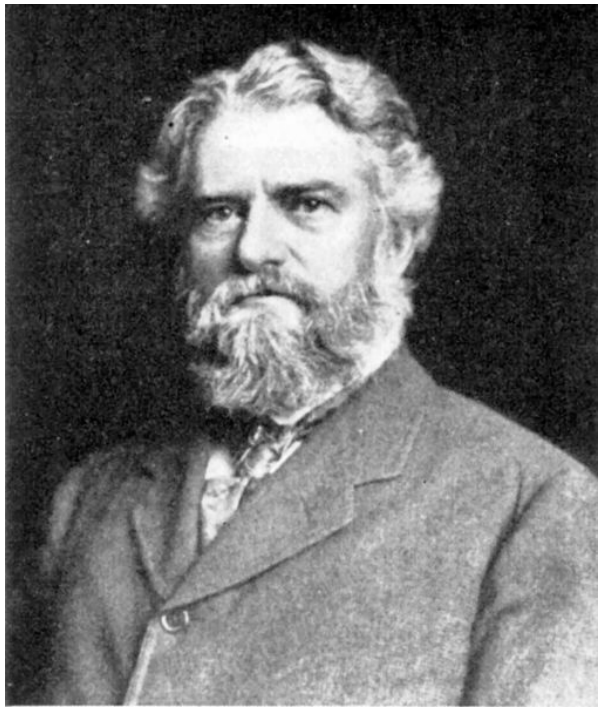
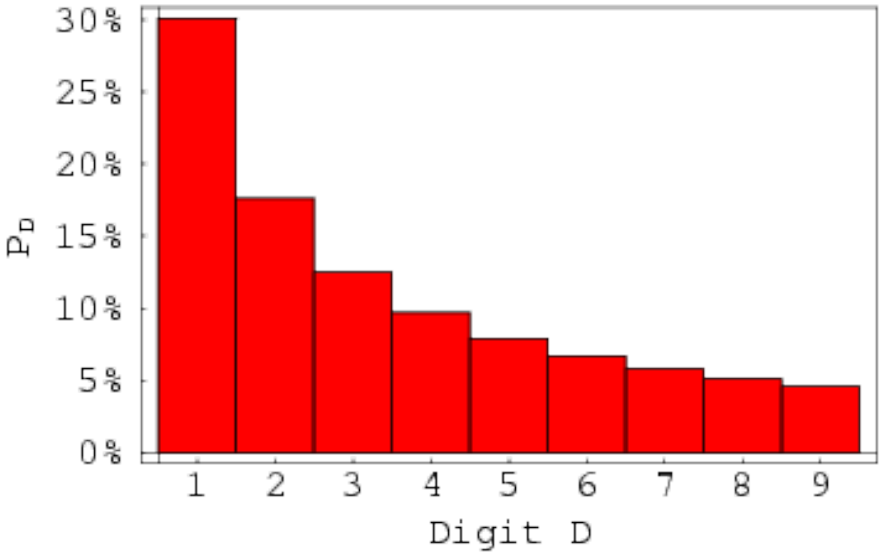
Az első számjegyek Benford törvénye

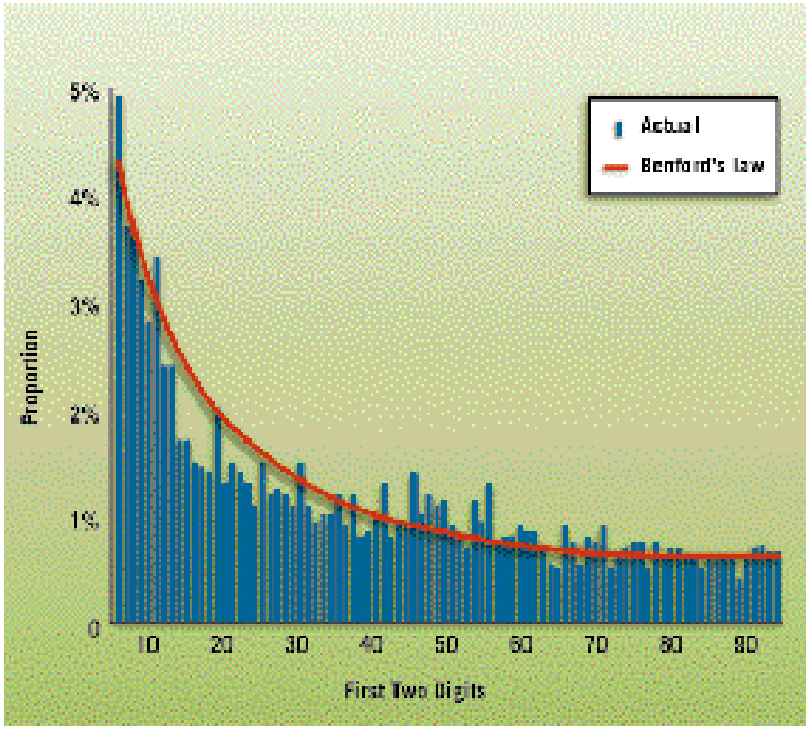
Frank Benford (1883-1948)
A General Electric fizikusa

Simon Newcomb (1835 – 1909)
asztronómus



$$P(d) = \frac{\log_{10}(1+1/d)}{\log_{10} B}$$





A híres arizonai csekk sikkasztási eset

The table lists the checks that a manager in the office of the Arizona State Treasurer wrote to divert funds for his own use. The vendors to whom the checks were issued were fictitious.

Date of Check	Amount
October 9, 1992	$ 1,927.48
	27,902.31
October 14, 1992	86,241.90
	72,117.46
	81,321.75
	97,473.96
October 19, 1992	93,249.11
	89,658.17
	87,776.89
	92,105.83
	79,949.16
	87,602.93
	96,879.27
	91,806.47
	84,991.67
	90,831.83
	93,766.67
	88,338.72
	94,639.49
	83,709.28
	96,412.21
	88,432.86
	71,552.16
TOTAL	$ 1,878,687.58

http://www.aicpa.org/pubs/jofa/may1999/nigrini.htm

11. előadás
Multinomiális kísérletek és kontingencia táblák

11-1 Áttekintés

11-2 Multinomiális kísérletek: az illeszkedés jósága

11-3 Kontingencia táblák: Függetlenség és homogenitás

11-1 & 11-2 fejezetek Áttekintés és multinomiális kísérletek: az illeszkedés jósága

Áttekintés

- **Kategoriális adatokkal foglalkozunk, vagy olyan kvantitatív adatokkal, amelyeket különböző kategóriákba lehet sorolni (gyakran bineknek vagy celláknak hívjuk).**
- **A χ^2 (khí-négyzet) teszt statisztika.**
- **Az illeszkedés vizsgálat (goodness of fit test) egy egydimenziós gyakorisági táblázat (egy sor vagy oszlop).**
- **A kontingencia tábla egy kétdimenzós gyakorisági táblázat (kettő vagy több oszlop és sor).**

Kulcsfogalmak

Adott, kategóriákba sorolt adatok esetén azt a hipotézist teszteljük, hogy az adatok eloszlása megegyezik valamilyen általunk feltételezett eloszlással.

A hipotézis teszt a khí-négyzet eloszlást használja a megfigyelt gyakoriságok és az általunk várt gyakoriságok összehasonlítására.

Definíció

Multinomiális kísérlet

Egy olyan kísérlet, ami az alábbi feltételeknek tesz eleget:

1. **A próbálkozások/kísérletek száma előre adott.**

2. **A próbálkozások/kísérletek függetlenek.**

3. **A kísérlet minden kimenetele egyértelműen besorolható pontosan egybe a lehetséges kategóriák közül.**

4. **A kísérletek során a kategóriák valószínűsége nem változik, állandó marad.**

Példa: A tömegek utolsó számjegye

Amikor az embereket megkérdezik, hogy mekkora a tömegük, gyakran mondanak a valóságosnál kisebb értékeket. Hogyan lehet eldönteni egy adathalmazról, hogy igazi mérésből származnak, vagy az emberek megkérdezéséből nyert értékek?

Példa: A tömegek utolsó jegye

Teszteljük azt a feltevést, hogy az 11-2. táblázatban található értékek ugyanazzal a gyakorisággal lépnek fel.

11-2. táblázat összesítés 80 hallgató tömegének utolsó számjegyei

Table 11-2
Last Digits of Weights

Last Digit	Frequency
0	35
1	0
2	2
3	1
4	4
5	24
6	1
7	4
8	7
9	2

Példa: folyt.

Ellenőrizzük, hogy a multinomiális kísérlet feltételei fennállnak-e.

1. **A kísérletek száma adott, 80.**
2. **A kísérletek függetlenek, mert valaki tömegének utolsó számjegye nincs hatással valaki más tömegének utolsó számjegyére.**
3. **Minden kimenet (utolsó számjegy) pontosan egy kategóriába sorolható. A kategóriák 0, 1, … , 9.**
4. **Végül, pedig nem változik a kimenetek valószínűsége a kísérlet során.**

Definíció

Illeszkedés vizsgálat

Az illeszkedés vizsgálatot annak tesztelésére használjuk, hogy a megfigyelt gyakoriságok illeszkednek a feltételezett gyakoriság eloszláshoz.

Illeszkedés vizsgálat

Jelölések

O jelöli egy kimenetel **megfigyelt gyakoriságát**.

E jelöli egy kimenetel **várt gyakoriságát**.

k jelöli a lehetséges kimenetek/**kategóriák számát**.

n jelöli a **kísérletek teljes számát**.

Várt gyakoriságok

Ha minden gyakoriság egyenlő:

$$E = \frac{n}{k}$$

az összes megfigyelt előfordulások száma elosztva a kategóriák számával

Várt gyakoriságok

Ha nem mindegyik gyakoriság egyforma:

$$E = n\,p$$

Meg kell szorozni a kategória valószínűséget az összes esetek számával.

Illeszkedés vizsgálat (teszt) multinomiális kísérletekben

Követelmények

1. **Az adatokat véletlenül választjuk ki**
2. **A minta minden kategóriára vonatkozó gyakoriság adataiból áll.**
3. **Minden kategóriában legalább 5 legyen a várt megfigyelések száma!**

Illeszkedés vizsgálat (teszt) multinomiális kísérletekben

Teszt statisztika

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Kritikus értékek

1. **A khí-négyzet táblázatot kell használnunk**

k – 1 szabadsági fokok számával, ahol k = a kategóriák száma.

2. **Az illeszkedés vizsgálatok mindig jobboldali tesztek.**

Illeszkedés vizsgálat (teszt) multinomiális kísérletekben

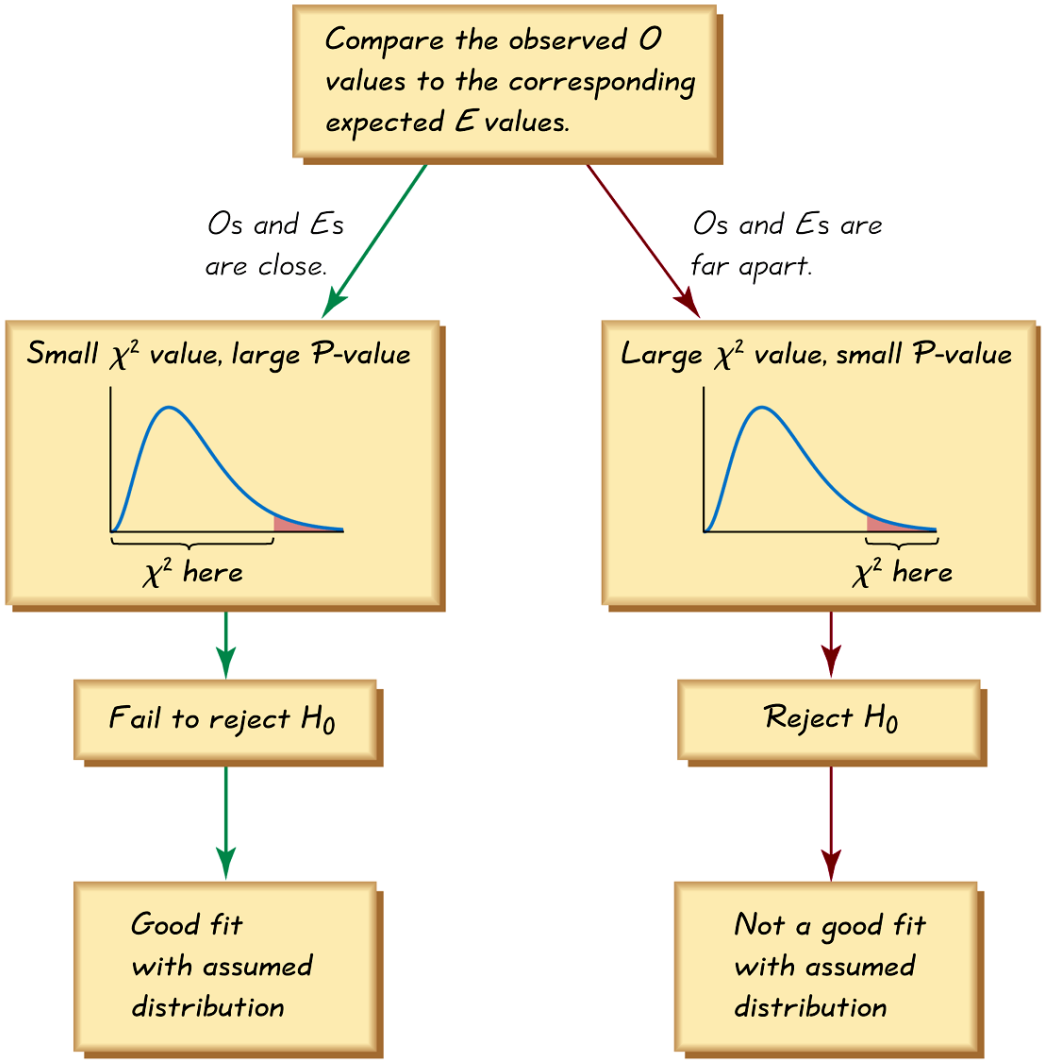
- A **közeli egyezés** a megfigyelt és a várt értékek között kicsi χ^2 és nagy P-értékre vezetnek.

- A **nagy eltérés** a megfigyelt és a várt értékek között nagy χ^2 és kis P-értékre vezetnek.

- Egy **szignifikánsan nagy** χ^2 érték a null hipotézis **elutasítását** fogja okozni, amennyiben a null hipotézis szerint nincs különbség a megfigyelt és a várt gyakoriságok között.

Kapcsolat a χ^2 teszt statisztika, P-érték, és az illeszkedés vizsgálat között

11-3. ábra



Példa: utolsó számjegy elemzés

Teszteljük azt a feltevést, hogy a 11-2. táblázatban a számjegyek nem ugyanazzal a gyakorisággal fordulnak elő.

H_0: $p_0 = p_1 = \ldots = p_9$

H_1: Legalább az egyik vsz. különbözik a többitől.

$\alpha = 0.05$

$k - 1 = 9$

$\chi^2_{.05,\,9} = 16.919$

Table 11-2
Last Digits of Weights

Last Digit	Frequency
0	35
1	0
2	2
3	1
4	4
5	24
6	1
7	4
8	7
9	2

Példa: utolsó számjegy elemzés

Teszteljük azt a feltevést, hogy a 11-2. táblázatban a számjegyek nem ugyanazzal a gyakorisággal fordulnak elő.

Ha a 80 számjegy egyenletesen oszlana el a 10 kategória között, akkor minden gyakoriságra 8-at várunk.

Table 11-2
Last Digits of Weights

Last Digit	Frequency
0	35
1	0
2	2
3	1
4	4
5	24
6	1
7	4
8	7
9	2

Példa: utolsó számjegy elemzés

Teszteljük azt a feltevést, hogy a 11-2. táblázatban a számjegyek nem ugyanazzal a gyakorisággal fordulnak elő.

Table 11-3 Calculating the χ^2 Test Statistic for the Last Digits of Weights

Last Digit	Observed Frequency O	Expected Frequency E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
0	35	8	27	729	91.1250
1	0	8	−8	64	8.0000
2	2	8	−6	36	4.500
3	1	8	−7	49	6.125
4	4	8	−4	16	2.000
5	24	8	16	256	32.000
6	1	8	−7	49	6.125
7	4	8	−4	16	2.000
8	7	8	−1	1	0.125
9	2	8	−6	36	4.500
	80	80			$\chi^2 = \sum \frac{(O - E)^2}{E} = 156.500$

(Except for rounding errors, these two totals must agree.)

Példa: utolsó számjegy elemzés

Teszteljük azt a feltevést, hogy a 11-2. táblázatban a számjegyek nem ugyanazzal a gyakorisággal fordulnak elő.

A 11-3. táblázat szerint, a teszt statisztika értéke
$\chi^2 = 156.500$.

Mivel a kritikus éték 16.919, elutasítjuk a null hipotézist, amely szerint a valószínűségek megegyeznek.

Elegendő evidencia van arra, hogy támogassuk azt a feltevést, hogy az utolsó számjegyek nem mind ugyanakkora gyakorisággal fordulnak elő.

Példa: Csalás detektálás

11-1. táblázat: Az első számjegyek statisztikája és a Brenford szabály.

Table 11-1 Benford's Law: Distribution of Leading Digits

Leading Digit	1	2	3	4	5	6	7	8	9
Benford's law: frequency distribution of leading digits	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%
Expected frequencies of leading digits from 784 checks following Benford's law	235.984	137.984	98.000	76.048	61.936	52.528	45.472	39.984	36.064
Observed leading digits of 784 actual checks analyzed for fraud	0	15	0	76	479	183	8	23	0

Példa: Csalás detektálás

Teszteljük azt a feltevést, hogy szignifikáns eltérés van a Benford szabály és a 784 db számla első számjegye között.

Observed Frequencies and Frequencies Expected with Benford's Law

Digit	Observed Frequency	Expected Frequency	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
1	0	235.984	−235.984	55688.4483	235.9840
2	15	137.984	−122.984	15125.0643	109.6146
3	0	98.000	−98.000	9604.0000	98.0000
4	76	76.048	−0.048	0.0023	0.0000
5	479	61.936	417.064	173942.3801	2808.4213
6	183	52.528	130.472	17022.9428	324.0737
7	8	45.472	−37.472	1404.1508	30.8795
8	23	39.984	−16.984	288.4563	7.2143
9	0	36.064	−36.064	1300.6121	36.0640
				Total: $\chi^2 = \sum \frac{(O - E)^2}{E}$	= 3650.2514

Példa: Csalás detektálás

Teszteljük azt a feltevést, hogy szignifikáns eltérés van a Benford szabály és a 784 db számla első számjegye között.

H_0: $p_1 = 0.301$, $p_2 = 0.176$, $p_3 = 0.125$, $p_4 = 0.097$, $p_5 = 0.079$, $p_6 = 0.067$, $p_7 = 0.058$, $p_8 = 0.051$ and $p_9 = 0.046$

H_1: Legalább egy gyakoriság eltér ezektől az arányoktól.

$\alpha = 0.01$

$k - 1 = 8$

$\chi^2_{.01,8} = 20.090$

Példa: Csalás detektálás

Teszteljük azt a feltevést, hogy szignifikáns eltérés van a Brenford szabály és a 784 db számla első számjegye között.

A teszt statisztika értéke $\chi^2 = 3650{,}251$.

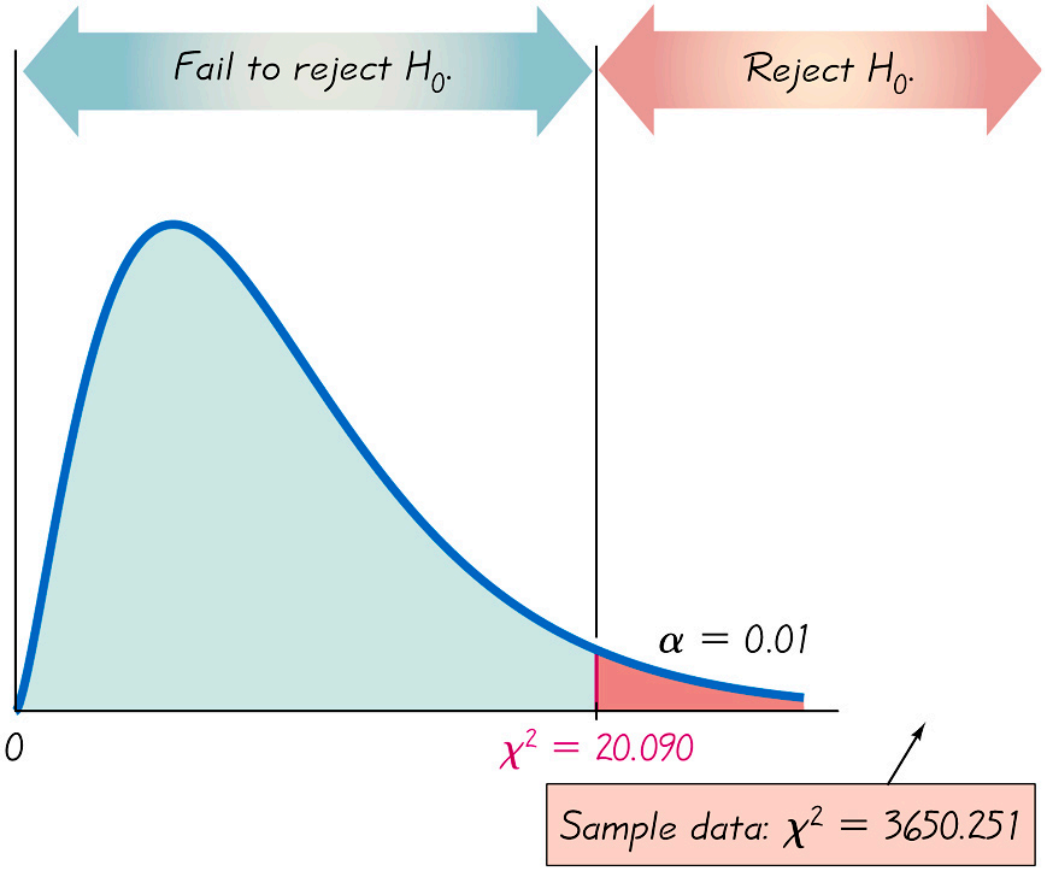
Mivel a kritikus érték 20,090 , elutasítjuk a null hipotézist.

Elég bizonyíték van a null hipotézis elutasítására - Elég bizonyíték van arra, hogy legalább az egyik arány eltér a várhatótól.

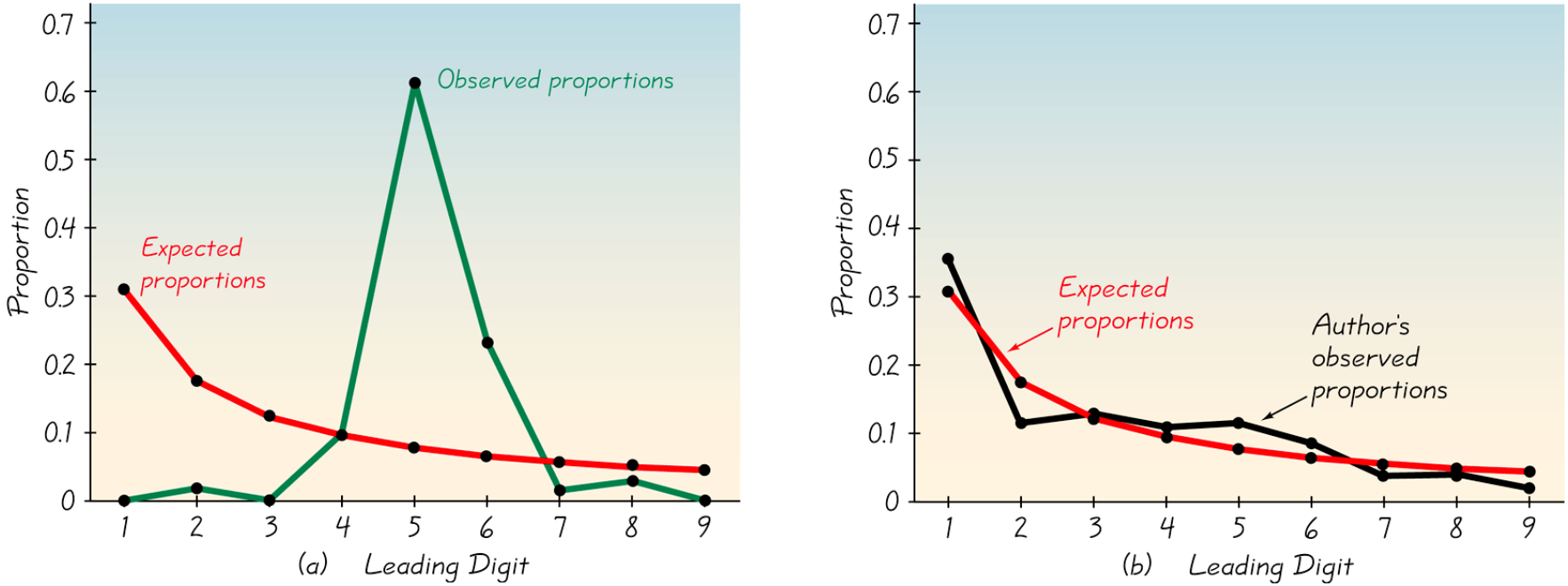
Példa: Csalás detektálás

Teszteljük azt a feltevést, hogy szignifikáns eltérés van a Benford szabály és a 784 db számla első számjegye között.

11-5. ábra



Példa: Csalás detektálás



11-6. ábra A megfigyelt és a Benford törvénynek megfelelő első számjegy eloszlások

Összefoglalás

Ebben a fejezetben megbeszéltük:

Multinomiális kísérletek: Illeszkedés jósága

Annak a hipotézisnek a tesztelése, hogy a megfigyelt gyakoriság eloszlás illeszkedik a feltételezett eloszláshoz.

11-3. fejezet
Kontingencia táblázatok: Függetlenség és homogenitás

Kulcsfogalmak

Ebben a fejezetben kontingencia vagy más néven két dimenziós gyakorisági táblázatokkal foglalkozunk.

Olyan eljárást mutatunk be, amivel vizsgálni lehet, hogy a sor és az oszlop változók függetlenek-e egymástól.

A homogenitás vizsgálatára ugyanezt a módszert használjuk, amellyel eldönthető, hogy különböző populációkban valamilyen tulajdonság ugyanolyan megoszlásban van-e jelen.

Definíció

Kontingencia táblázat

(vagy kétdimenziós gyakorisági táblázat)

Egy kontingencia táblázat olyan táblázat, melyekben a gyakoriságok két változóhoz tartoznak.

(Az egyik változó kategorizálja az oszlopokat, a másik a sorokat.)

A kontingencia táblázatok minimum 2 X 2-esek.

Esettanulmány motorosokról

A bukósisak színe és a baleseti sérülések között van-e valamilyen kapcsolat?

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll (nem sérült)	**491**	**377**	**31**	**899**
Balesetes (sérült v. meghalt)	**213**	**112**	**8**	**333**
Oszlopösszeg	**704**	**489**	**39**	**1232**

Definíció

Függetlenség vizsgálat (teszt)

A függetlenség vizsgálat azt a null hipotézist teszteli, hogy nincs kapcsolat az oszlop és a sor változó között a kontingencia táblában. A null hipotézis az, hogy a „sor és oszlop változók függetlenek”.

Követelmények

1. **A minta adatokat véletlenül választjuk ki és két dimenziós gyakorisági táblázatban helyezzük el.**

2. **A null hipotézis H_0 az, hogy a sor és oszlop változók függetlenek; az alternatív hipotézis H_1 az, hogy az oszlop és sor változók függenek egymástól.**

3. **A kontingencia táblában minden várható gyakoriság E legalább 5. (Nem feltétel, hogy a megfigyelt esetek száma legalább 5 legyen. Nem feltétel, hogy a populáció normális eloszlású legyen.)**

Függetlenségi teszt

Teszt statisztika

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Kritikus értékek

1. **A khí-négyzet eloszlás táblázatából**

   szabadsági fokok száma = $(r - 1)(c - 1)$

   r a sorok, c az oszlopok száma

2. **A függetlenségi teszt mindig jobboldali.**

Feltételezett/várható gyakoriság

$$E = \frac{(\text{sor összeg})\,(\text{oszlop összeg})}{(\text{összes eset})}$$

A megfigyelt gyakoriságok teljes száma az egész táblázatban

Függetlenségi teszt

Ez a procedúra nem alkalmas arra, hogy direkt ok-okozati kapcsolatot mutassunk ki a változók között.

A függőség csak azt jelenti, hogy kapcsolat van a két változó között.

A kontingencia tábla várható gyakorisága

$$E = \text{összes eset} \cdot \frac{\text{sorösszeg}}{\text{összes eset}} \cdot \frac{\text{oszlopösszeg}}{\text{összes eset}}$$

$$E = n \cdot p$$

(cella valószínűség)

$$E = \frac{(\text{sorösszeg})\,(\text{oszlopösszeg})}{(\text{összes eset})}$$

Eset tanulmány

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll (nem sérült)	491	377	31	899
Balesetes	213	112	8	333
Oszlopösszeg	704	489	39	1232

A bal felső cellára:

$$E = \frac{\text{(sorösszeg) (oszlopösszeg)}}{\text{(összes eset)}}$$

$$E = \frac{(899)(704)}{1232} = 513.714$$

Esettanulmány

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll	491	377	31	899
Várt esetszám	513.714			
Balesetes	213	112	8	333
Várt esetszám				
Oszlopösszeg	704	489	39	1232

$$E = \frac{(\text{sorösszeg})\,(\text{oszlopösszeg})}{(\text{összes eset})}$$

$$E = \frac{(899)(704)}{1232} = 513.714$$

Esettanulmány

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll	**491**	**377**	**31**	**899**
Várt	**513.714**	**356.827**	**28.459**	
Balesetes	**213**	**112**	**8**	**333**
Várt	**190.286**	**132.173**	**10.541**	
Oszlopösszeg	**704**	**489**	**39**	**1232**

Kiszámítottuk a várható esetszámot.

A bal felső cella interpretálása: azt mondhatjuk, hogy 491 fekete sisakos motoros sérült meg, de 513.714 lenne a várható szám, ha a sérülések függetlenek lennének a sisak színétől.

folyt.

A 0.05 szignifikancia szintet használva teszteljük azt a feltevést, hogy a csoport (kontroll vagy balesetes) független a sisak színétől.

H_0: Az, hogy valaki a kontroll vagy a balesetes csoportba esik független a sisak színétől.

H_1: A csoport és a szín összefüggnek.

folyt.

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll	**491**	**377**	**31**	**899**
Várható	**513.714**	**356.827**	**28.459**	
Balesetes	**213**	**112**	**8**	**333**
Várható	**190.286**	**132.173**	**10.541**	
Oszlopösszeg	**704**	**489**	**39**	**1232**

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(491-513.714)^2}{513.714} + \ldots + \frac{(8-10.541)^2}{10.541}$$

$$\chi^2 = 8.775$$

folyt.

H_0: Sor és oszlop változók függetlenek.

H_1: Sor és oszlop változók összefüggnek.

A teszt statisztika $\chi 2 = 8.775$

$\alpha = 0.05$

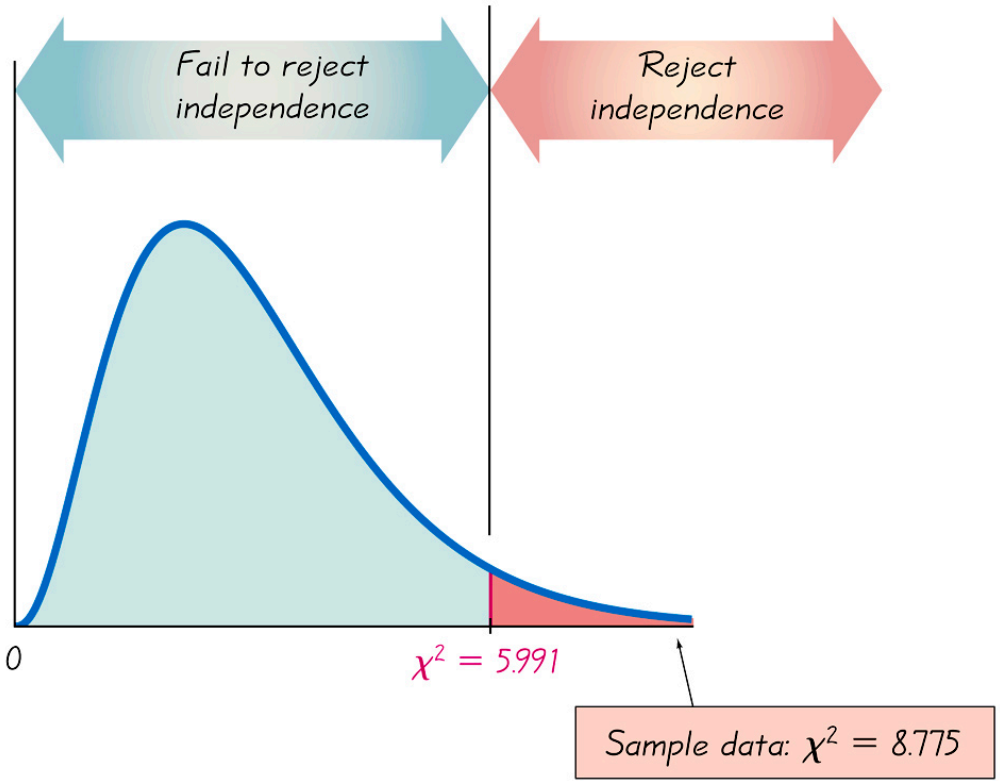
A szabadsági fokok száma:

$(r-1)(c-1) = (2-1)(3-1) = 2.$

A kritikus érték a táblázatból $\chi^2_{.05,2} = 5.991.$

folyt.

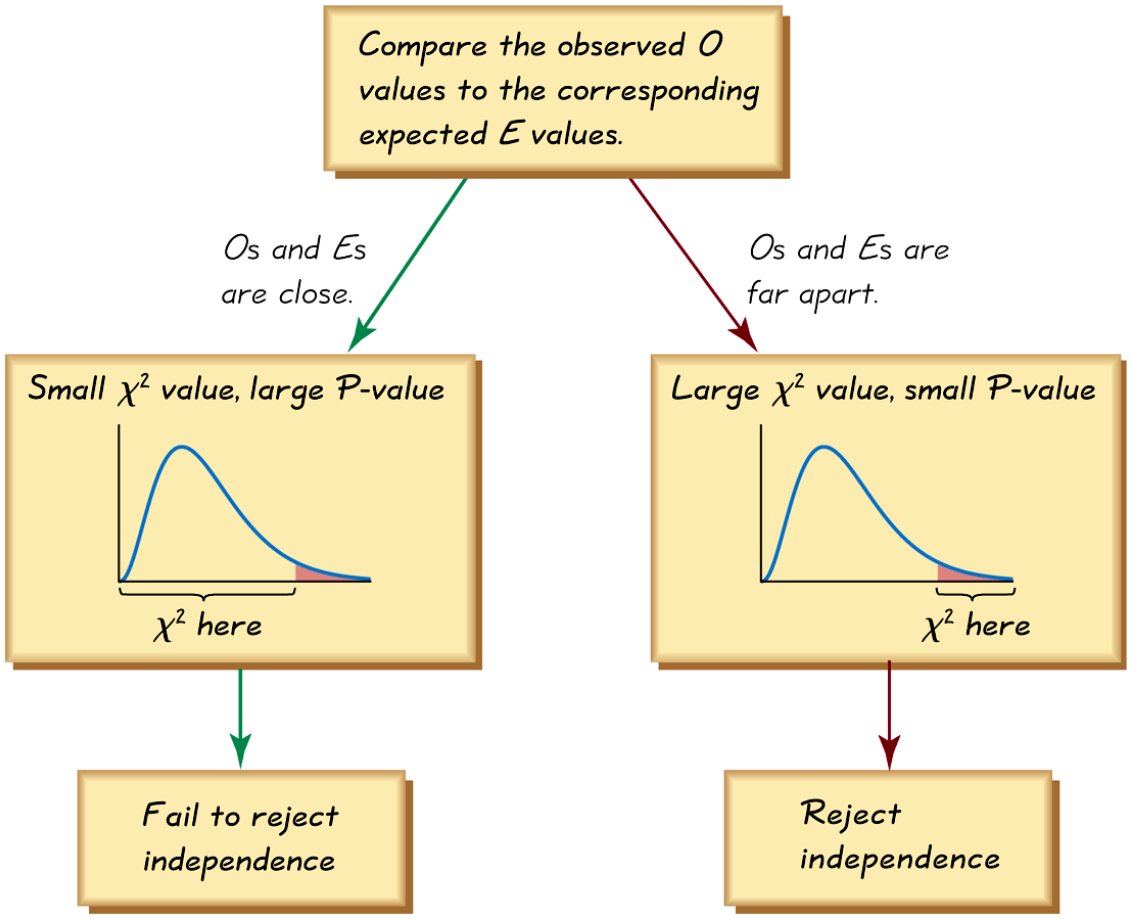
11-4. ábra



Elvetjük a null hipotézist. Úgy tűnik, van kapcsolat a sisak színe és a motorozás biztonsága között.

A tesztelés menete

11-8. ábra



Definíció

Homogenitás vizsgálat

A homogenitás vizsgálatban, azt a feltevést teszteljük, hogy *különböző populációk* bizonyos tulajdonságokat ugyanolyan arányban tartalmaznak.

Mi a különbség a homogenitás és a függetlenség vizsgálat között:

Egy *előre meghatározott* minta elemszámot használunk mindkét populációból (homogenitás vizsgálat), vagy egy nagy mintát használtunk, amiből a sor és az oszlopösszegek véletlenül jönnek ki (függetlenség vizsgálat)?

Példa: A nemek hatása

Az 11-6. táblázatot használva 0.05 szignifikancia szint mellett teszteljük, van-e hatása a kérdező nemének a férfi válaszolók válaszaira.

Table 11-6 Gender and Survey Responses

	Gender of Interviewer	
	Man	Woman
Men who agree	560	308
Men who disagree	240	92

Példa: folyt

H_0: Azok aránya, akik egyetértenek/nem értenek egyet ugyanakkora a férfi és a női kérdezők esetén is.

H_1: Az arányok különböznek.

Példa: folyt.

Minitab

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

            C1       C2  Total
    1      560      308    868
        578.67   289.33
         0.602    1.204

    2      240       92    332
        221.33   110.67
         1.574    3.149

Total      800      400   1200

Chi-Sq = 6.529, DF = 1, P-Value = 0.011
```

Összefoglalás

Ebben a fejezetben megvitattuk:

Kontingencia táblázatokat, ahol kategoriális adatok sorokba és oszlopokba vannak rendezve.

*** Függetlenség vizsgálattal teszteljük azt a feltételezést, hogy a sor és az oszlop változók függetlenek.**

*** Homogenitás vizsgálattal teszteljük azt a feltételezést, hogy két populáció valamilyen tulajdonságot ugyanolyan arányban tartalmaz.**

11-4. fejezet
McNemar teszt párokba rendezett adatokra

Kulcsfogalmak

Eddig független folyamatokból származó adatokkal foglalkoztunk.

Hogyan kezeljünk olyan adatokat, amelyek párokba rendezetten keletkeznek? Ekkor nem teljesül a függetlenség feltétele.

Ilyenkor használjuk McNemar párosított adatokra kidolgozott tesztjét.

Példa: két kezelést kapott minden páciens, néhányan gyógyultak, néhányan nem.

Table 11-9 2 × 2 Table with Frequency Counts from Matched Pairs

		Treatment X	
		Cured	Not Cured
Treatment Y	Cured	*a*	*b*
	Not cured	*c*	*d*

Melyik kezelés az eredményesebb?
Vessük össze a b és c számokat!

Table 11-9 2 × 2 Table with Frequency Counts from Matched Pairs

		Treatment X	
		Cured	Not Cured
Treatment Y	Cured	*a*	*b*
	Not cured	*c*	*d*

Nullhipotézis: b aránya = c aránya

Table 11-9 2 × 2 Table with Frequency Counts from Matched Pairs

		Treatment X	
		Cured	Not Cured
Treatment Y	Cured	*a*	*b*
	Not cured	*c*	*d*

Párokba rendezett előfordulás: minden páciens megkapta mindkét kezelést

Előfeltételek

1. **Véletlen mintavételezés**
2. **A mintáról párokba rendezett előfordulási gyakoriságokat rögzítettünk**
3. **Nominális adatok, amiket két osztályba sorolhatunk minden párra.**
4. **A kontingencia-tábla nem-diagonális elemei legalább 10-szer előfordultak**
   példánkban: $b+c \geq 10$
   azaz a különböző eredményre vezető esetek legalább 10-szer fordultak elő

Teszt

Null hipotézis: b, c relatív gyakorisága egyenlő

Próba statisztika:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$$

Kritikus tartomány: jobboldali teszt

Szabadsági fokok száma = 1

Példa: kezelések összehasonlítása

Table 11-10 Clinical Trials of Treatments for Athlete's Foot

		Treatment with Pedacream	
		Cured	Not Cured
Treatment with Fungacream	Cured	12	8
	Not cured	40	20

Párokba rendezett előfordulás: egyik lábon Fungakrém, másik lábon Pedakrém kezelés

Hipotézis-vizsgálat

H_0: az alábbi két relatív gyakoriság egyenlő:

1. **Pedakrémmel kezelt lábon gyógyult és a Fungakrémmel kezelt lábon nem javuló betegek aránya**
2. **Fungakrémmel kezelt lábon gyógyult és a Pedakrémmel kezelt lábon nem javuló betegek aránya**

Kérdés: az adatok alapján mondhatjuk-e, hogy eltérést látunk a két arány között?

Hipotézis-vizsgálat

A teszt előfeltételei teljesülnek:

1. **Véletlenszerűen kiválasztott mintából kaptuk az adatokat**
2. **Minden megfigyelésünk egyértelműen kategorizálható két változó-érték valamelyikére (egyik változó = Peda- vagy Fungakrém, másik változó = gyógyult/nem gyógyult)**
3. $b = 8$ **és** $c = 40$, **tehát** $b + c \geq 10$.

Teszt-statisztika

Teszt-statisztika:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|8-40|-1)^2}{8+40} = 20.021$$

Kritikus érték táblázatból:

$$\chi^2 = 3.841$$

Mivel a kritikus értéknél nagyobb a statisztikánk értéke, ezért elfogadhatatlanul kicsi a valószínűsége, hogy a két kezelés egyforma

Elutasítjuk a null hipotézist.

Fogalmak

A teszt nem használja azokat az értékeket, amiknél az értékek egyezőek voltak.

Definíció: negatív párosításnak (diszkordáns párnak) nevezzük azokat az értékpárokat, amelyekre eltérő kategóriákba esnek a párosított értékek (Példánkban: gyógyult/nem gyógyult)

Összefoglalás

Ebben a fejezetben megvitattuk:

McNemar tesztet, ahol párosított értékeket vizsgálunk.

*** Itt az adatok 2x2-es táblázatba vannak rendezve, ahol minden érték két kategória szerint van osztályozva**
*** A teszt csak a diszkordáns (eltérő kategóriájú) párok előfordulásait használja fel**