20	21	22	23
16	368	467	566	659	742	812	868	911	942	963
17	281	371	468	564	655	736	805	861	905	937
18	208	287	375	469	562	651	731	799	855	899
19	150	215	292	378	469	561	647	725	793	849
20	105	157	221	297	381	470	559	644	721	787
21	072	111	163	227	302	384	471	558	640	716
22	048	077	117	169	232	306	387	472	556	637
23	031	052	082	123	175	238	310	389	472	555
24	020	034	056	087	128	180	243	314	392	473
25	012	022	038	060	092	134	185	247	318	394

$\lambda = np$	k_{max}									
	24	25	26	27	28	29	30	31	32	33
16	978	987	993	996	998	999	999	1000		
17	959	975	985	991	995	997	999	999	1000	
18	932	955	972	983	990	994	997	998	999	1000
19	893	927	951	969	980	988	993	996	998	999
20	843	888	922	948	966	978	987	992	995	997
21	782	838	883	917	944	963	976	985	991	994
22	712	777	832	877	913	940	959	973	983	989
23	635	708	772	827	873	908	936	956	971	981
24	554	632	704	768	823	868	904	932	953	969
25	473	553	629	700	763	818	863	900	929	950

$\lambda = np$	k_{max}									
	34	35	36	37	38	39	40	41	42	43
19	999	1000								
20	999	999	1000							
21	997	998	999	999	1000					
22	994	996	998	999	999	1000				
23	988	993	996	997	999	999	1000			
24	979	987	992	995	997	998	999	999	1000	
25	966	978	985	991	994	997	998	999	999	1000

Normál eloszlás

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \qquad \Phi(x) = \int_{-\infty}^{x} \varphi(u)\, du$$

5. táblázat

x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$
0,00	0,3989	0,5000	0,60	0,3332	0,7257	1,20	0,1942	0,8849
01	3989	5040	61	3312	7291	21	1919	8869
02	3989	5080	62	3292	7324	22	1895	8888
03	3988	5120	63	3271	7357	23	1872	8907
04	3986	5160	64	3251	7389	24	1849	8925
05	3984	5199	65	3230	7422	25	1826	8944
06	3982	5239	66	3209	7454	26	1804	8962
07	3980	5279	67	3187	7486	27	1781	8980
08	3977	5319	68	3166	7517	28	1758	8997
09	3973	5359	69	3144	7549	29	1736	9015
0,10	0,3970	0,5398	0,70	0,3123	0,7580	1,30	0,1714	0,9032
11	3965	5438	71	3101	7611	31	1691	9049
12	3961	5478	72	3079	7642	32	1669	9066
13	3956	5517	73	3056	7673	33	1647	9082
14	3951	5557	74	3034	7703	34	1626	9099
15	3945	5596	75	3011	7734	35	1604	9115
16	3939	5636	76	2989	7764	36	1582	9131
17	3932	5675	77	2966	7794	37	1561	9147
18	3925	5714	78	2943	7823	38	1539	9162
19	3918	5753	79	2920	7852	39	1518	9177
0,20	0,3910	0,5793	0,80	0,2897	0,7881	1,40	0,1497	0,9192
21	3902	5832	81	2874	7910	41	1476	9207
22	3894	5871	82	2850	7939	42	1456	9222
23	3885	5910	83	2827	7967	43	1435	9236
24	3876	5948	84	2803	7995	44	1415	9251
25	3867	5987	85	2780	8023	45	1394	9265
26	3857	6026	86	2756	8051	46	1374	9279
27	3847	6064	87	2732	8078	47	1354	9292
28	3836	6103	88	2709	8106	48	1334	9306
29	3825	6141	89	2685	8133	49	1315	9319
0,30	0,3814	0,6179	0,90	0,2661	0,8159	1,50	0,1295	0,9332
31	3802	6217	91	2637	8186	51	1276	9345
32	3790	6265	92	2613	8212	52	1257	9357
33	3778	6293	93	2589	8238	53	1238	9370
34	3765	6331	94	2565	8264	54	1219	9382

x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$
35	3752	6368	95	2541	8289	55	1200	9394
36	3739	6406	96	2516	8315	56	1182	9406
37	3725	6443	97	2492	8340	57	1163	9418
38	3712	6480	98	2468	8365	58	1145	9429
39	3697	6517	99	2444	8389	59	1127	9441
0,40	0,3683	0,6557	1,00	0,2420	0,8413	1,60	0,1109	0,9452
41	3668	6591	01	2396	8438	61	1092	9463
42	3653	6628	02	2371	8461	62	1074	9474
43	3637	6664	03	2347	8485	63	1057	9484
44	3621	6700	04	2323	8508	64	1040	9495
45	3605	6736	05	2299	8531	65	1023	9505
46	3589	6772	06	2275	8554	66	1006	9515
47	3572	6808	07	2251	8577	67	0989	9525
48	3555	6844	08	2227	8599	68	0973	9535
49	3538	6879	09	2203	8621	69	0957	9545
0,50	0,3521	0,6915	1,10	0,2179	0,8643	1,70	0,0940	0,9554
51	3503	6950	11	2155	8665	71	0925	9564
52	3485	6985	12	2131	8686	72	0909	9573
53	3467	7019	13	2107	8708	73	0893	9583
54	3448	7054	14	2083	8729	74	0878	9591
55	3429	7088	15	2059	8749	75	0863	9599
56	3410	7123	16	2036	8770	76	0848	9608
57	3391	7157	17	2012	8790	77	0833	9616
58	3372	7190	18	1989	8810	78	0818	9625
59	3352	7224	19	1965	8830	79	0804	9633
1,80	0,0790	0,9641	2,20	0,0355	0,9861	2,80	0,0079	0,9974
81	0775	9649	22	0339	9868	82	0075	9976
82	0761	9656	24	0325	9875	84	0071	9977
83	0748	9664	26	0310	9881	86	0067	9979
84	0734	9671	28	0297	9887	88	0063	9980
85	0721	9678	30	0283	9893	90	0060	9981
86	0707	9686	32	0270	9898	92	0056	9982
87	0694	9693	34	0258	9904	94	0053	9984
88	0681	9699	36	0246	9909	96	0050	9985
89	0669	9706	38	0235	9913	98	0047	9986

x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$
1,90	0,0656	0,9713	2,40	0,0224	0,9918	3,00	0,00443	0,99865
91	0644	9719	42	0213	9922	10	00327	99903
92	0632	9729	44	0203	9927	20	00238	99931
93	0620	9732	46	0194	9931	30	00172	99951
94	0608	9738	48	0184	9934	40	00123	99966
95	0596	9744	50	0175	9938	50	00087	99976
96	0584	9750	52	0167	9941	60	00061	99984
97	0573	9756	54	0158	9945	70	00042	99989
98	0562	9761	56	0151	9948	80	00029	99993
99	0551	9767	58	0143	9951	90	00020	99995
2,00	0,0540	0,9772	2,60	0,0136	0,9953	4,00	0,000134	0,999968
02	0519	9783	62	0129	9956	50	000016	999997
04	0498	9793	64	0122	9959	5,00	000002	999997
06	0478	9803	66	0116	9961			
08	0459	9812	68	0110	9963			
10	0440	9821	70	0104	9965			
12	0422	9830	72	0099	9967			
14	0404	9838	74	0093	9969			
16	0387	9846	76	0088	9971			
18	0371	9854	78	0084	9973			

e^x, e^{-x} értékei

6. táblázat

x	*e* hatványai							
	e^x	e^{-x}	x	e^x	e^{-x}	x	e^x	e^{-x}
0,00	1,0000	1,0000	0,50	1,6487	0,6065	1,0	2,718	0,3679
0,01	1,0101	0,9900	0,51	1,6653	0,6005	1,1	3,004	0,3329
0,02	1,0202	0,9802	0,52	1,6820	0,5945	1,2	3,320	0,3012
0,03	1,0305	0,9704	0,53	1,6989	0,5886	1,3	3,669	0,2725
0,04	1,0408	0,9608	0,54	1,7160	0,5827	1,4	4,055	0,2466
0,05	1,0513	0,9512	0,55	1,7333	0,5769	1,5	4,482	0,2231
0,06	1,0618	0,9418	0,56	1,7507	0,5712	1,6	4,953	0,2019
0,07	1,0725	0,9324	0,57	1,7683	0,5655	1,7	5,474	0,1827
0,08	1,0833	0,9231	0,58	1,7860	0,5599	1,8	6,050	0,1653
0,09	1,0942	0,9139	0,59	1,8040	0,5543	1,9	6,686	0,1496
0,10	1,1052	0,9048	0,60	1,8221	0,5488	2,0	7,389	0,1353
0,11	1,1163	0,8958	0,61	1,8404	0,5434	2,1	8,166	0,1225

	e hatványai							
x	e^x	e^{-x}	x	e^x	e^{-x}	x	e^x	e^{-x}
0,12	1,1275	0,8869	0,62	1,8589	0,5379	2,2	9,025	0,1108
0,13	1,1388	0,8781	0,63	1,8776	0,5326	2,3	9,974	0,1003
0,14	1,1503	0,8694	0,64	1,8965	0,5273	2,4	11,023	0,0907
0,15	1,1618	0,8607	0,65	1,9155	0,5220	2,5	12,182	0,0821
0,16	1,1735	0,8521	0,66	1,9348	0,5169	2,6	13,464	0,0743
0,17	1,1853	0,8437	0,67	1,9542	0,5117	2,7	14,880	0,0672
0,18	1,1972	0,8353	0,68	1,9739	0,5066	2,8	16,445	0,0608
0,19	1,2096	0,8270	0,69	1,9937	0,5016	2,9	18,174	0,0550
0,20	1,2214	0,8187	0,70	2,0138	0,4966	3,0	20,086	0,0498
0,21	1,2337	0,8106	0,71	2,0340	0,4916	3,1	22,20	0,0450
0,22	1,2461	0,8025	0,72	2,0544	0,4868	3,2	24,53	0,0408
0,23	1,2586	0,7945	0,73	2,0751	0,4819	3,3	27,11	0,0369
0,24	1,2713	0,7866	0,74	2,0960	0,4771	3,4	29,96	0,0334
0,25	1,2840	0,7788	0,75	2,1170	0,4724	3,5	33,12	0,0302
0,26	1,2969	0,7711	0,76	2,1383	0,4677	3,6	36,60	0,0273
0,27	1,3100	0,7634	0,77	2,1598	0,4630	3,7	40,45	0,0247
0,28	1,3231	0,7558	0,78	2,1815	0,4584	3,8	44,70	0,0224
0,29	1,3364	0,7483	0,79	2,2034	0,4538	3,9	49,40	0,0202
0,30	1,3497	0,7408	0,80	2,2255	0,4493	4,0	54,60	0,0183
0,31	1,3634	0,7334	0,81	2,2479	0,4449	4,1	60,34	0,0166
0,32	1,3771	0,7261	0,82	2,2705	0,4404	4,2	66,69	0,0150
0,33	1,3910	0,7189	0,83	2,2933	0,4360	4,3	73,70	0,0136
0,34	1,4050	0,7118	0,84	2,3164	0,4317	4,4	81,45	0,0123
0,35	1,4191	0,7047	0,85	2,3397	0,4274	4,5	90,02	0,0111
0,36	1,4333	0,6977	0,86	2,3632	0,4232	4,6	99,48	0,0101
0,37	1,4477	0,6907	0,87	2,3869	0,4190	4,7	109,95	0,0091
0,38	1,4623	0,6839	0,88	2,4109	0,4148	4,8	121,51	0,0082
0,39	1,4770	0,6771	0,89	2,4351	0,4107	4,9	134,29	0,0074
0,40	1,4918	0,6703	0,90	2,4596	0,4066	5,0	148,41	0,0067
0,41	1,5068	0,6637	0,91	2,4843	0,4025	5,1	164,0	0,0061
0,42	1,5220	0,6570	0,92	2,5093	0,3985	5,2	181,3	0,0055
0,43	1,5373	0,6505	0,93	2,5345	0,3946	5,3	200,3	0,0050
0,44	1,5527	0,6440	0,94	2,5600	0,3906	5,4	221,4	0,0045
0,45	1,5683	0,6376	0,95	2,5857	0,3867	5,5	244,7	0,0041
0,46	1,5841	0,6313	0,96	2,6117	0,3829	5,6	270,4	0,0037
0,47	1,6000	0,6250	0,97	2,6380	0,3791	5,7	298,9	0,0034
0,48	1,6161	0,6188	0,98	2,6645	0,3753	5,8	330,3	0,0030
0,49	1,6323	0,6126	0,99	2,6912	0,3716	5,9	365,0	0,0027
0,50	1,6487	0,6065	1,00	2,7183	0,3679	6,0	403,4	0,0025

χ^2-eloszlás

(χ^2 értékek a χ^2-próba alkalmazásához)

7. táblázat

Szabadság-fok	Valószínűség százalékban						
	90,0	95,0	97,5	99,0	99,5	99,9	99,95
1	2,71	3,84	5,02	6,63	7,88	10,8	12,1
2	4,61	5,99	7,38	9,21	10,6	13,8	15,2
3	6,25	7,81	9,35	11,3	12,8	16,3	17,7
4	7,78	9,49	11,1	13,3	14,9	18,5	20,0
5	9,24	11,1	12,8	15,1	16,7	20,5	22,1
6	10,6	12,6	14,4	16,8	18,5	22,5	24,1
7	12,0	14,1	16,0	18,5	20,3	24,3	26,0
8	13,4	15,5	17,5	20,1	22,0	26,1	27,9
9	14,7	16,9	19,0	21,7	23,6	27,9	29,7
10	16,0	18,3	20,5	23,2	25,2	29,6	31,4
11	17,3	19,7	21,9	24,7	26,8	31,3	33,1
12	18,5	21,0	23,3	26,2	28,3	32,9	34,8
13	19,8	22,4	24,7	27,7	29,8	34,5	36,5
14	21,1	23,7	26,1	29,1	31,3	36,1	38,1
15	22,3	25,0	27,5	30,6	32,8	37,7	39,7
16	23,5	26,3	28,8	32,0	34,3	39,3	41,3
17	24,8	27,6	30,2	33,4	35,7	40,8	42,9
18	26,0	28,9	31,5	34,8	37,2	42,3	44,4
19	27,2	30,1	32,9	36,2	38,6	43,8	46,0
20	28,4	31,4	34,2	37,6	40,0	45,3	47,5
21	29,6	32,7	35,5	38,9	41,4	46,8	49,0
22	30,8	33,9	36,8	40,3	42,8	48,3	50,5
23	32,0	35,2	38,1	41,6	44,2	49,7	52,0
24	33,2	36,4	39,4	43,0	45,6	51,2	53,5
25	34,4	37,7	40,6	44,3	46,9	52,6	54,9
26	35,6	38,9	41,9	45,6	48,3	54,1	56,4
27	36,7	40,1	43,2	47,0	49,6	55,5	57,9
28	37,9	41,3	44,5	48,3	51,0	56,9	59,3
29	39,1	42,6	45,7	49,6	52,3	58,3	60,7
30	40,3	43,8	47,0	50,9	53,7	59,7	62,2

t-eloszlás

Táblázat a t értékek kiszámítására adott %-ú statisztikai biztonság és adott n számú mérés esetén (Student-eloszlás)

8. táblázat

$n=f+1$	Statisztikai biztonság					
	80%	90%	95%	98%	99%	99,9%
2	3,078	6,314	12,706	31,821	63,657	636,619
3	1,886	2,920	4,303	6,965	9,925	31,598
4	1,638	2,353	3,182	4,541	5,841	12,941
5	1,533	2,132	2,776	3,747	4,604	8,610
6	1,476	2,015	2,571	3,365	4,032	6,859
7	1,440	1,943	2,447	3,143	3,707	5,959
8	1,415	1,895	2,365	2,998	3,499	5,405
9	1,397	1,860	2,306	2,896	3,355	5,041
10	1,383	1,833	2,262	2,821	3,250	4,781
11	1,372	1,812	2,228	2,764	3,169	4,587
12	1,363	1,796	2,201	2,718	3,106	4,437
13	1,356	1,782	2,179	2,681	3,055	4,318
14	1,350	1,771	2,160	2,650	3,012	4,221
15	1,345	1,761	2,145	2,624	2,977	4,140
16	1,341	1,753	2,131	2,602	2,947	4,073
17	1,337	1,746	2,120	2,583	2,921	4,015
18	1,333	1,740	2,110	2,567	2,898	3,965
19	1,330	1,734	2,101	2,552	2,878	3,922
20	1,328	1,729	2,093	2,539	2,861	3,883
21	1,325	1,725	2,086	2,528	2,845	3,850
22	1,323	1,721	2,080	2,518	2,831	3,819
23	1,321	1,717	2,074	2,508	2,819	3,792
24	1,319	1,714	2,069	2,500	2,807	3,767
25	1,318	1,711	2,064	2,492	2,797	3,745
26	1,316	1,708	2,060	2,485	2,787	3,725
27	1,315	1,706	2,056	2,479	2,779	3,707
28	1,314	1,703	2,052	2,473	2,771	3,690
29	1,313	1,701	2,048	2,467	2,763	3,674
30	1,311	1,699	2,045	2,462	2,756	3,659
31	1,310	1,697	2,042	2,457	2,750	3,646
41	1,303	1,684	2,021	2,423	2,704	3,551
61	1,296	1,671	2,000	2,390	2,660	3,460
121	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,291

F-eloszlás

Az F-próba kritikus értékei 95% biztonsági szinten (f_1 a nagyobbik empirikus szórásnégyzetű minta elemszámának, f_2 pedig a másik minta elemszámának eggyel kisebbített értéke)

9. táblázat

f_2	f_1								
	1	2	3	4	5	6	7	8	9
1	161	200	216	225	230	234	237	239	241
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,35
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24
32	4,15	3,30	2,90	2,67	2,51	2,40	2,32	2,25	2,19
36	4,11	3,26	2,86	2,63	2,48	2,36	2,28	2,21	2,15
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97
200	3,89	3,04	2,65	2,41	2,26	2,14	2,05	1,98	1,92
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88

f_2	f_1 = 10	12	16	20	24	30	50	100	∞
1	242	244	246	248	249	250	252	253	254
2	19,39	19,41	19,43	19,44	19,45	19,46	19,47	19,49	19,50
3	8,78	8,74	8,69	8,66	8,64	8,62	8,58	8,56	8,53
4	5,96	5,91	5,84	5,80	5,77	5,74	5,70	5,66	5,63
5	4,74	4,68	4,60	4,56	4,53	4,50	4,44	4,40	4,36
6	4,06	4,00	3,92	3,87	3,84	3,81	3,75	3,71	3,67
7	3,63	3,57	3,49	3,44	3,41	3,38	3,32	3,28	3,23
8	3,34	3,28	3,20	3,15	3,12	3,08	3,03	2,98	2,93
9	3,13	3,07	2,98	2,93	2,90	2,86	2,80	2,76	2,71
10	2,97	2,91	2,82	2,77	2,74	2,70	2,64	2,59	2,54
11	2,86	2,79	2,70	2,65	2,61	2,57	2,50	2,45	2,40
12	2,76	2,69	2,60	2,54	2,50	2,46	2,40	2,35	2,30
13	2,67	2,60	2,51	2,46	2,42	2,38	2,32	2,26	2,21
14	2,60	2,53	2,44	2,39	2,35	2,31	2,24	2,19	2,13
15	2,55	2,48	2,39	2,33	2,29	2,25	2,18	2,12	2,07
16	2,49	2,42	2,33	2,28	2,24	2,20	2,13	2,07	2,01
17	2,45	2,38	2,29	2,23	2,19	2,15	2,08	2,02	1,96
18	2,41	2,34	2,25	2,19	2,15	2,11	2,04	1,98	1,92
19	2,38	2,31	2,21	2,15	2,11	2,07	2,00	1,94	1,88
20	2,35	2,28	2,18	2,12	2,08	2,04	1,96	1,90	1,84
22	2,30	2,23	2,13	2,07	2,03	1,98	1,91	1,84	1,78
24	2,26	2,18	2,09	2,02	1,98	1,94	1,86	1,80	1,73
26	2,22	2,15	2,05	1,99	1,95	1,90	1,82	1,76	1,69
28	2,19	2,12	2,02	1,96	1,91	1,87	1,78	1,72	1,65
32	2,14	2,07	1,97	1,91	1,86	1,82	1,74	1,67	1,59
36	2,10	2,03	1,93	1,87	1,82	1,78	1,69	1,62	1,55
40	2,07	2,00	1,90	1,84	1,79	1,74	1,66	1,59	1,51
60	1,99	1,92	1,81	1,75	1,70	1,65	1,56	1,48	1,39
100	1,92	1,85	1,75	1,68	1,63	1,57	1,48	1,39	1,28
200	1,87	1,80	1,69	1,62	1,57	1,52	1,42	1,32	1,19
∞	1,83	1,75	1,64	1,57	1,52	1,46	1,35	1,24	1,00

A mérések számának megállapítása normál eloszlás esetén

Adott $\frac{q}{s^*}\left(=\frac{t}{\sqrt{n}}\right)$ esetén adott biztonsági szinthez szükséges mérések számának, n-nek az értékei 100-ig

10. táblázat

$f=n-1$	95% biztonság $\frac{t}{\sqrt{n}}$	99% biztonság $\frac{t}{\sqrt{n}}$	99,9% biztonság $\frac{t}{\sqrt{n}}$
1	12,71	7	636,62
2	3,04	3,73	22
3	1,86	2,3	7,5
4	1,39	1,83	4,3
5	1,15	1,52	3,08
6	1	1,34	2,44
7	0,9	1,18	2,05
8	0,82	1,07	1,78
9	0,75	1	1,58
10	0,706	0,94	1,45
11	0,665	0,885	1,34
12	0,63	0,836	1,25
13	0,60	0,79	1,17
14	0,575	0,763	1,1
15	0,55	0,730	1,05
16	0,53	0,704	1,01
17	0,512	0,680	0,96
18	0,496	0,657	0,926
19	0,48	0,637	0,892
20	0,467	0,617	0,862
21	0,455	0,603	0,833
22	0,442	0,576	0,810
23	0,432	0,560	0,78
24	0,415	0,556	0,766
25	0,412	0,548	0,748
26	0,404	0,535	0,730
27	0,393	0,522	0,713
28	0,387	0,514	0,695
29	0,38	0,502	0,680
30	0,373	0,428	0,668
40	0,318	0,378	0,562
50	0,283	0,344	0,489
60	0,259	0,316	0,458
70	0,238	0,295	0,411
80	0,223	0,288	0,382
90	0,21	0,278	0,360
100	0,198	0,262	0,339

Kolmogorov-féle egymintás próba

11. táblázat

n	0,95	0,99	n	0,95	0,99
			21	0,2827	0,3443
8	0,4543	0,5419	22	0,2809	0,3367
9	0,4300	0,5133	23	0,2749	0,3295
10	0,4093	0,4889	24	0,2693	0,3229
11	0,3912	0,4677	25	0,2640	0,3166
12	0,3754	0,4491	26	0,2591	0,3106
13	0,3614	0,4325	27	0,2544	0,3050
14	0,3489	0,4176	28	0,2499	0,2997
15	0,3376	0,4042	29	0,2457	0,2947
16	0,3273	0,3920	30	0,2417	0,2899
17	0,3180	0,3809	35	0,2243	0,2690
18	0,3094	0,3706	40	0,2101	0,2521
19	0,3014	0,3612	45	0,1984	0,2380
20	0,2941	0,3524	50	0,1884	0,2260

Kolmogorov–Szmirnov-féle kétmintás próba

12. táblázat

n	0,95	0,99	n	0,95	0,99	n	0,95	0,99
7	6	6	15	8	9	23	10	11
8	6	7	16	8	10	24	10	12
9	6	7	17	8	10	25	10	12
10	7	8	18	9	10	26	10	12
11	7	8	19	9	10	27	10	12
12	7	8	20	9	11	28	11	13
13	7	9	21	9	11	29	11	13
14	8	9	22	9	11	30	11	13

A Wilcoxon-féle kétmintás próba kritikus értékei 97,5%-os egyoldali, ill. 95%-os kétoldali szintre

13. táblázat

	m															
n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	77
14	13	17	22	26	31	36	40	45	50	55	59	64	69	74	78	83
15	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	17	22	28	34	39	45	51	57	63	69	75	81	87	93	99	105
18	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

A Wilcoxon-féle kétmintás próba kritikus értékei 99,5%-os egyoldali, ill. 99%-os kétoldali szintre

	m															
n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	7	10	13	17	20	24	27	31	34	38	42	45	49	53	57	60
14	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	10	15	19	24	29	34	39	44	49	55	60	65	70	75	81	86
18	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	12	17	22	28	33	39	45	51	57	63	69	74	81	87	93	99
20	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

A Kolmogorov-próba.
A $K(z)$ Kolmogorov-függvény értékei

14. táblázat

y	$K(y)$	y	$K(y)$	y	$K(y)$	y	$K(y)$	y	$K(y)$	y	$K(y)$
		0,61	0,1492	1,00	0,7300	1,40	0,9603	1,80	0,9^{2}693	2,20	0,9^{3}874
		0,62	0,1632	1,01	0,7406	1,41	0,9625	1,81	0,9^{2}715	2,21	0,9^{3}886
		0,63	0,1777	1,02	0,7508	1,42	0,9645	1,82	0,9^{2}735	2,22	0,9^{3}896
		0,64	0,1927	1,03	0,7608	1,43	0,9665	1,83	0,9^{2}753	2,23	0,9^{4}04
		0,65	0,2080	1,04	0,7704	1,44	0,9684	1,84	0,9^{2}770	2,24	0,9^{4}12
		0,66	0,2236	1,05	0,7798	1,45	0,9701	1,85	0,9^{2}787	2,25	0,9^{4}20
0,28	0,0^{5}1	0,67	0,2396	1,06	0,7889	1,46	0,9718	1,86	0,9^{2}802	2,26	0,9^{4}26
0,29	0,0^{5}4	0,68	0,2558	1,07	0,7976	1,47	0,9734	1,87	0,9^{2}814	2,27	0,9^{4}34
0,30	0,0^{5}9	0,69	0,2722	1,08	0,8061	1,48	0,9750	1,88	0,9^{2}830	2,28	0,9^{4}40
0,31	0,0^{4}21	0,70	0,2888	1,09	0,8143	1,49	0,9764	1,89	0,9^{2}842	2,29	0,9^{4}44
0,32	0,0^{4}46	0,71	0,3055	1,10	0,8223	1,50	0,9778	1,90	0,9^{2}854	2,30	0,9^{4}49
0,33	0,0^{4}91	0,72	0,3223	1,11	0,8299	1,51	0,9791	1,91	0,9^{2}864	2,31	0,9^{4}54
0,34	0,0^{3}171	0,73	0,3391	1,12	0,8373	1,52	0,9803	1,92	0,9^{2}874	2,32	0,9^{4}58
0,35	0,0^{3}303	0,74	0,3560	1,13	0,8445	1,53	0,9815	1,93	0,9^{2}884	2,33	0,9^{4}62
0,36	0,0^{3}511	0,75	0,3728	1,14	0,8514	1,54	0,9826	1,94	0,9^{2}892	2,34	0,9^{4}65
0,37	0,0^{3}826	0,76	0,3896	1,15	0,8580	1,55	0,9836	1,95	0,9^{3}004	2,35	0,9^{4}68
0,38	0,0^{2}128	0,77	0,4064	1,16	0,8644	1,56	0,9846	1,96	0,9^{3}079	2,36	0,9^{4}70
0,39	0,0^{2}193	0,78	0,4230	1,17	0,8706	1,57	0,9855	1,97	0,9^{3}149	2,37	0,9^{4}73
0,40	0,0^{2}281	0,79	0,4395	1,18	0,8765	1,58	0,9864	1,98	0,9^{3}213	2,38	0,9^{4}76
0,41	0,0^{2}397	0,80	0,4558	1,19	0,8822	1,59	0,9873	1,99	0,9^{3}273	2,39	0,9^{4}78
0,42	0,0^{2}548	0,81	0,4720	1,20	0,8877	1,60	0,9880	2,00	0,9^{3}329	2,40	0,9^{4}80
0,43	0,0^{2}738	0,82	0,4880	1,21	0,8930	1,61	0,9888	2,01	0,9^{3}380	2,41	0,9^{4}82
0,44	0,0^{2}973	0,83	0,5038	1,22	0,8981	1,62	0,9895	2,02	0,9^{3}428	2,42	0,9^{4}84
0,45	0,0126	0,84	0,5194	1,23	0,9030	1,63	0,9^{2}015	2,03	0,9^{3}474	2,43	0,9^{4}86
0,46	0,0160	0,85	0,5347	1,24	0,9076	1,64	0,9^{2}078	2,04	0,9^{3}516	2,44	0,9^{4}87
0,47	0,0200	0,86	0,5497	1,25	0,9121	1,65	0,9^{2}136	2,05	0,9^{3}552	2,45	0,9^{4}88
0,48	0,0247	0,87	0,5645	1,26	0,9164	1,66	0,9^{2}192	2,06	0,9^{3}588	2,46	0,9^{4}89
0,49	0,0300	0,88	0,5791	1,27	0,9205	1,67	0,9^{2}244	2,07	0,9^{3}620	2,47	0,9^{5}
0,50	0,0360	0,89	0,5933	1,28	0,9245	1,68	0,9^{2}293	2,08	0,9^{3}650	2,48	0,9^{5}1
0,51	0,0428	0,90	0,6073	1,29	0,9283	1,69	0,9^{2}339	2,09	0,9^{3}680	2,49	0,9^{5}2
0,52	0,0503	0,91	0,6209	1,30	0,9319	1,70	0,9^{2}383	2,10	0,9^{3}705	2,50	0,9^{5}25
0,53	0,0585	0,92	0,6343	1,31	0,9354	1,71	0,9^{2}423	2,11	0,9^{3}723	2,55	0,9^{5}56
0,54	0,0675	0,93	0,6473	1,32	0,9387	1,72	0,9^{2}461	2,12	0,9^{3}750	2,60	0,9^{5}74
0,55	0,0772	0,94	0,6601	1,33	0,9418	1,73	0,9^{2}497	2,13	0,9^{3}770	2,65	0,9^{5}84
0,56	0,0876	0,95	0,6725	1,34	0,9449	1,74	0,9^{2}531	2,14	0,9^{3}790	2,70	0,9^{6}0
0,57	0,0986	0,96	0,6846	1,35	0,9477	1,75	0,9^{2}562	2,15	0,9^{3}806	2,75	0,9^{6}4
0,58	0,1104	0,97	0,6964	1,36	0,9505	1,76	0,9^{2}592	2,16	0,9^{3}822	2,80	0,9^{6}7
0,59	0,1228	0,98	0,7079	1,37	0,9531	1,77	0,9^{2}620	2,17	0,9^{3}838	2,85	0,9^{6}82
0,60	0,1357	0,99	0,7191	1,38	0,9556	1,78	0,9^{2}646	2,18	0,9^{3}852	2,90	0,9^{7}
				1,39	0,9580	1,79	0,9^{2}670	2,19	0,9^{3}864	2,95	0,9^{7}4
										3,00	0,9^{7}7

A korrelációs együttható kritikus értékei

15. táblázat

Pontok száma, n	Szabadságfok $f=n-2$	80%	90%	95%	99%	99,9%
3	1	0,951	,988	,997	1,000	1,000
4	2	0,800	,900	,950	,990	,999
5	3	0,687	,805	,878	,959	,991
6	4	0,608	,729	,811	,917	,974
7	5	0,551	,669	,755	,875	,951
8	6	0,507	,621	,707	,834	,925
9	7	0,472	,582	,666	,798	,898
10	8	0,443	,549	,632	,765	,872
11	9	0,419	,521	,602	,735	,847
12	10	0,398	,479	,576	,708	,823
13	11	0,380	,476	,553	,684	,801
14	12	0,365	,457	,532	,661	,780
15	13	0,351	,441	,514	,641	,760
16	14	0,338	,426	,497	,623	,742
17	15	0,327	,412	,482	,606	,725
18	16	0,317	,400	,468	,590	,708
19	17	0,308	,389	,456	,575	,693
20	18	0,299	,378	,444	,561	,679
21	19	0,291	,369	,433	,549	,665
22	20	0,284	,360	,423	,537	,652
23	21	0,277	,352	,413	,526	,640
24	22	0,271	,344	,404	,515	,629
25	23	0,265	,337	,396	,505	,618
26	24	0,260	,330	,388	,496	,607
27	25	0,255	,323	,381	,487	,597
28	26	0,250	,317	,374	,479	,588
29	27	0,245	,311	,367	,471	,579
30	28	0,241	,306	,361	,463	,570
31	29	0,237	,301	,355	,456	,562
32	30	0,233	,296	,349	,449	,554
42	40	0,202	,257	,304	,393	,490
62	60	0,165	,211	,250	,325	,408
122	120	0,117	,150	,178	,232	,294

IRODALOM

Arató–Knuth: Sztochasztikus folyamatok elemei. BME,
Bp., Tankönyvkiadó, 1970.

Ágoston–Nagy–Orosz: Méréses módszerek a pedagógiában.
Bp., Tankönyvkiadó, 1974.

Bácskai–Huszti–Meszéna–Mikó–Szép: A gazdasági kockázat és mérésének módszerei.
Bp., Közgazdasági és Jogi Könyvkiadó, 1976.

Balogh–Dukáti–Sallay: Minőségellenőrzés és megbizhatóság.
Bp., Műszaki Könyvkiadó, 1980.

Bányainé–Perczelné: Tartósított termékek statisztikai minőségellenőrzése.
Bp., Mezőgazdasági Kiadó, 1983.

Bethea–Duran–Boullion: Statistical Methods for Engineers and Scientists.
Marcel Dekker Inc., New York and Basel, 1975.

Bognárné–Mogyoródi–Prékopa–Rényi–Szász: Valószínűségszámítási feladatgyűjtemény.
Bp., Tankönyvkiadó, 1971.

Bognárné–Nemetz–Tusnády: Ismerkedés a véletlennel.
Bp., Tankönyvkiadó, 1980.

Csörgő–Révész: Strong Approximations in Probability and Statistics.
Bp., Akadémiai Kiadó, 1981.

Denkinger: Valószínűségszámítás.
Bp., Tankönyvkiadó, 1978.

Denkinger: Valószínűségszámítás példatár.
Bp., Tankönyvkiadó, 1979.

Éltető–Meszéna–Ziermann: Sztochasztikus módszerek és modellek.
Bp., Közgazdasági és Jogi Könyvkiadó, 1982.

Éltető: Mintavétel véges alapsokaságból.
Bolyai János Matematikai Társulat, Bp., 1970.

Éltető–Ziermann: Matematikai statisztika.
Bp., Tankönyvkiadó, 1961.

Ezekiel–Fox: Korreláció- és regresszióanalízis.
Bp., Közgazdasági és Jogi Könyvkiadó, 1970.

Fekete: Matematika és számítástechnika I., II. Bánki Donát Gépipari Műszaki Főiskola.
Bp., Műszaki Könyvkiadó, 1985–86.

Felix–Bláha: Matematikai statisztika a vegyiparban.
Bp., Műszaki Könyvkiadó, 1964.

Füstös–Meszéna–Simonné: Sokváltozós adatelemzés statisztikai módszerei.
Bp., Akadémiai Kiadó, 1985.

Gergely–Hódi–Kelemen–Lukács–Sain–Sólyom: A lánctörtektől a számítógépekig.
Bp., Tankönyvkiadó, 1975.

Gnedenko: The Theory of Probability.
Mir Publichers, Moscow, 1973.

Gnedenko–Kowalenko: Einführung in die Bedienungstheorie.
Berlin, Akademie Verlag, 1971.

Gnyegyenko–Beljajev–Szolovjev: A megbízhatóságelmélet matematikai módszerei.
Bp., Műszaki Könyvkiadó, 1970.

Gulyás: Valószínűségszámítás és matematikai statisztika.
Bp., ELTE, Tankönyvkiadó, 1978.

Gulyás: Fejezetek a matematikai statisztika meteorológiai alkalmazásából.
Bp., Országos Meteorológiai Szolgálat, 1981.

Hajtman: Bevezetés a matematikai statisztikába.
Bp., Akadémiai Kiadó, 1971.

Kolmogorov: A valószínűségszámítás alapfogalmai.
Bp., Gondolat, 1982.

Kovacsics: Statisztika.
Bp., Tankönyvkiadó, 1974.

Köves–Párniczky: Általános statisztika.
Bp., Közgazdasági és Jogi Könyvkiadó, 1973.

Lázár: Általános energiagazdálkodás.
Bp., Műszaki Könyvkiadó, 1968.

Lukács: Matematikai példatár. Könnyűipari Műszaki Főiskola.
Bp., Tankönyvkiadó, 1977.

Lukács: Matematikai statisztika számítógépes alkalmazásokkal.
Bp., KSH Statisztikai Könyvkiadó, 1978.

Lukács: Operációkutatás.
Bp., Tankönyvkiadó, 1982.

Lukács: Matematika IV. Bánki Donát Gépipari Műszaki Főiskola.
Bp., Statisztikai Könyvkiadó, 1981.

Medgyessi–Takács: Valószínűségszámítás.
Bp., Tankönyvkiadó, 1957.

Meszéna–Ziermann: Valószínűségelmélet és matematikai statisztika.
Bp., Közgazdasági és Jogi Könyvkiadó, 1981.

Meszéna (szerk.): Sztochasztikus módszerek a döntés-előkészítésben.
Bp., Tankönyvkiadó, 1984.

Monostory: Valószínűségelmélet és matematikai statisztika. BME.,
Bp., Tankönyvkiadó, 1975.

Monostory: Valószínűségelmélet és matematikai statisztika példatár. BME.,
Bp., Tankönyvkiadó, 1981.

Moroney: Számoktól a tényekig.
Bp., Gondolat, 1970.

Morrison: Multivariate Statistical Methods,
New York, Mc Graw-Hill Book Company, 1976.

Mundruczó: Alkalmazott regressziószámítás.
Bp., Akadémiai Kiadó, 1981.

Prékopa: Valószínűségelmélet.
Bp., Műszaki Könyvkiadó, 1962.

Rasch: Elementare Einführung in die mathematische Statistik.
Berlin, VEB Deutscher Verlag der Wissenschaften, 1970.

Reimann–V. Nagy: Hidrológiai statisztika.
Bp., Tankönyvkiadó, 1984.

Reimann–Tóth: Valószínűségszámítás és matematikai statisztika.
Bp., Tankönyvkiadó, 1985.

Reimann: Ismerkedés a valószínűségszámítással.
Bp., Zrínyi Katonai Kiadó, 1972.

Rejtő–Pachné–Révész: Matematika.
Bp., Mezőgazdasági Kiadó, 1972.

Rényi: Valószínűségszámítás.
Bp., Tankönyvkiadó, 1966.

Révész–Fritz: Az alakfelismerés statisztikus módszerei.
Bp., MTA Matematikai Kutató Intézet, 1974.

Sarkadi–Vincze: Mathematical Methods of Statistical Quality Control.
Bp., Akadémiai Kiadó, 1974.

Sachs: Statisztikai módszerek.
Bp., Mezőgazdasági Kiadó, 1985.

Smirnow–Dunin–Barkowski: Mathematische Statistik in der Technik.
Berlin, VEB Deutscher Verlag der Wissenschaften, 1969.

Solt: Valószínűségszámítás.
Bp., Műszaki Könyvkiadó, 1969.

Sváb: Többváltozós módszerek a biometriában.
Bp., Mezőgazdasági Kiadó, 1979.

Sváb: Biometriai módszerek a kutatásban.
Bp., Mezőgazdasági Kiadó, 1981.

Sweschnikow: Wahrscheinlichkeitsrechnung und Mathematische Statistik in Aufgaben.
Leipzig, Teubner Verlagsgesellschaft, 1970.

Székely: Paradoxonok a véletlen matematikájában.
Bp., Műszaki Könyvkiadó, 1982.

Tatsuoka: Multivariate Analysis.
New York, John Wiley Sons, Inc., 1971.

Tomkó: A Markov-folyamatok elemei és néhány operációkutatási vonatkozása.
Bp., Bolyai János Matematikai Társulat, 1968.

Tomkó: Sztochasztikus folyamatok.
Debrecen, Kossuth Lajos Tudományegyetem, 1977.

Tóth: Ipari energiagazdálkodás.
Bp., Bánki Donát Gépipari Műszaki Főiskola, 1976.

Vincze: Valószínűségszámítás.
Bp., Tankönyvkiadó, 1972.

Vincze: Matematikai statisztika.
Bp., ELTE, Tankönyvkiadó, 1973.

Vincze: Matematikai statisztika ipari alkalmazásokkal.
Bp., Műszaki Könyvkiadó, 1968.

Vincze (szerk.): Statisztikai minőségellenőrzés.
Bp., Közgazdasági és Jogi Könyvkiadó, 1958.

Yuran: Minőségtervezés, -szabályozás, -ellenőrzés.
Bp., Műszaki Könyvkiadó, 1966.

Yule–Kendall: Bevezetés a statisztika elméletébe.
Bp., Közgazdasági és Jogi Könyvkiadó, 1964.

Néhány ajánlható zsebszámológéppel és számítógéppel foglalkozó kiadvány a számítástechnikai részek iránt érdeklődőknek:

Alcock: Ismerd meg a BASIC nyelvet!
Bp., Műszaki Könyvkiadó, 1983.

Alkotó Ifjúság Egyesülés Számítástechnikai Iroda: Matematikai statisztikai programcsomag ZX Spectrum személyi számítógéphez.
Bp., 1984.

Baki: Zsebszámológépek programozása.
Bp., Műszaki Könyvkiadó, 1985.

Bodor–Gerő: A BASIC programozás technikája.
Bp., SZÁMALK, 1983.

Fekete: Számítástechnika.
Bp., Bánki Donát Gépipari Műszaki Főiskola, Műszaki Könyvkiadó, 1983.

Fekete (szerk.): Számítógép-programozási útmutató és példatár.
Bp., Bánki Donát Gépipari Műszaki Főiskola, Műszaki Könyvkiadó, 1981.

Kipnis: Statistik in BASIC.
Berlin, Vogel Verlag, 1985.

Kocsis: Tv-BASIC. SZÁMALK.
Budapest, 1985.

Lukács–Brückner: A folyamatábráktól a programozásig.
Bp., Tankönyvkiadó, 1986.

TÁRGYMUTATÓ

Különvettük egy-egy címszónál, hogy az *a)* az elméleti összefoglalókban, *b)* a feladatoknál szerepel-e. Az utóbbi oldalszámok * jelzés után következnek.

A, Á

B

C, CS

D

E, É

F

G, GY

H

I

J

K

L

M

N

O, Ö

P

R

S

SZ

T

U

V

W

Kiadja a Műszaki Könyvkiadó
Felelős kiadó: Bérczi Sándor ügyvezető igazgató

Felelős szerkesztő: Szokol Ágnes
Műszaki vezető: Abonyi Ferenc
Műszaki szerkesztő: Trencséni Ágnes
A borítót tervezte: Kováts Tibor
A könyv ábráit rajzolta: Maklári Zoltánné
A könyv formátuma: Fr5
Terjedelem: 29,0 (A/5) ív
Azonossági szám: 10276
Készült az MSZ 5601:1983 és 5602:1983 szerint

Nyomta és kötötte az Oláh Nyomdaipari Kft.
Felelős vezető: Oláh Miklós