Elemi statisztika fizikusoknak

Pollner Péter

Biológiai Fizika Tanszék

pollner@elte.hu

6. Előadás
A normális eloszlás

A fejezet példája:

Nemrég Baltimore belső kikötőjében elsüllyedt egy vízitaxi. A 25 rajta tartózkodó ember közül 5-en meghaltak, 16-an megsebesültek. A vizsgálat kimutatta, hogy a biztonságos össz utas tömeg 1600 kg lett volna. Feltéve, hogy egy utas átlagos tömege 64 kg, 25 utas felvétele volt engedélyezve. A 64 kg-os átlagot 44 évvel ezelőtt állapították meg, amikor az emberek sokkal könnyebbek voltak. (Az elsüllyedt hajó 25 utasának átlagos tömege 76 kg volt.) Az eset után az USA-ban a közlekedési eszközökön 80 kg-ra emelték. Így 1600 kg össztömeg esetén már csak 20 utast szabad felengedni.

6-3. fejezet
A normális eloszlás alkalmazásai

Kulcsfogalmak

Ebben a fejezetben átnézzük, hogy hogyan kell olyan normális eloszlásokkal dolgozni, amelyeknek nem 0 az átlaguk és nem 1 a szórásuk.

A legfontosabb, hogy egyszerűen átkonvertálhatunk egy nem standard eloszlást úgy, hogy az eredmény standard normális eloszlás legyen és így a korábban használt módszereket alkalmazni tudjuk.

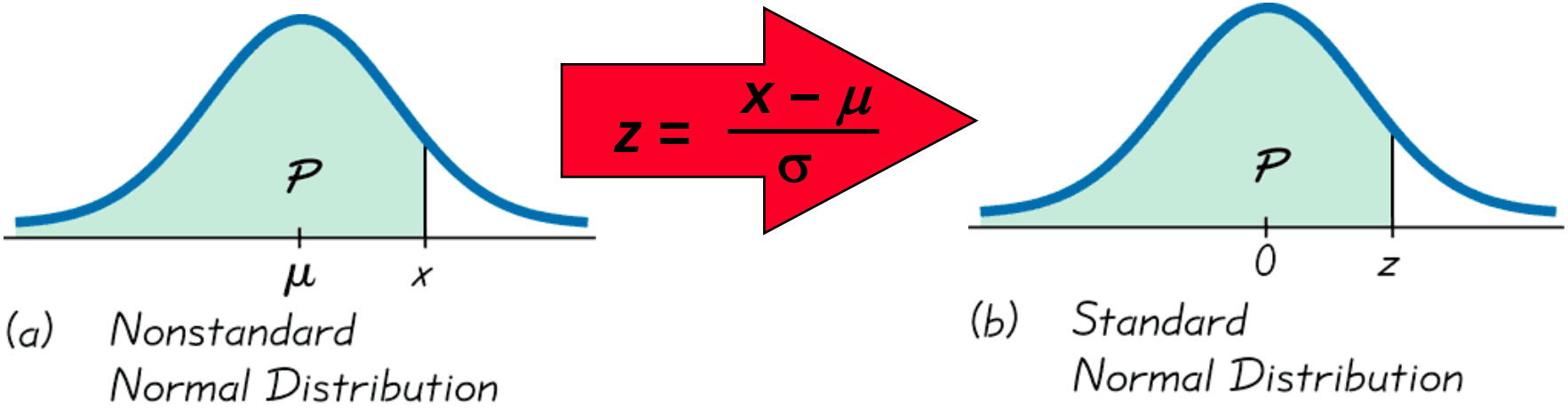
Konverziós formula (standardizálás)

6-2. képlet

$$z = \frac{x - \mu}{\sigma}$$

$$x = \mu + \sigma \cdot z$$

Konvertálás nem-standardból standardba



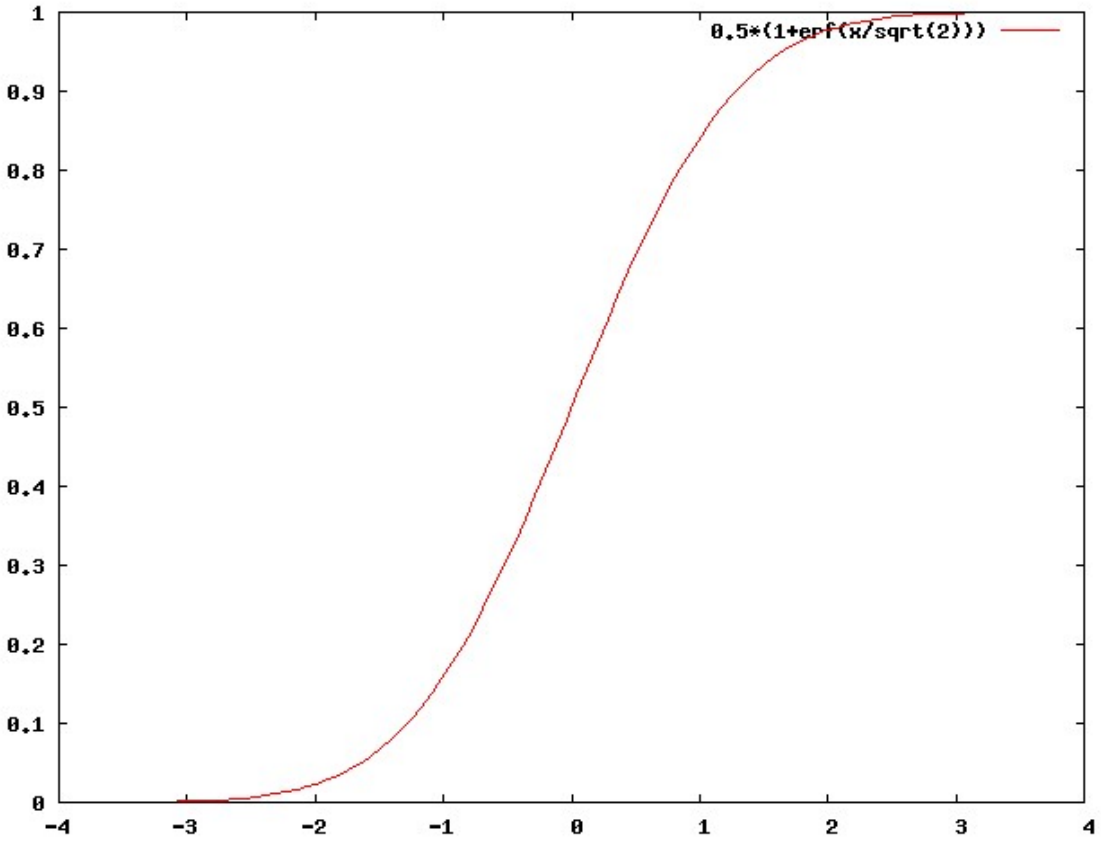
(a) Nonstandard Normal Distribution

(b) Standard Normal Distribution

6-12. ábra

A hiba függvény

$$\Pr(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right).$$



Példa – a vízitaxi utasainak súlyeloszlása

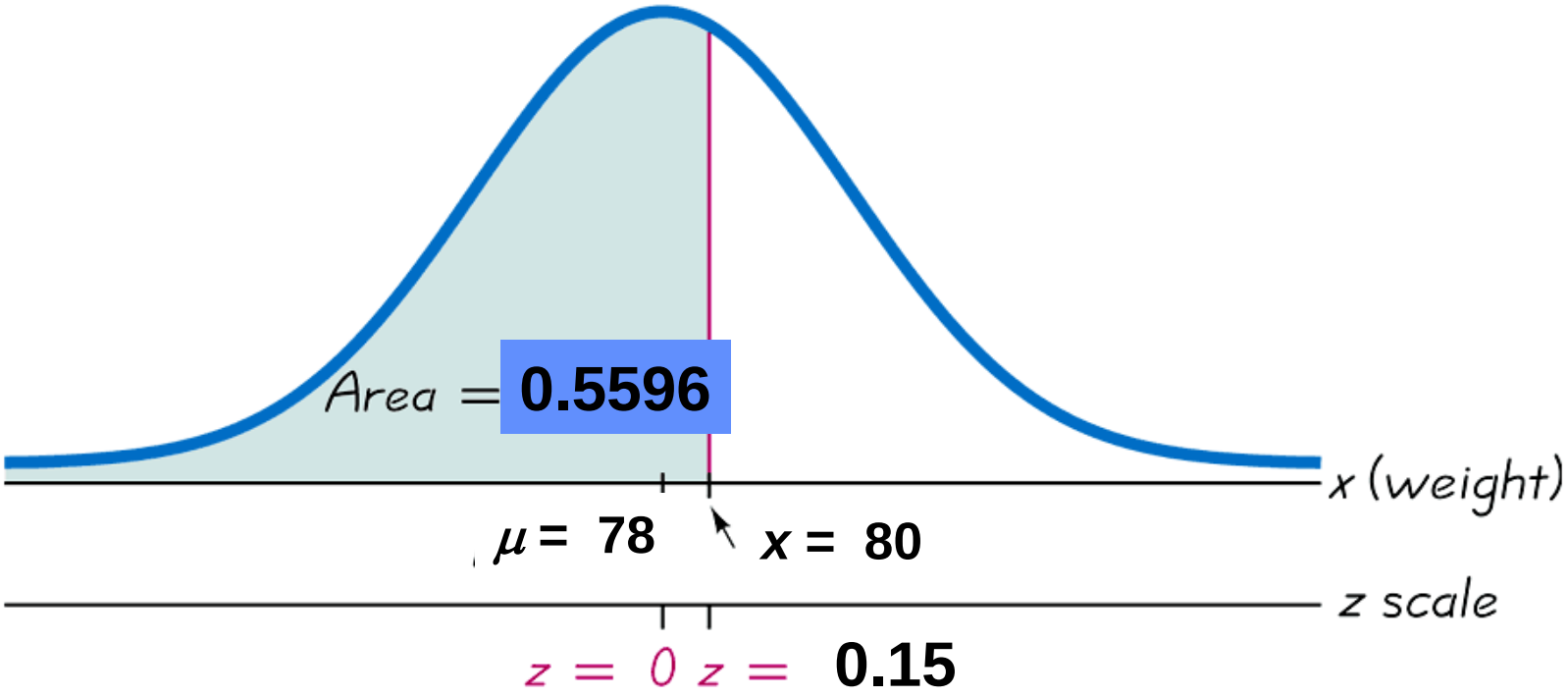
A fejezet elején a vízitaxi megengedett utas tömege 1600 kg volt és az átlagos utas tömegét 64 kg-nak feltételezték. Tegyük fel a legrosszabb esetet, hogy az összes utas férfi. És tegyük fel, hogy a férfiak tömege normális eloszlást követ 78 kg-os átlaggal és 13 kg szórással. Ha véletlenül választunk egyet, mi a valószínűsége annak, hogy tömege kisebb mint 80 kg?

Példa - folyt

$\mu = 78$

$\sigma = 13$

$$z = \frac{80 - 78}{13} = 0.15$$



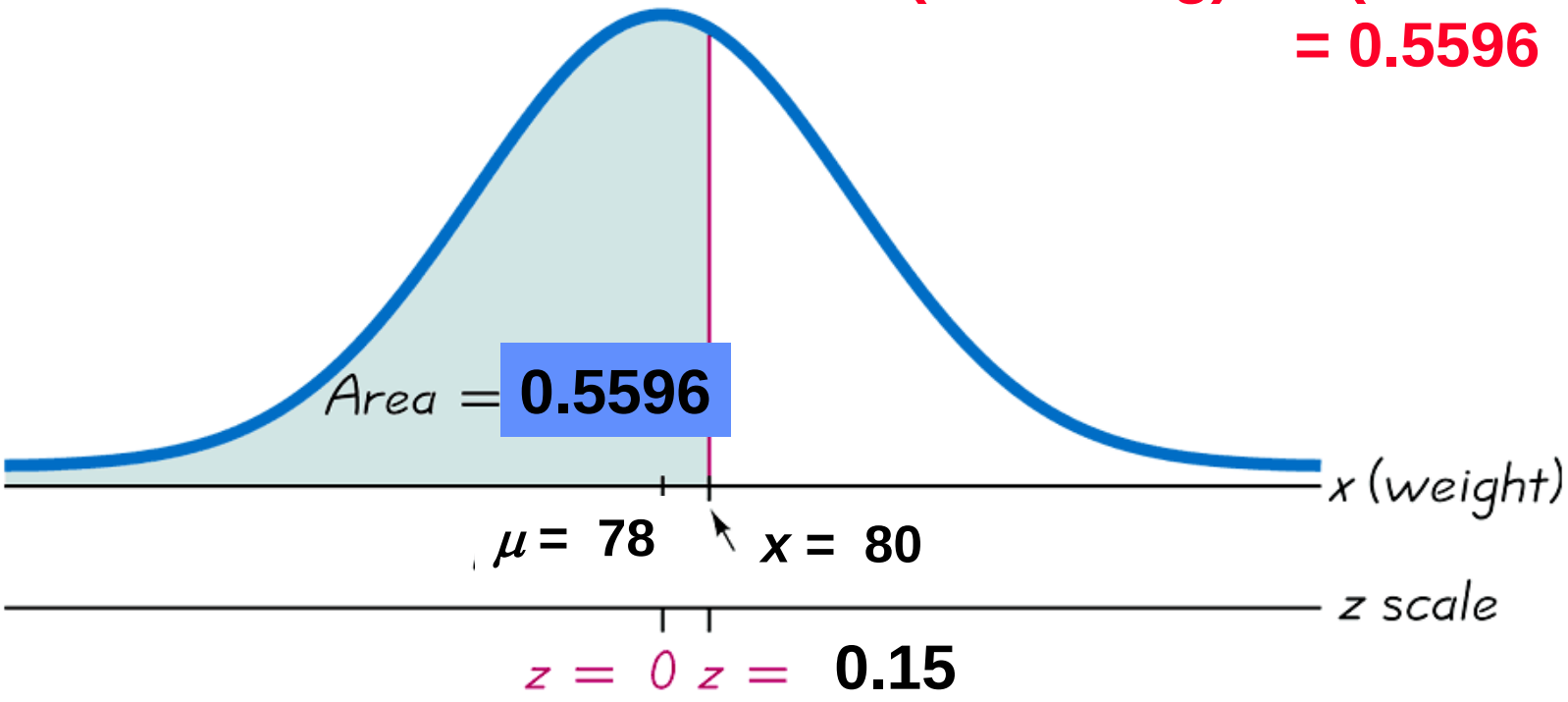
6-13. ábra

Példa - folyt

μ = 78

σ = 13

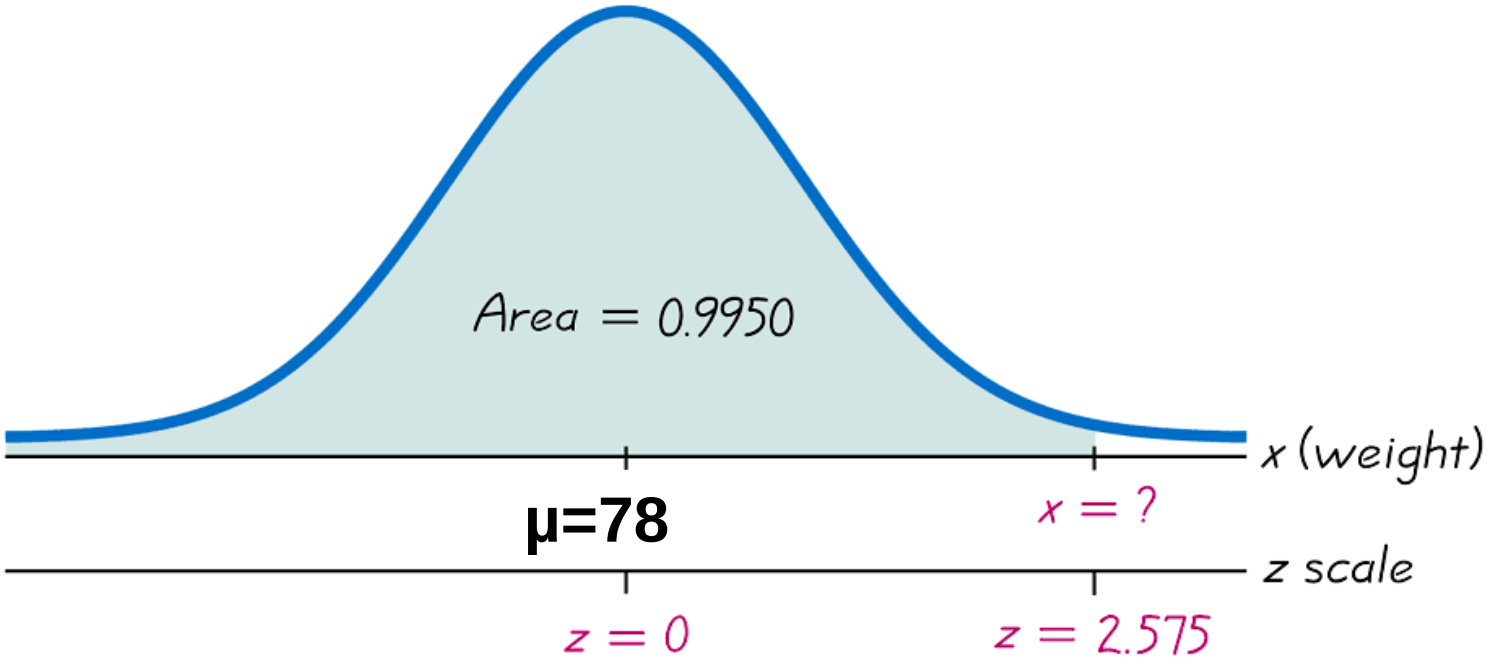
P (x < 80 kg) = P(z < 0.15)
= 0.5596



6-13. ábra

Példa – A legkönnyebb és a legnehezebb

A példa adatait használva határozzuk meg mekkora az a súly, ami a legkönnyebb 99.5%-ot elválasztja a legnehezebb 0.5%-tól?

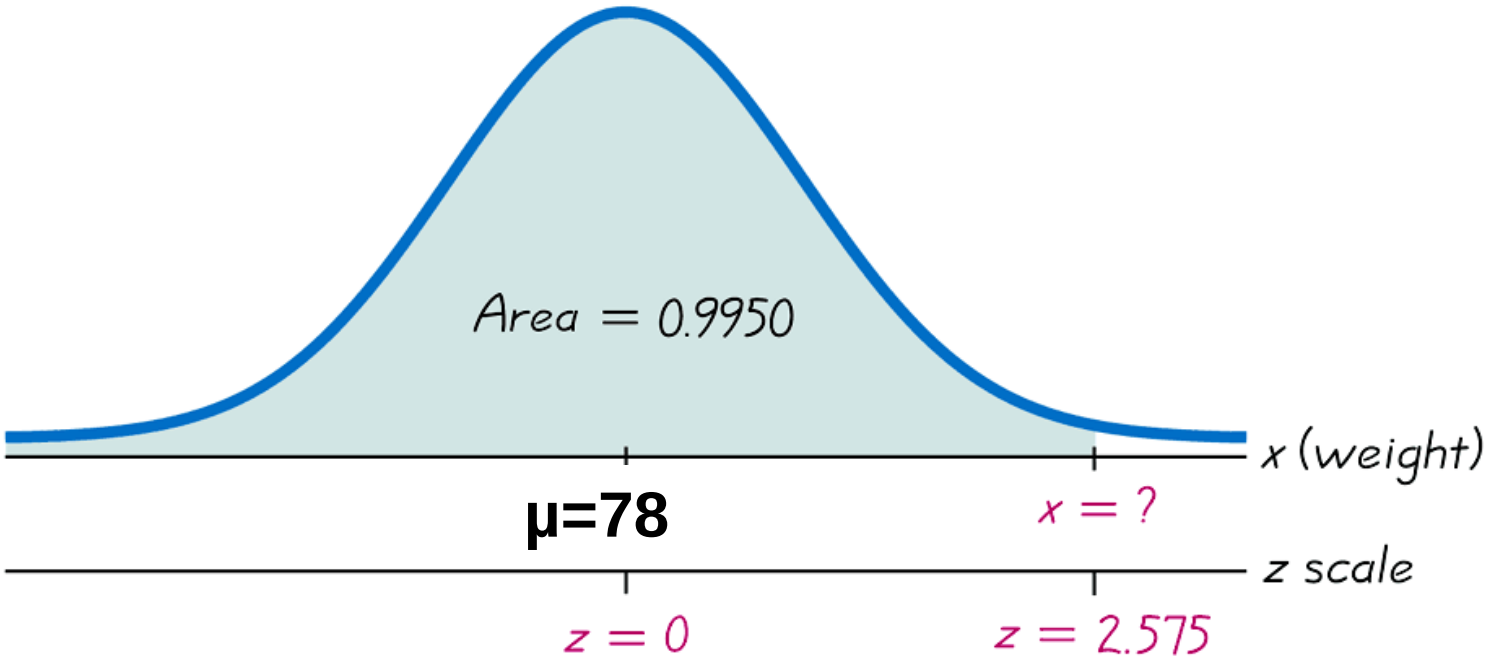


Példa – folyt

$x = \mu + (z \bullet \sigma)$

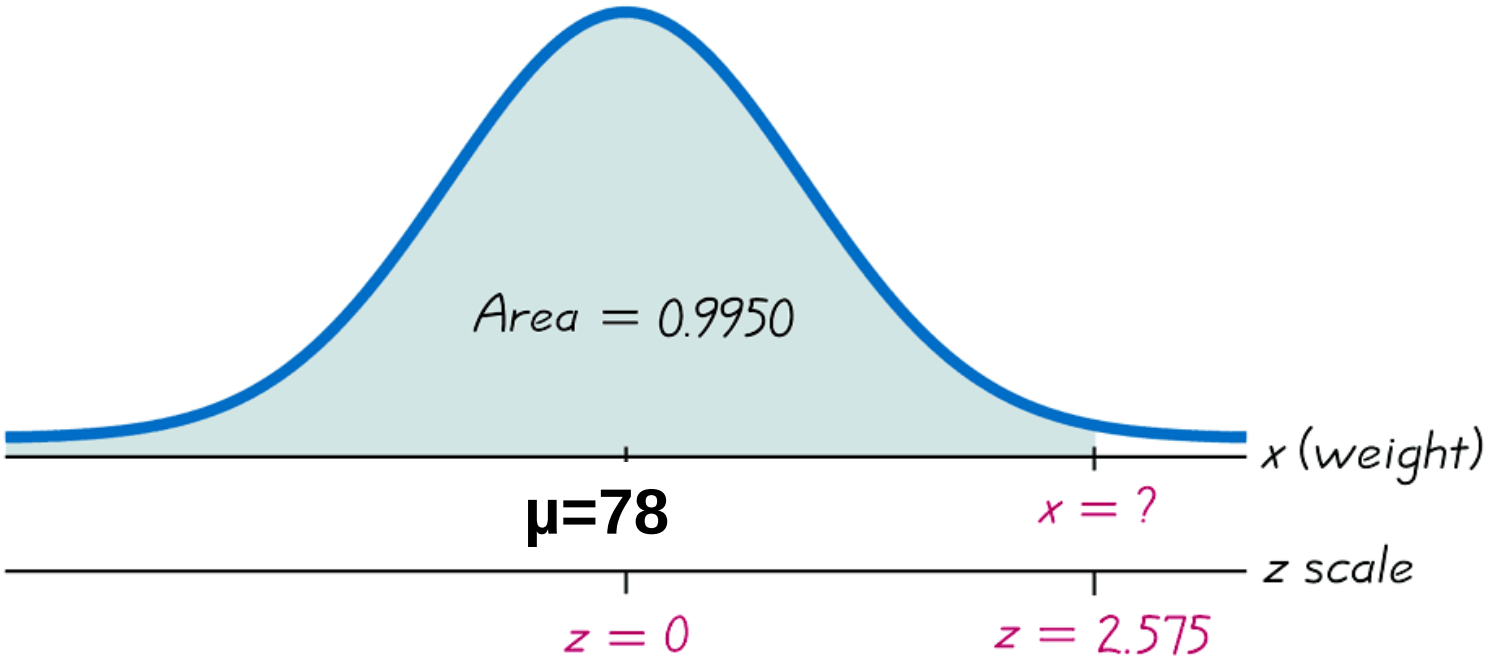
$x = 78 + (2.575 \bullet 13)$

$x = 111{,}475$



Példa – folyt.

Kb. 111 kg a választópont a 99.5% legkönnyebb és a 0.5% legnehezebb között.



Összefoglalás

Ebben a fejezetben megvitattuk:

- **A nem standard normális eloszlást.**
- **A standard normálisba konvertálást.**

6-4. fejezet
A statisztikák eloszlásai és becslések

Kulcsfogalmak

A fejezet célja, hogy bevezessük a statisztika eloszlását, ami az adott statisztika értékeinek eloszlása abban az esetben, amikor az értékeket a populációból kiválasztott minden lehetséges adott elemszámú mintára kiszámítjuk.

Látni fogjuk, hogy bizonyos statisztikák jobbak mint mások a populáció paramétereinek becslésére.

Definíció

❖ **A statisztika eloszlása (mint például a minta arány vagy a minta átlag eloszlása) a statisztika minden lehetséges értékének eloszlása abban az esetben, amikor értékét a populáció minden lehetséges *n* elemszámú mintájára kiszámítjuk.**

Definíció

❖ **Az arány eloszlása valami mintabeli arányának eloszlása, a populáció minden lehetséges *n* elemszámú mintájában.**

Tulajdonságok

- **A minta arányok a populációs arányhoz tartanak. (Azaz a lehetséges minták arányainak átlaga egyenlő az „igazi” populációs aránnyal.)**
- **Bizonyos feltételek mellett a mintabeli arányok eloszlása normális eloszlással közelíthető.**

Definíció

❖ **Az átlag eloszlása a minták átlagainak eloszlása abban az esetben, ha a populációból vett összes lehetséges *n* elemszámú mintát vesszük. (Az átlag eloszlását általában táblázatosan megadott valószínűség eloszlásként, hisztogramként vagy képlettel prezentáljuk.)**

Definíció

- **A statisztika értéke, mint például a minta átlag $\bar{x}$, függ a mintába kerülő konkrét értékektől, és általában mintáról mintára változik. A statisztikának ezt a variabilitását minta variabilitásnak nevezzük.**

Becslő függvények (becslések)

Bizonyos statisztikák sokkal jobbak, mint mások a populáció paramétereinek becslésére. A következő példa ezt mutatja be.

Példa

A populáció álljon az 1, 2, és 5 értékekből. Véletlenülszerűen, visszatevéssel választunk 2 elemszámú mintákat. Összesen 9 minta lehetséges.

a. Minden mintára megkeressük az átlagot, a mediánt, a terjedelmet, a varianciát és a szórást.

b. Mindegyik statisztikára számítsuk ki ezek átlagát.

Table 6-7 Sampling Distributions of Statistics (for Samples of Size 2 Drawn with Replacement from the Population 1, 2, 5)

Sample	Mean $\bar{x}$	Median	Range	Variance s^2	Standard Deviation s	Proportion of Odd Numbers	Probability
1, 1	1.0	1.0	0	0.0	0.000	1	1/9
1, 2	1.5	1.5	1	0.5	0.707	0.5	1/9
1, 5	3.0	3.0	4	8.0	2.828	1	1/9
2, 1	1.5	1.5	1	0.5	0.707	0.5	1/9
2, 2	2.0	2.0	0	0.0	0.000	0	1/9
2, 5	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 1	3.0	3.0	4	8.0	2.828	1	1/9
5, 2	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 5	5.0	5.0	0	0.0	0.000	1	1/9
Mean of Statistic Values	8/3	8/3	16/9	26/9	1.3	2/3	
Population Parameter	8/3	2	4	26/9	1.7	2/3	
Does the sample statistic target the population parameter?	Yes	No	No	Yes	No	Yes	

Interpretáció

Láthatjuk, hogy bizonyos statisztikák jók abban az értelemben, hogy a populáció paramétereihez tartanak. Az ilyen statisztikákat torzítatlan becsléseknek nevezik.

Olyan statisztikák, melyek a populációs paraméterekhez tartanak: átlag, variancia, részarány

Olyan statisztikák, melyek nem tartanak a populáció paramétereihez: medián, terjedelem, szórás

A populáció és a minta szórása közti különbség: az átlag ingadozása

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$$

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$

2-4. képlet

Példa: 3, 6, 9

N=3 $\mu = 6$ $\sigma^2 = ((3-6)^2 + (6-6)^2 + (9-6)^2)/3 = 6.0$

n=2

3,6 és 6,3 $\overline{x} = 4.5$ $s^2 = ((3-4.5)^2 + (6-4.5)^2)/1 = 4.5$

6,9 és 9,6 $\overline{x} = 7.5$ $s^2 = ((6-7.5)^2 + (9-7.5)^2)/1 = 4.5$

3,9 és 9,3 $\overline{x} = 6$ $s^2 = ((3-6)^2 + (9-6)^2)/1 = 18.0$

3,3 $\overline{x} = 3$ $s^2 = ((3-3)^2 + (3-3)^2)/1 = 0.0$

6,6 $\overline{x} = 6$ $s^2 = ((6-6)^2 + (6-6)^2)/1 = 0.0$

9,9 $\overline{x} = 9$ $s^2 = ((9-9)^2 + (9-9)^2)/1 = 0.0$

$(4.5+4.5+4.5+4.5+18.0+18.0+0.0+0.0+0.0)/9=54.0/9=6.0$

Összefoglalás

Ebben a fejezetben megvitattuk:

- **Statisztika eloszlását.**
- **Az arány eloszlását.**
- **Az átlag eloszlását.**
- **A minta variabilitását.**
- **Becsléseket.**

6-5. fejezet
A központi határeloszlás tétel

Kulcsfogalmak

Ebben a fejezetben megalapozzuk a populáció paramétereinek becslését és a hipotézis vizsgálatokat, melyről a következő előadások szólnak majd.

Központi határeloszlás tétel

Adott:

1. Az *x* véletlen változónak μ átlaga és σ szórással rendelkező eloszlása van (ami vagy normális vagy sem).
2. Egyszerű *n* elemszámú véletlen mintákat választunk a populációból. (A mintákat úgy választjuk, hogy bármely *n* elemszámú mintát ugyanazzal az eséllyel választunk ki.) A minták egymástól függetlenek.

Központi határeloszlás tétel – folyt.

Konklúziók:

1. A minta átlag eloszlása $\bar{x}$,
   ahogy a minta méretét növeljük,
   a normális eloszláshoz tart.

2. A minta átlagok átlaga μ.

3. A minta átlagok szórása pedig $\sigma/\sqrt{n}$.

Általános gyakorlati tanácsok

1. **Általában ha a minta *n* mérete nagyobb mint 30, akkor a minta átlagok eloszlását meglehetősen jól lehet normális eloszlással közelíteni. A közelítés egyre jobb, ahogy *n* növekszik.**
2. **Ha az eredeti populáció maga is normális eloszlású, akkor a minta átlagok eloszlása mindig normális bármely *n-re* (nem csak a 30-nál nagyobb értékek esetén).**

Jelölés

a minta átlagok átlaga

$$\mu_{\bar{x}} = \mu$$

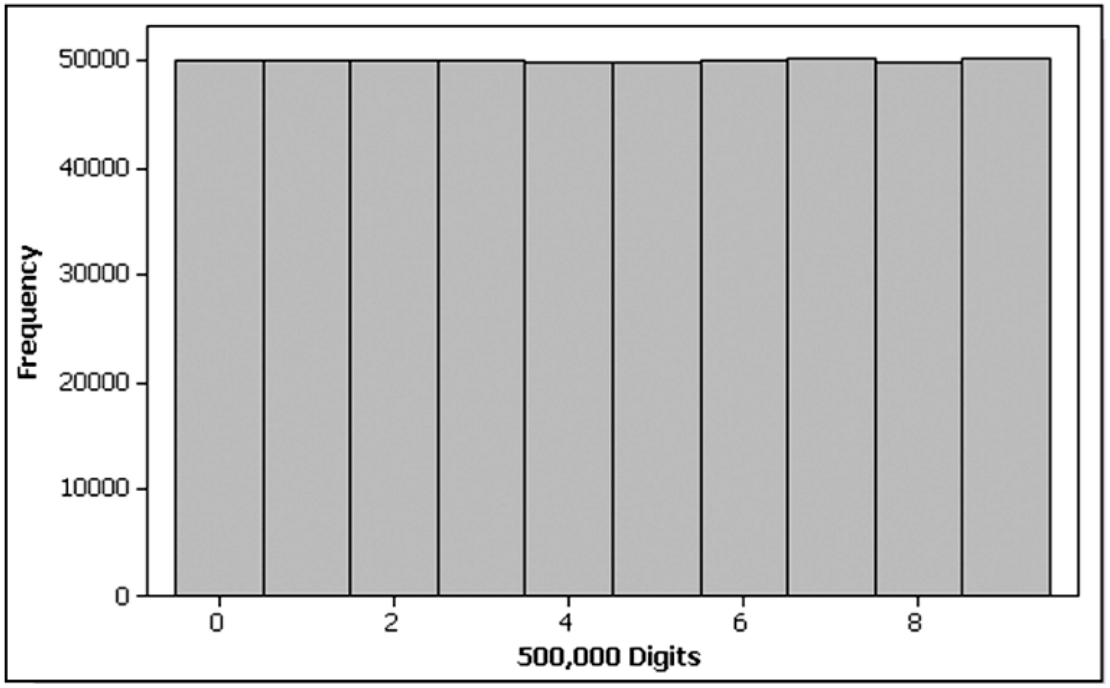
a minta átlagok szórása

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(gyakran az átlag standard hibájának is nevezik)

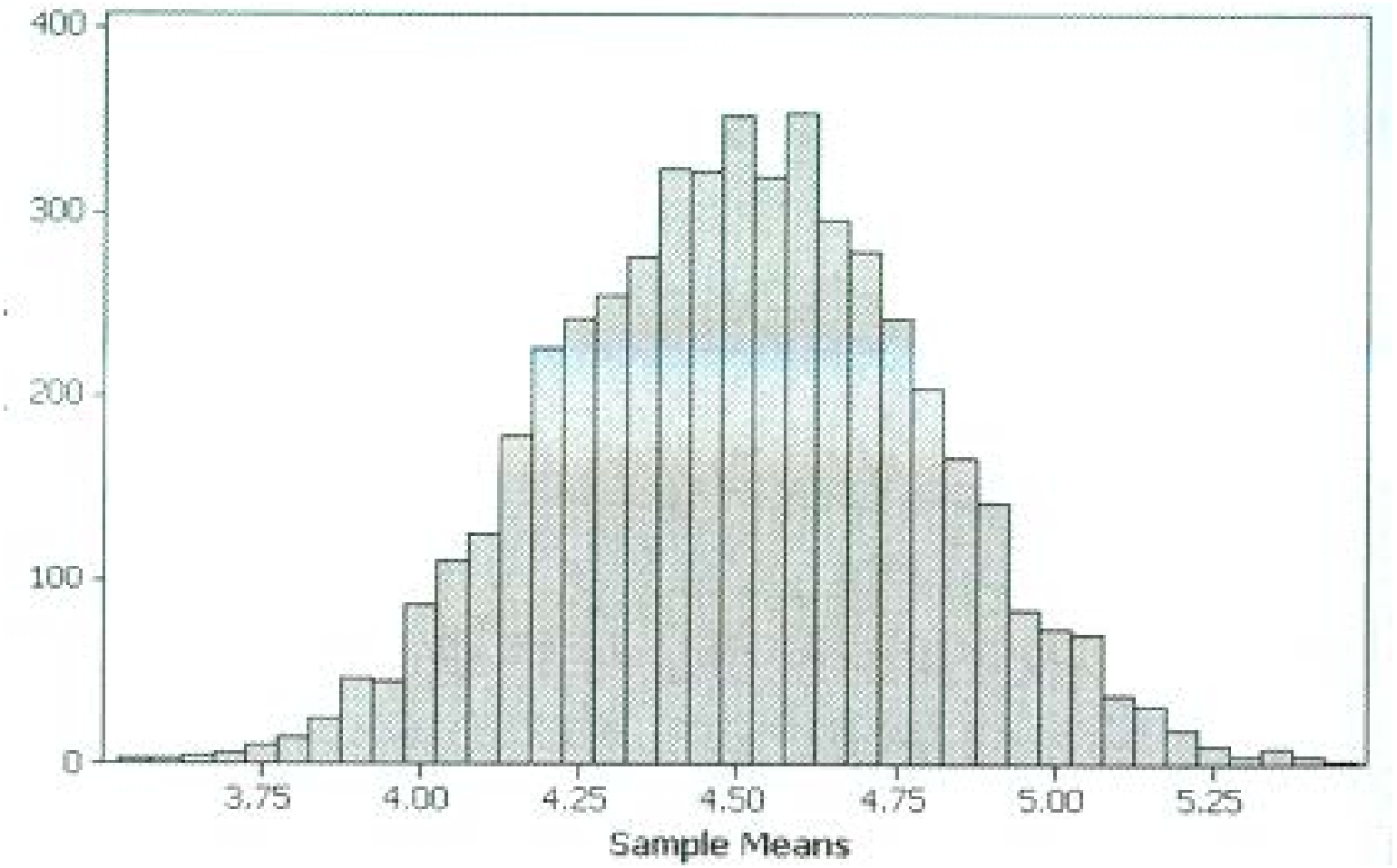
Szimuláció véletlen számokkal

Generáljunk 500,000 véletlen 0 és 9 közötti egész számot, csoportosítsuk 5000 mintába, mindegyikben 100 számmal. Keresd meg mindegyik minta átlagát!



Annak ellenére, hogy az eredeti 500,000 szám egyenletesen oszlik el, az 5000 minta átlag eloszlása normális eloszlás lesz!

5000 db 100 elemű minta átlagainak eloszlása



Fontos felismerés

Ahogy a minta nagyság nő, a minta átlag eloszlása egyre inkább normális lesz.

Példa – vízitaxi biztonság

A férfiak egy adott populációjának tömege normális eloszlású, átlagosan 78 kg a súlya 13 kg szórással,

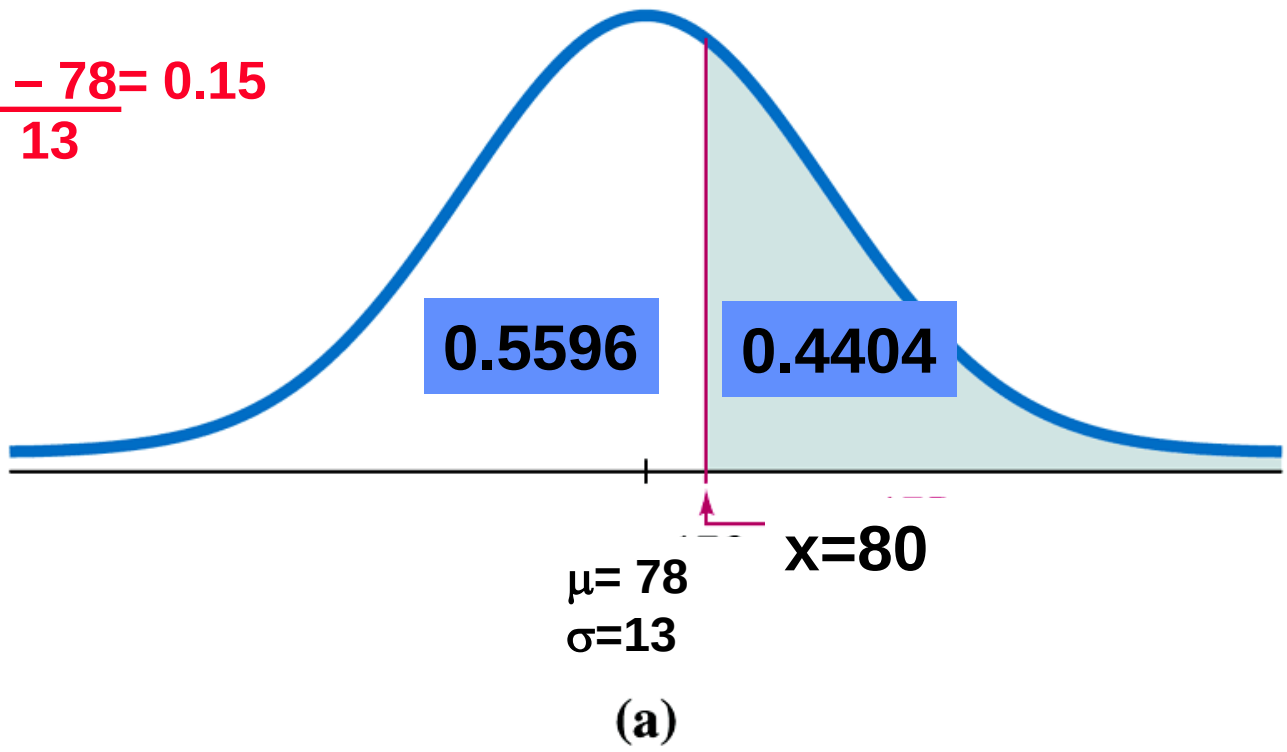
a) ha kiválasztunk egy férfit, mi a valószínűsége annak, hogy a tömege több mint 80 kg.

b) ha 20 különböző férfit véletlenül választunk, számítsuk ki, hogy mi annak a valószínűsége, hogy átlagsúlyuk meghaladja a kritikus 80 kg-ot.

Példa – folyt.

a) egy embert kiválasztva határozzuk meg, hogy mi a valószínűsége annak, hogy tömege több mint 80 kg.

$$z = \frac{80 - 78}{13} = 0.15$$

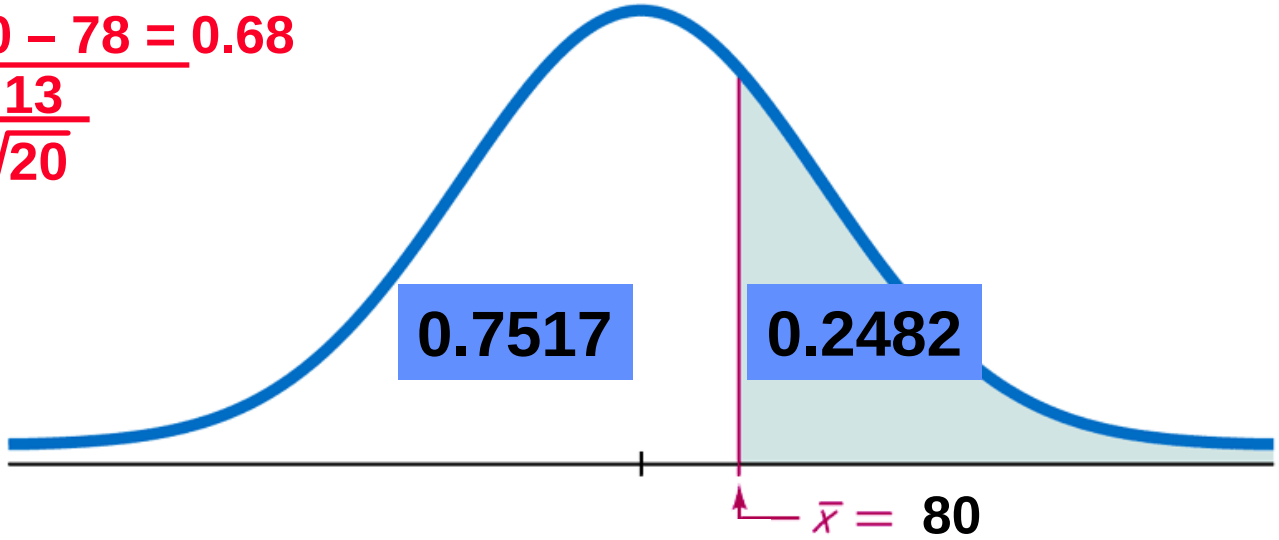


(a)

Példa – folyt

b) ha 20 különböző férfit választunk véletlenül, számítsuk ki annak a valószínűségét, hogy átlagsúlyuk több mint 80 kg.

$$z = \frac{80 - 78}{\frac{13}{\sqrt{20}}} = 0.68$$



$\mu_{\bar{x}} =$ 78

$(\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} =$ 13/√20=2,906

(b)

Példa – folyt.

a) egy véletlenül kiválasztott férfinál annak a valószínűsége, hogy 80 kg-nál nehezebb

$$P(x > 80) = 0.4404$$

b) véletlenül kiválasztott 20 férfi esetén annak a valószínűsége, hogy átlagosan nehezebbek mint 80 kg

$$P(\bar{x} > 80) = 0.2482$$

Egyvalaki esetén sokkal valószínűbb, hogy 80 kg-nál nagyobb, mint hogy 20 férfi esetében az átlaguk nagyobb, mint 80 kg.

Az eredmények értelmezése

Ha a biztonságos kapacitás 1600 kg, akkor elég nagy esélye van annak (24%-os valószínűsége), hogy 20 férfi tömege ezt meg fogja haladni!

Véges populációs korrekció

Ha visszatevés nélkül mintavételezünk, és a minta *n* mérete nagyobb mint 5%-a a véges *N* elemű populációnak, akkor a mintaátlag szórását korrigálnunk kell az alábbi faktorral:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

véges populációs korrekciós faktor

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A központi határeloszlás tételt.**
- **Praktikus megfontolásokat.**
- **A mintaméret hatását.**
- **Véges populációs korrekciót.**

6-6. fejezet
A binomiális közelítése normálissal

Kulcsfogalmak

Ebben a fejezetben megmutatjuk, hogy hogyan lehet egy binomiális eloszlást normális eloszlással közelíteni.

Ha az $np \geq 5$ és az $nq \geq 5$ feltételek egyszerre teljesülnek, akkor a binomiális eloszlást egy $\mu = np$ átlagú és $\sigma = \sqrt{npq}$ szórású normális eloszlással jól közelíthető.

Példa

- Egy Boeing 767-300 repülőn 213 ülőhely van.
- A nők átlag tömege 65 kg, a férfiaké 78 kg.
- Ha 122 férfinél több van, akkor vigyázni kell az utasok ültetésére
- Tegyük fel, hogy 50-50% a férfi és nő utasok valószínűsége
- Mi annak a valószínűsége, hogy legalább 122 férfi utas van a gépen.
- Az eloszlás binomiális, de nekünk most 92 esetre kellene kiszámítanunk …

Áttekintés

Binomiális eloszlás

1. A véletlen kísérletek száma állandó.
2. A kísérletek függetlenek.
3. Minden kísérletnek két kimenete van.
4. A siker valószínűsége állandó a kísérletek során.

.

A binomiális közelítése normális eloszlással

$np \geq 5$

$nq \geq 5$

ekkor $\mu = np$ és $\sigma = \sqrt{npq}$

és a véletlen változó

eloszlása  .

(normal)

A binomiális normálissal való közelítése

1. **Bizonyosodj meg, hogy $np \geq 5$ és $nq \geq 5$ tényleg fennáll.**

2. **Számítsd ki a μ és σ paraméterek értékeit a $\mu = np$ és $\sigma = \sqrt{npq}$ képlettel.**

3. **Azonosítsd x diszkrét értékeit (a sikerek számát). A diszkrét x értéket helyettesítsük az $x - 0.5$ -től $x + 0.5$ – ig intervallummal. (Ld. folytonossági korrekciók még ebben a fejezetben.) Rajzoljuk meg a normális görbét μ , σ, paraméterekkel.**

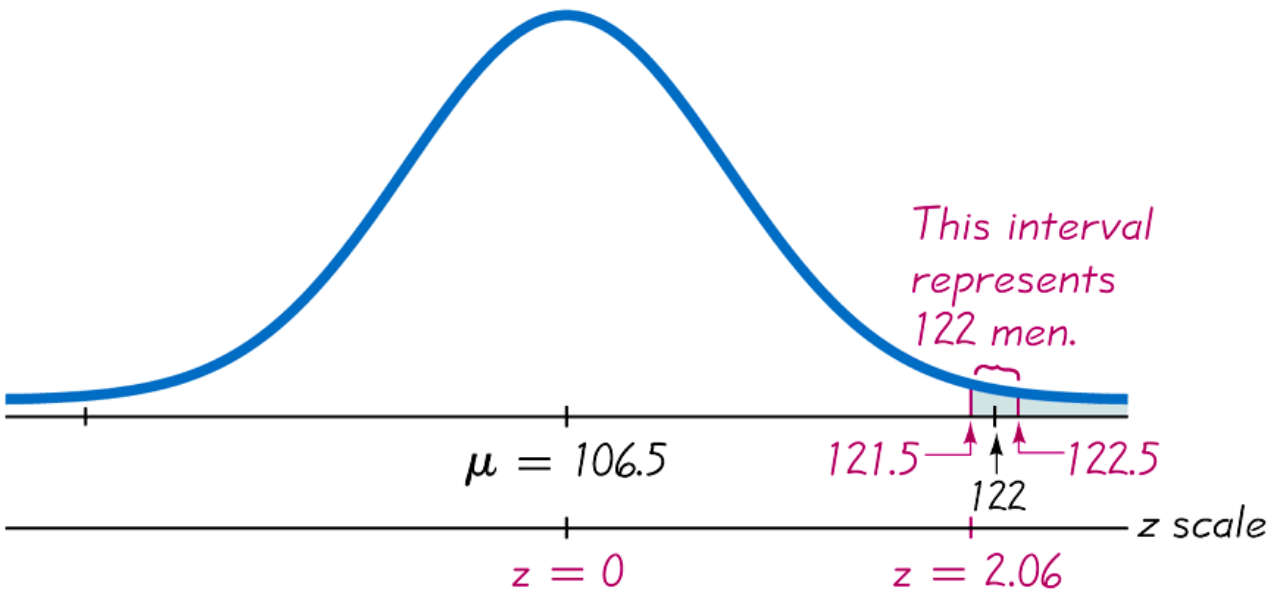
A binomiális normálissal való közelítése Folyt.

4. Helyettesítsük *x* –et vagy *x* – 0.5 –el, vagy *x* + 0.5 -el, a feladatnak megfelelően.

5. Az *x* – 0.5 vagy *x* + 0.5 értéket (a feladatnak megfelelően) használva *x helyett,* keresd meg a kívánt valószínűséget úgy, hogy először a megfelelő *z* értékhez kikeresed a tőle balra fekvő területet.

Példa – A férfiak száma az utasok között

A “legalább 122 férfi” valószínűségének meghatározása 213 utas esetén



6-21. ábra

Definíció

Amikor a normális eloszlást használjuk (ami egy folytonos eloszlás) arra, hogy a binomiálist közelítsük (ami pedig diszkrét), egy folytonossági korrekciót kell végrehajtanunk és a diszkrét egész *x-et* a

$$x - 0.5 \text{ -tól } x + 0.5 \text{ –ig}$$

intervallummal kell helyettesíteni

(hozzá kell adni és levonni 0.5-öt).

A folytonossági korrekció menete

1. **Ha a binomiálist normálissal közelíted, mindig használd a folytonossági korrekciót.**

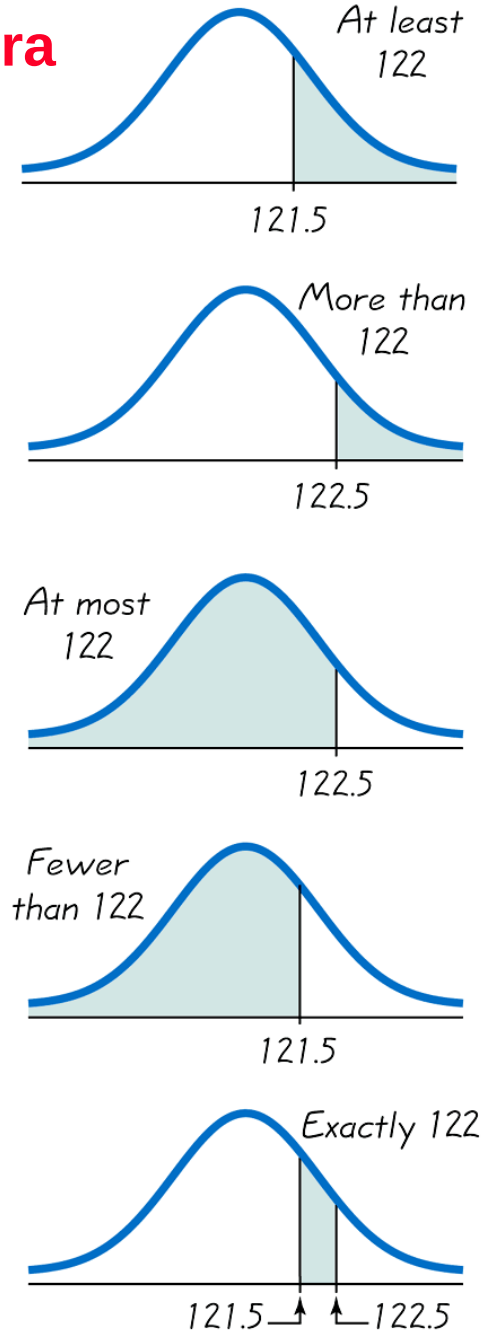
2. **Először keresd meg a diszkrét egész *x*-*et* a binomiális problémánál.**

3. **Rajzolj egy normális eloszlást, *μ* átlag köré, és rajzolj egy függőleges *x*-re centrált sávot *x* – 0.5 és *x* + 0.5 határokkal. Példánkban *x* = 122, rajzoljunk be egy sávot 121.5-nél és 122.5-nél. A berajzolt terület reprezentálja a diszkrét egész *x* érték valószínűségét.**

- folyt.

4. **Aztán gondold meg, hogy *x* maga benne van-e abban a valószínűségben, amit ki akarsz számítani. Utána gondold meg, hogy a „legalább *x*”, „legfeljebb *x*”, „több mint *x*”, „kevesebb mint *x*”, vagy „pontosan *x*” valószínűségére van-e szükséged. Satírozd be a sávtól balra vagy jobbra eső területet és a sávot magát is akkor, és csak akkor ha *x* maga is benne van. A teljes besatírozott terület adja a keresett valószínűséget, amit keresünk.**

6-22. ábra



X = **legalább** 122
(tartalmazza 122-t és felette)

X = **több mint** 122
(nincs benne a 122)

X = **legfeljebb** 122
(tartalmazza 122-t és alatta)

X = **kevesebb mint** 122
(nem tartalmazza 122-t)

X = **pontosan** 122

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A binomiális normálissal való közelítését.**
- **A normális közelítés procedúráját.**
- **A folytonossági korrekciókat.**

6-7. fejezet
A normalitás vizsgálata

Kulcsfogalmak

Ebben a fejezetben meghatározzuk, hogy valamilyen eloszlás mikor tekinthető normálisnak.
A kritériumok eddig:

- **a hisztogram vizuális megfigyelése és a haranggörbével való összehasonlítása**
- **az outlierek azonosításan**

Most:
a normális kvantilis-kvantilis plot módszer

Módszerek az adatok normalitásának vizsgálatára

1. **Hisztogram*:*** **Készíts hisztogramot. Ha eltér a haranggörbétől, akkor vesd el a normalitást.**

2. **Outlierek*:*** **Keresd meg az outliereket. Ha több mint egyet találsz, vesd el a normalitást.**

3. **Normál QQ plot: Ha a hisztogram alapvetően szimmetrikus, és legfeljebb egy outlier van, készítsd el a normál QQ plotot az alábbi módon:**

Definíció

- **Normál QQ plot (vagy normál valószínűség plot) egy pontokból (*x*,*y*) álló grafikon, ahol**
  - **az *x* érték az eredeti minta adatokból áll**
  - **az *y* érték a standard normális eloszlásból származó kvantilis értéknek megfelelő *z* érték.**

- folyt

3. Normál QQ plot

a. Rendezd sorba az adatokat a legkisebbtől a legnagyobbik irányában.

b. A *n* elemű minta esetén, minden érték a minta 1/*n*-ed részét jelenti. Használva az *n értékét*, határozd meg az 1/2*n*, 3/2*n*, 5/2*n*, 7/2*n*, … területeket. Ezek lesznek a megfelelő minta értéktől balra esés valószínűségei.

c. Felhasználva a standard normális eloszlást (táblázat , szoftver vagy kalkulátor) számítsd ki a fenti területekhez tartozó *z* értékeket.

- folyt

d. Párosítsd a kiszámított z értékeket az x értékekkel, majd készítsd el az (*x*, *y*) grafikont, ahol *x* az eredeti adatok és *y* a megfelelő *z* érték.

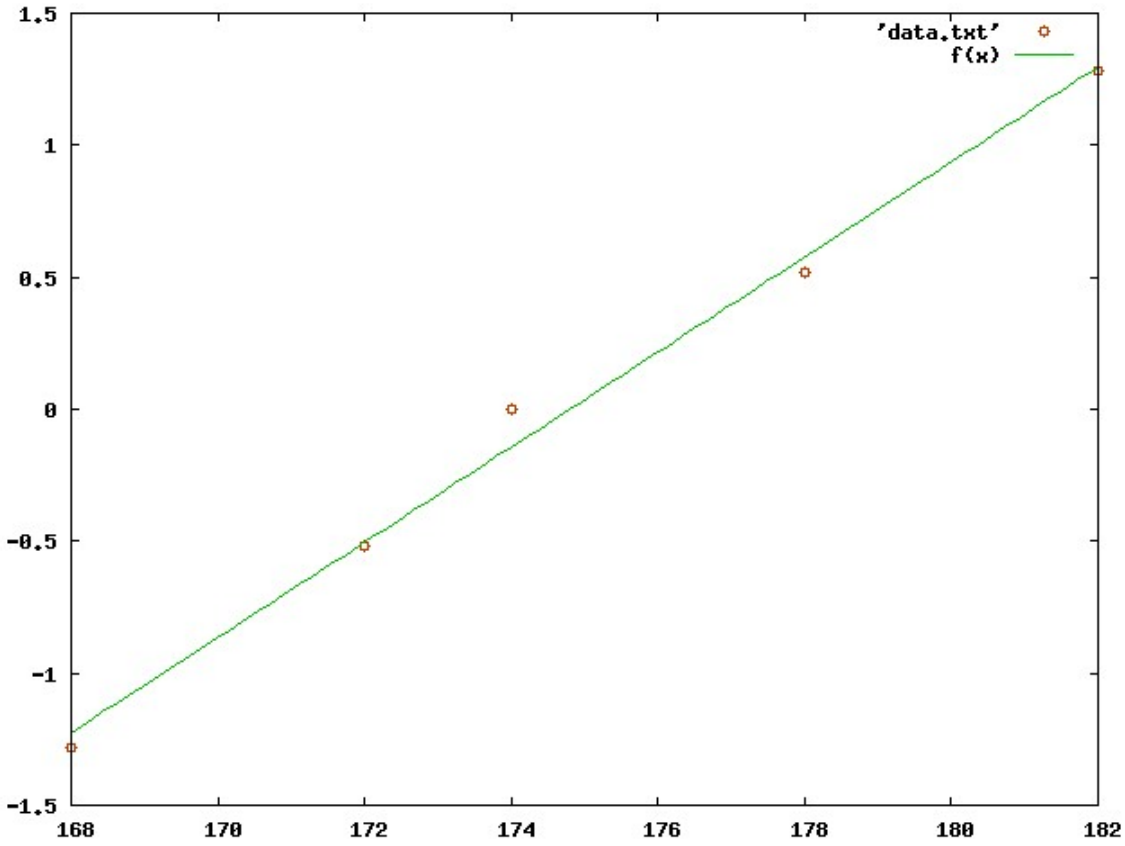
e. Vizsgáld meg az így készített QQ plotot az alábbi kritériumok alapján:

Ha az adatok nem fekszenek egy egyenesen, vagy valamilyen szisztematikus, de nem egyenes alakzatot öltenek, akkor az adatok nem normális eloszlással rendelkező populációból származnak. Ha az adatok elfogadhatóan közel vannak egy egyeneshez, akkor a populáció normálisnak tűnik.

Példa

- Vegyünk emberek magasságának adatait
- Elég pl. 5-öt 178, 168, 182, 172, 175
- n=5 minden adat 1/5-öde a teljesnek
- területek: 0.1,0.3,0.5,0.7 és 0.9
  (nem 0-tól 1-ig megy, hanem 1/2n-től 1-1/2n-ig)
- z= -1.28, -0.52, 0, 0.52 és 1.28
- (x,y)= (168, -1.28) (172, -0.52) (175,0)
  (178,0.52) (182, 1.28)

Példa



Interpretáció: Mivel a pontok elfogadhatóan közel vannak egy egyeneshez és nem látszik bennük semmilyen más szisztematikus eltérés, arra következtetünk, hogy az eredeti adatok egy normális populációból származnak.

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A normál QQ plotot.**
- **Azt a procedúrát, amivel eldönthetjük, hogy az adatok normális eloszlásúak-e.**