Valószínűségszámítás és Statisztika

10. előadás

2017. november 26.

Konfidenciaintervallum

- Olyan intervallum, mely legalább 1-α valószínűséggel tartalmazza a keresett paramétert: $P_\theta\left(T_1(X) < \theta < T_2(X)\right) \geq 1-\alpha$
- Példák:
  - [0, θ] intervallumon egyenletes eloszlású minta esetén a paraméterre.
  - Normális eloszlás várható értékére

Példa (normális eloszlás)

- A Gyorskenyér Kft automata kenyérsütő készülékei egyszerre 100 kenyeret sütnek ki. Ezek tömegei grammban mérve $N(m,10^2)$ eloszlással közelíthetőek, ahol m a kezelő beállításától függ. Egy ellenőrzésnél megmérték mind a 100 kenyér tömegét. Az átlag 990 g volt. Készítsünk 95%-os megbízhatóságú konfidencia intervallumot m-re!

Konfidencia intervallum normális eloszlás várható értékére (ismert szórás esetén)

$$\xi_1,\ldots,\xi_n \sim N(m,\sigma^2),\ \sigma \text{ ismert},\ \Phi(u_y)=y \Rightarrow$$

$$P\left(\bar{\xi}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}<m<\bar{\xi}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha,$$

$$P\left(m>\bar{\xi}-u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha,$$

$$P\left(m<\bar{\xi}+u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

Konfidencia intervallum várható értékre (ismert szórás esetén)

$$\xi_1,\ldots,\xi_n,\ E\xi_i = m,\ D^2\xi_i = \sigma^2,\ \sigma \text{ ismert} \Rightarrow$$

$$P\left(\overline{\xi} - \sqrt{\frac{1}{\alpha}}\frac{\sigma}{\sqrt{n}} < m < \overline{\xi} + \sqrt{\frac{1}{\alpha}}\frac{\sigma}{\sqrt{n}}\right) \geq 1-\alpha.$$

α	$u_{1-\alpha/2}$	$\sqrt{\frac{1}{\alpha}}$
10%	1,64	3,16
5%	1,96	4,47
2,50%	2,24	6,32
1%	2,58	10,00

Konfidencia intervallum ”sok” megfigyelés esetén

$$\xi_1, \ldots, \xi_n,\ D^2\xi_i = \sigma^2 \text{ ismert} \Rightarrow$$

$$P\left(\bar{\xi} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < m < \bar{\xi} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \sim 1 - \alpha.$$

Példák (milyen valószínűséggel születik fiúgyermek?)

- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.
- Fiúk relatív gyakorisága így 0,5141.

$$p(1-p) \le \frac{1}{4} \Rightarrow$$

$$P\left(\bar{\xi} - \frac{u}{2\sqrt{n}} < p < \bar{\xi} + \frac{u}{2\sqrt{n}} \right) \sim 2\Phi(u) - 1$$

Esetünkben $0{,}9973$ valószínűséggel $0{,}5132 \le p \le 0{,}5150$

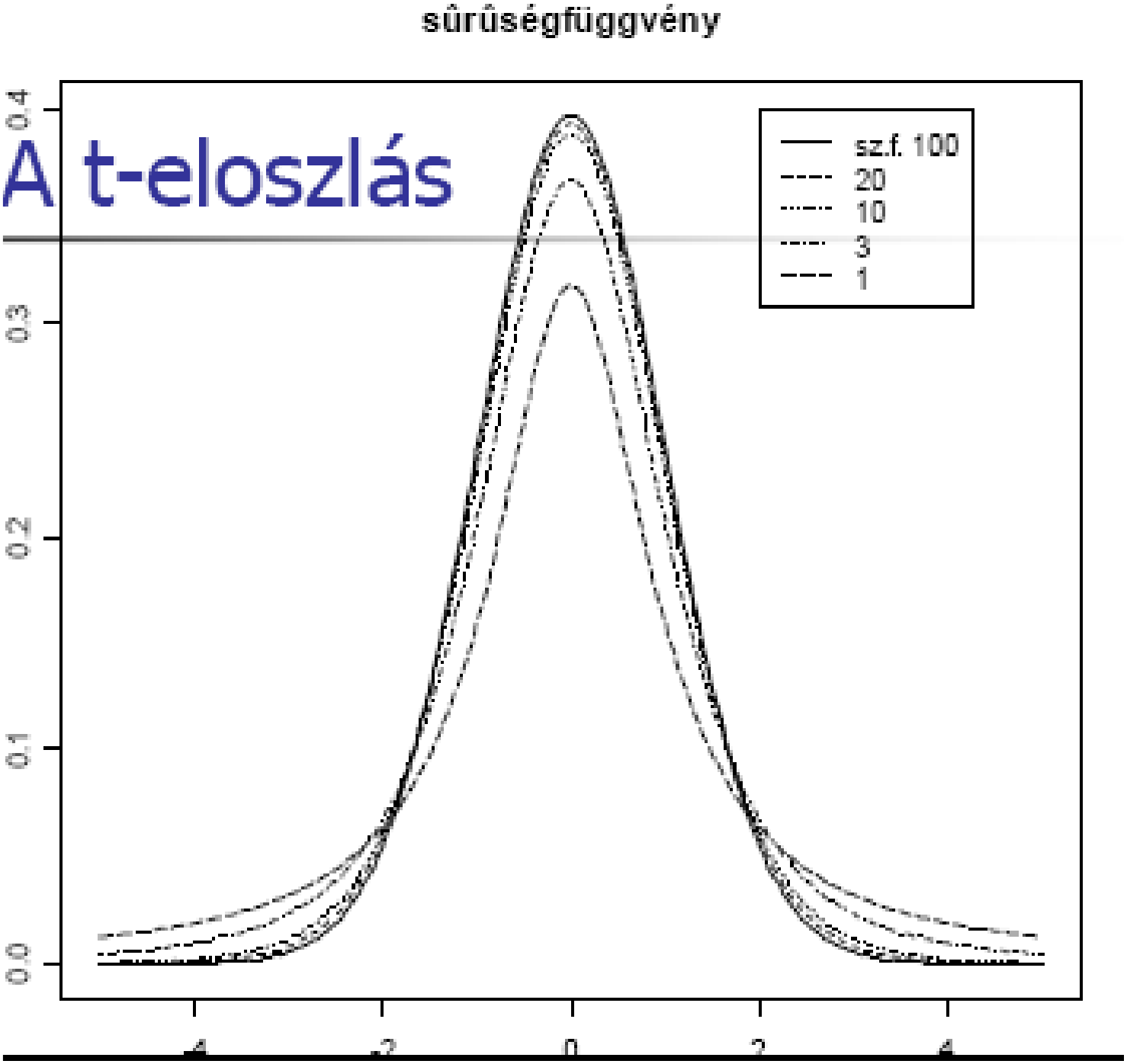
Konfidencia intervallum normális eloszlás várható értékére (ismeretlen szórás esetén)

- Ha a szórás nem ismert, becsüljük
- Tétel (biz. nélkül): normális eloszlású minta esetén a mintaátlag és a tapasztalati szórás független
- *n*-1 szabadságfokú t (Student) eloszlás:

$$X_0, X_1, \ldots, X_n \ \textbf{\textit{független}}\ N(0,1)$$

$$\frac{X_0}{\sqrt{\left(X_1^2 + \ldots + X_n^2\right)/n}} \sim t_n$$

A t-eloszlás



Konfidencia intervallum normális eloszlás várható értékére (ismeretlen szórás esetén) (folyt.)

$$\xi_1,\ldots,\xi_n \sim N(m,\sigma^2), \tilde{\sigma}^2 = \left(\left(\xi_1-\bar{\xi}\right)^2+\ldots+\left(\xi_n-\bar{\xi}\right)^2\right)/(n-1) \Rightarrow$$

$$\frac{\sqrt{n}\left(\bar{\xi}-m\right)}{\sqrt{\tilde{\sigma}^2}} \sim t_{n-1}$$

$$P(t_{n-1} < t_{n-1,y}) = y$$

$$P\left(\bar{\xi}-t_{n-1,1-\alpha/2}\frac{\tilde{\sigma}}{\sqrt{n}} < m < \bar{\xi}+t_{n-1,1-\alpha/2}\frac{\tilde{\sigma}}{\sqrt{n}}\right)=1-\alpha,$$

$$P\left(m > \bar{\xi}-t_{n-1,1-\alpha}\frac{\tilde{\sigma}}{\sqrt{n}}\right)=1-\alpha,$$

$$P\left(m < \bar{\xi}+t_{n-1,1-\alpha}\frac{\tilde{\sigma}}{\sqrt{n}}\right)=1-\alpha$$

Példa (kenyér. folyt.)

- Tegyük fel most, hogy nem ismerjük Gyorskenyér Kft kenyereinek szórását. Az átlag 990 g volt.
- Ismert 10 szórásnál 991,6 g volt a 95%-os megbízhatóságú felső konfidencia határ.
- Amennyiben a korrigált tapasztalati szórás is 10, akkor ez a határ csak kis mértékben változik (991,8 g).
- Azonban 50-es korrigált tapasztalati szórásnál ez az érték 999 g-ra változik.

u és *t* együtthatók összehasonlítása

$$u_{1-5\%} = 1,64 \qquad (\Phi(1,64) = 95\%)$$

n	$t_{n-1,1-5\%}$
2	6,31
3	2,92
4	2,35
5	2,13
10	1,83
20	1,73
50	1,68
100	1,66
1000	1,65

Hipotézisvizsgálat

- H_0 nullhipotézis (jelezni akarjuk, ha nem igaz) $\theta \in \Theta_0$.
- H_1 ellenhipotézis $\theta \in \Theta_1$.
- Elsőfajú hiba: H_0 igaz, de elutasítjuk
- Másodfajú hiba: H_0 hamis, de elfogadjuk
- Példák:
  - 2 kocka közül melyikkel dobunk?
  - mekkora a fejdobás valószínűsége?

Alapfogalmak

- Emlékeztető: $\mathbf{X}$ mintatér: a minta lehetséges értékeinek halmaza.
- $\mathbf{X} = \mathbf{X}_e \cup \mathbf{X}_k$
- $\mathbf{X}_k$: azon lehetséges értékek halmaza, amelyek megfigyelése esetén elutasítjuk a nullhipotézist.
- Gyakran statisztika segítségével határozzuk meg:

$$T(\mathbf{x}) = \begin{cases} 1 & , \ \mathbf{x} \in \mathbf{X}_k \\ 0 & , \ \mathbf{x} \notin \mathbf{X}_k \end{cases}$$

Lehetséges döntések táblázata

		Aktuális helyzet	
		A nullhipotézis igaz	A nullhipotézis hamis
Döntés:	Elfogadjuk a nullhipotézist	Helyes döntés	Másodfajú hiba
	Elutasítjuk a nullhipotézist	Elsőfajú hiba	Helyes döntés

Elsőfajú hiba valószínűsége

α a próba terjedelme, ha minden $\vartheta \in \Theta_0$-ra

$P_\vartheta\left(\xi \in \mathrm{X}_k\right) \le \alpha$

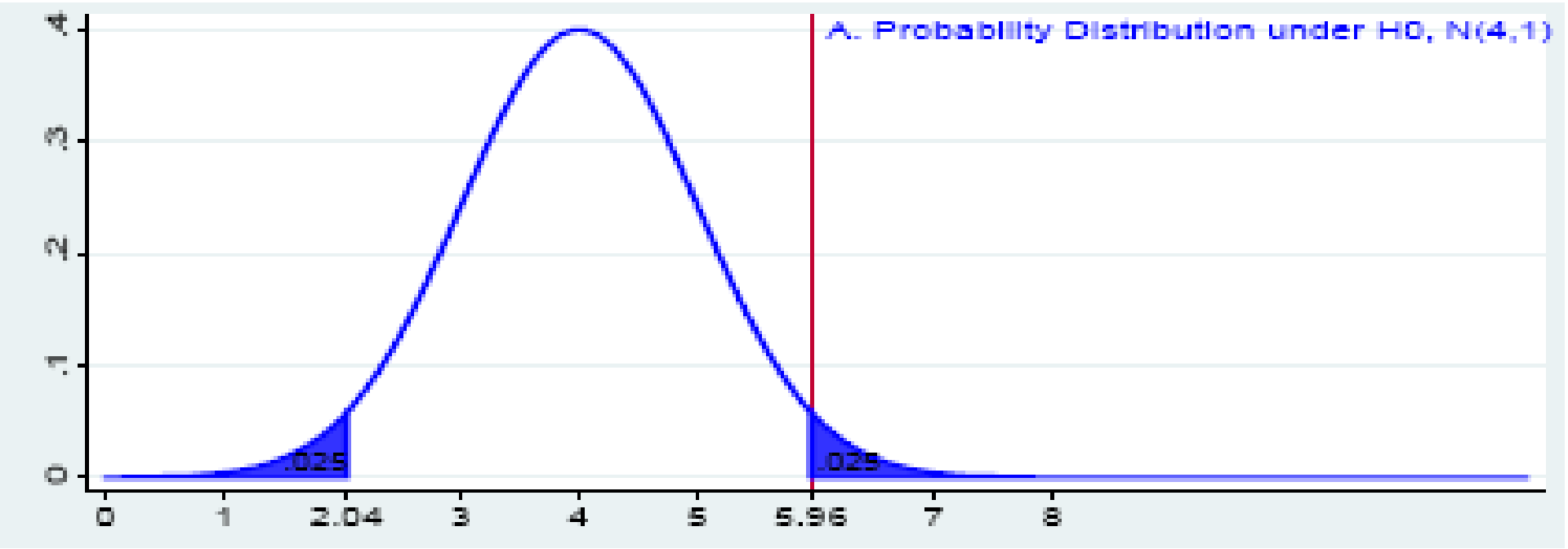
α a próba szignifikanciaszintje

(másképp: a próba pontos terjedelme),

$\sup_{\vartheta \in \Theta_0} P_\vartheta\left(\xi \in \mathrm{X}_k\right) = \alpha$

Példa (egyetlen megfigyelés)

H_0: a megfigyelés N(4,1) eloszlású



Példa (sörök megkülönböztetése)

- Ki tudják-e választani a különböző sört?
- 24 emberen kísérleteztek.

$$H_0 : p = 1/3,\ H_1 : p > 1/3$$

Az eloszlás H_0 esetén

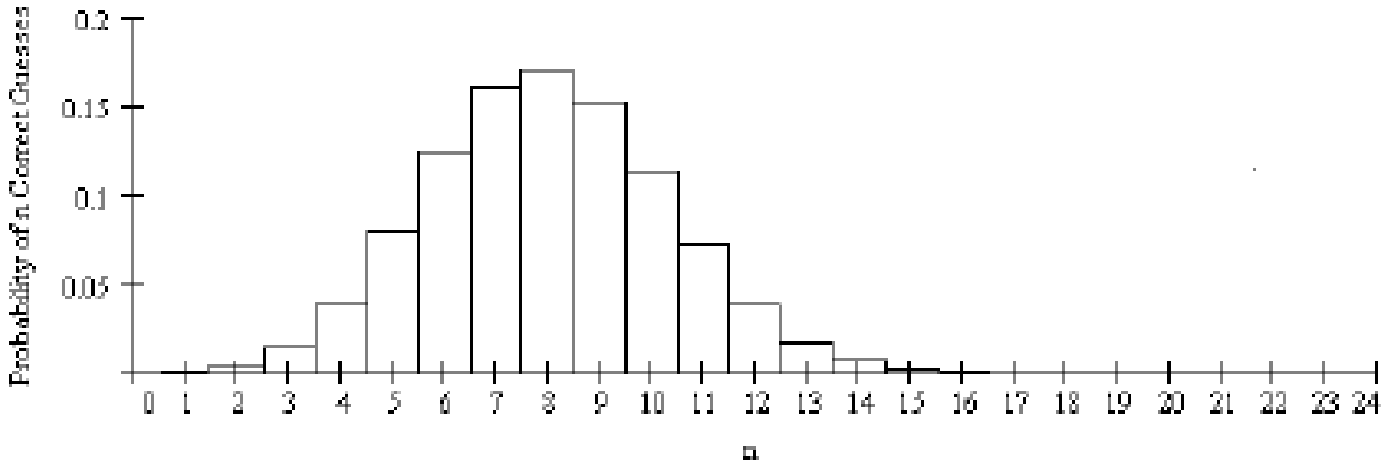


Figure 6: Distribution of Number of Correct Guesses with $p = \frac{1}{3}$

Kritikus tartomány megválasztása

$$P(\text{type I error}) = P(\text{Rejecting } H_0 | H_0 \text{ is true})$$

$$= P\left(y \geq y_c | p = \frac{1}{3}\right)$$

$$= \sum_{y=y_c}^{24} \binom{24}{y} \left(\frac{1}{3}\right)^y \left(\frac{2}{3}\right)^{24-y}$$

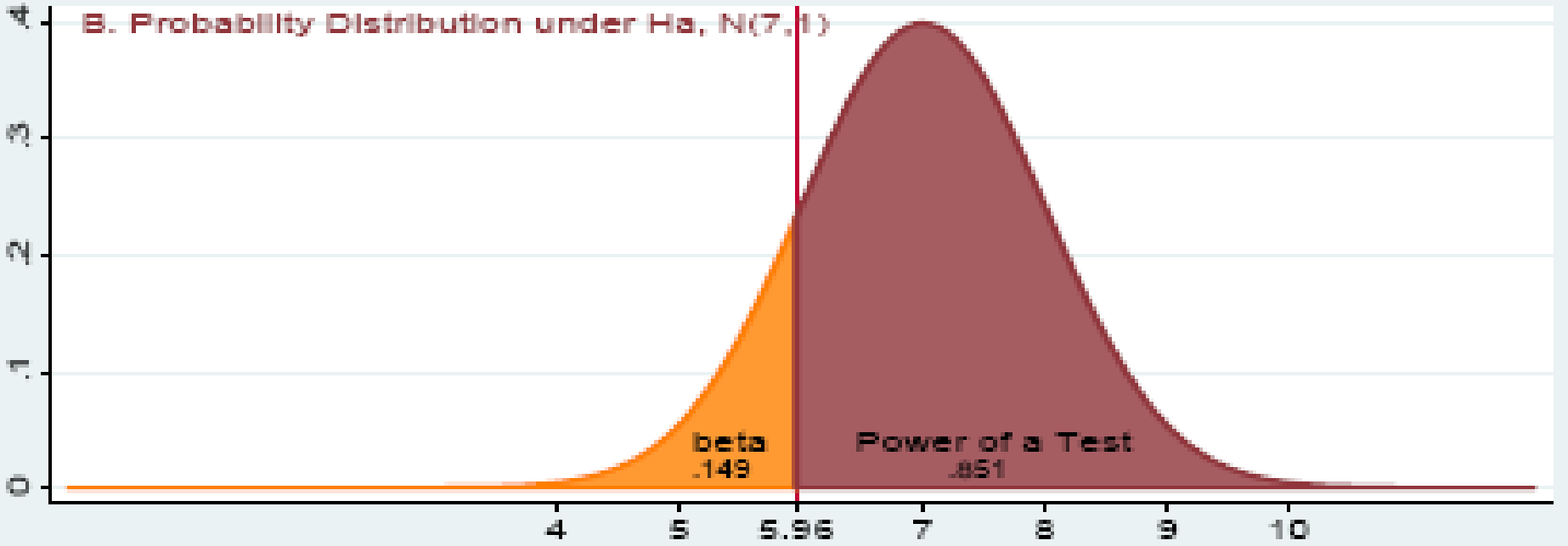
$$p\text{-value} = \sum_{y=11}^{24} \binom{24}{y} \left(\frac{1}{3}\right)^y \left(\frac{2}{3}\right)^{24-y} = 0.14$$

$$y_c = 12,\ P(\text{type I error}) = 0.0677 > 0.05$$

$$y_c = 13,\ P(\text{type I error}) = 0.0284 < 0.05$$

Másodfajú hiba valószínűsége

$$P_{\vartheta}\left(\xi \in \mathrm{X}_{e}\right), \vartheta \in \Theta_{1}$$



Példa (sörös)

- p=0.5 esetén a másodfajú hiba valószínűsége

$$= P[Y \leq 12 \mid p = 0.5]$$
$$= 0.581$$

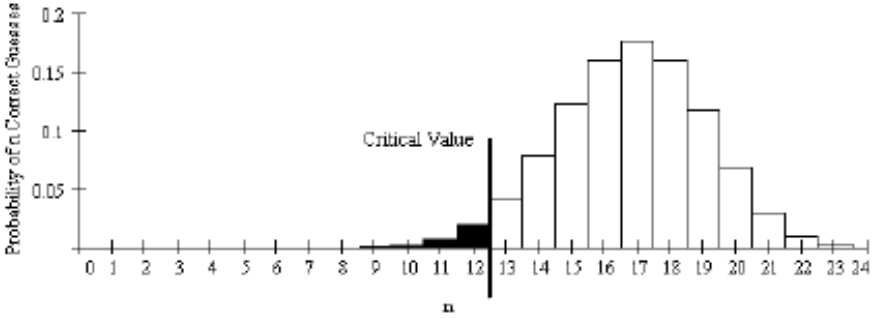


Figure 9: Distribution of Number of Correct Guesses with $p = 0.7$

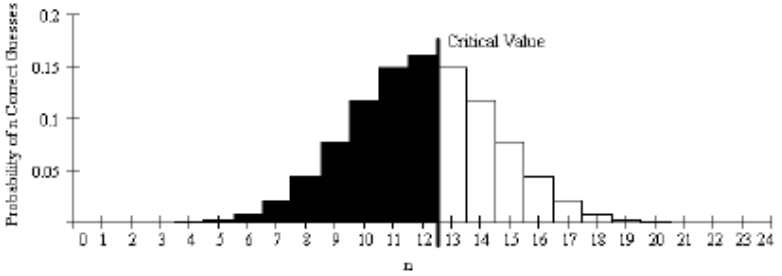
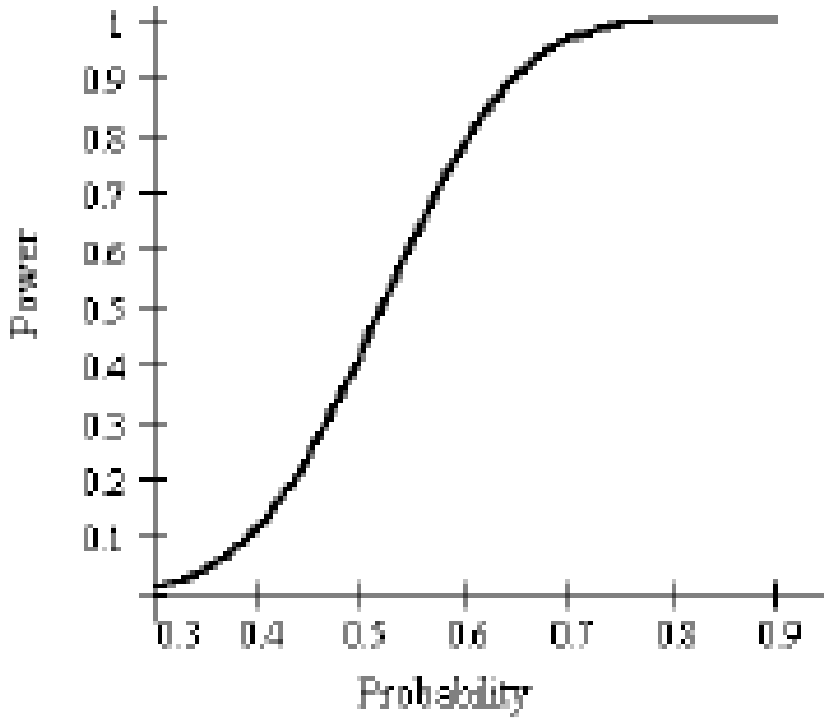


Figure 8: Distribution of Number of Correct Guesses with $p = \frac{1}{2}$

Erőfüggvény

A próba erőfüggvénye

$$\beta(\vartheta)=P_{\vartheta}\left(\xi \in \mathrm{X}_k\right)=1\text{-}P_{\vartheta}\left(\xi \in \mathrm{X}_e\right), \vartheta \in \Theta_1$$



Véletlenített próba

- Eddig adott megfigyelés esetén egyértelmű volt a döntésünk:

$$T(\mathbf{x}) = \begin{cases} 1 & , \ \mathbf{x} \in \mathbf{X}_k \\ 0 & , \ \mathbf{x} \notin \mathbf{X}_k \end{cases}$$

Véletlenített próba esetén sorsolhatunk is:

$$\Psi(\mathbf{x}) = \begin{cases} 1 & , \ \text{ha } T(\mathbf{x}) > c \\ \gamma & , \ \text{ha } T(\mathbf{x}) = c \\ 0 & , \ \text{ha } T(\mathbf{x}) < c \end{cases}$$

Elsőfajú hiba valószínűsége véletlenített próba esetén

$\vartheta \in \Theta_0$-ra az elsőfajú hiba valószínűsége:

$$P_\vartheta\left(T(\xi) > c\right) + \gamma P_\vartheta\left(T(\xi) = c\right) = E_\vartheta\left(\psi(\xi)\right)$$

α a próba terjedelme, ha minden $\vartheta \in \Theta_0$-ra

$$E_\vartheta\left(\psi(\xi)\right) \le \alpha$$

α a próba szignifikanciaszintje

(másképp: a próba pontos terjedelme),

$$\sup_{\vartheta \in \Theta_0} E_\vartheta\left(\psi(\xi)\right) = \alpha$$

Legerősebb próba egyszerű hipotézis esetében

Egyszerű H_0 és $H_1 : |\Theta_0| = |\Theta_1| = 1$.

ψ a legerősebb α-terjedelmű próba, ha:

$$P_{\vartheta_0}\left(T(\xi) > c\right) + \gamma P_{\vartheta_0}\left(T(\xi) = c\right) = E_{\vartheta_0}\left(\psi(\xi)\right) \le \alpha,$$

továbbá minden más α-terjedelmű ψ' próbára, annak másodfajú hibavalószínűsége nagyobb:

$$E_{\vartheta_1}\left(1-\psi(\xi)\right) \le E_{\vartheta_1}\left(1-\psi'(\xi)\right).$$

A legerősebb próba

- A legegyszerűbb eset: H_0 és H_1 is egyszerű (egyelemű). A valószínűséghányados (vh.) próba:

$$T(\mathbf{x}) = \begin{cases} 1 & \dfrac{L_1(\mathbf{x})}{L_0(\mathbf{x})} > c \\ \gamma & \dfrac{L_1(\mathbf{x})}{L_0(\mathbf{x})} = c \\ 0 & \dfrac{L_1(\mathbf{x})}{L_0(\mathbf{x})} < c \end{cases}$$

- Állítás (Neyman-Pearson lemma): a vh. próba legerősebb a saját terjedelmével. Minden $0<\alpha<1$-hez létezik ilyen terjedelmű vh. próba. Minden legerősebb próba ilyen alakú.

Paraméteres próbák

- Lényeg: valamilyen, véges sok valós paraméterrel leírható modellt tételezünk fel a mintáról.
- Példa:
  - Normális
  - indikátor

  eloszlású minta
- A feladat: a paraméter(ek)re vonatkozó hipotézis vizsgálata.

Próbák a normális eloszlás várható értékére: u-próba.

- H_0: $m=m_0$, H_1 $m\neq m_0$. Ha ismert a szórás (u-próba):

$$u = \sqrt{n}\,\frac{\overline{X} - m_0}{\sigma}$$

- Kritikus tartomány: $|u|>u_{\alpha/2}$. ($u_{\alpha/2}$ a standard normális eloszlás $1-\alpha/2$ kvantilise)
- Tulajdonságok:
  - torzítatlan
  - konzisztens
- Ha egyoldali az ellenhipotézis, akkor a kritikus tartomány $u>u_\alpha$ ($m>m_0$), illetve $u<-u_\alpha$ alakú ($m<m_0$). Ezek legerősebb próbák!

U-próba

$\xi_1, ..., \xi_n \sim N(m, \sigma^2)$, m ismeretlen, σ ismert.

$H_0: \; m = m_0$

$H_1: \; m \neq m_0$ (kétoldali ellenhipotézis)

$H_1': \; m < m_0$ (egyoldali ellenhipotézis)

$H_1'': \; m > m_0$ (egyoldali ellenhipotézis)

$$U = \frac{\bar{\xi} - m_0}{\sigma} \sqrt{n}$$

$$H_0 \Rightarrow U \sim N(0,1)$$

$$H_1 \Rightarrow U \sim N\left(\frac{m - m_0}{\sigma} \sqrt{n}, 1 \right)$$

U – próba (kétoldali ellenhipotézis)

$$\Phi(u_y) = y$$

$$\mathrm{X}_k = \left\{ \mathbf{x} : \left| \frac{\bar{x} - m_0}{\sigma} \sqrt{n} \right| \geq u_{1-\alpha/2} \right\} \Rightarrow$$

$$P_{m_0}\left(\xi \in \mathrm{X}_k\right) = P_{m_0}\left(\left|U\right| \geq u_{1-\alpha/2}\right) = 1 - \Phi\left(u_{1-\alpha/2}\right) + \Phi\left(-u_{1-\alpha/2}\right) =$$

$$= 1 - \left(1 - \alpha/2\right) + 1 - \left(1 - \alpha/2\right) = \alpha.$$

$$\beta(m) = P_m\left(\xi \in \mathrm{X}_k\right) = P_m\left(\left|U\right| \geq u_{1-\alpha/2}\right) =$$

$$1 - P_m\left(-u_{1-\alpha/2} < U < u_{1-\alpha/2}\right) = 1 - P_m\left(-u_{1-\alpha/2} < \frac{\bar{\xi} - m}{\sigma}\sqrt{n} + \frac{m - m_0}{\sigma}\sqrt{n} < u_{1-\alpha/2}\right) =$$

$$1 - P_m\left(-u_{1-\alpha/2} - \frac{m - m_0}{\sigma}\sqrt{n} < \frac{\bar{\xi} - m}{\sigma}\sqrt{n} < u_{1-\alpha/2} - \frac{m - m_0}{\sigma}\sqrt{n}\right) =$$

$$1 - \Phi\left(u_{1-\alpha/2} - \frac{m - m_0}{\sigma}\sqrt{n}\right) + \Phi\left(-u_{1-\alpha/2} - \frac{m - m_0}{\sigma}\sqrt{n}\right) \xrightarrow[n\to\infty]{} 1, m \neq m_0$$

Próbák a normális eloszlás várható értékére: t próba.

- H_0: $m=m_0$, H_1 $m\neq m_0$. Ha nem ismert a szórás (t-próba):

$$t = \sqrt{n}\,\frac{\overline{X} - m_0}{\hat{\sigma}}$$

- ahol

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 /(n-1)}$$

- Kritikus tartomány: $|t|>t_{1-\alpha/2,\ n-1}$. (H_0 esetén a próbastatisztika n-1 szabadságfokú, t-eloszlású.)
- Ha egyoldali az ellenhipotézis, akkor a kritikus tartomány $t>t_{1-\alpha,\ n-1}$ ($m>m_0$), illetve $t<-t_{1-\alpha,\ n-1}$ alakú ($m<m_0$). Ezek is legerősebb próbák!

Megjegyzések

- A kétoldali esetre kapott próba nem a legerősebb (ilyenkor nincs is ilyen).
- Ha a minta elemszáma nagy, a t-próba helyett az u-próba is használható (ekkor még a normális eloszlásúságra sincs szükség a centrális határeloszlás tétel miatt).

Kétoldali próbák és konfidencia intervallumok

- A normális eloszlásnál a várható értékre vonatkozó α terjedelmű próbánál láttuk, hogy a H_0: $m=m_0$ hipotézist a H_1: $m\neq m_0$ hipotézissel szemben pontosan akkkor fogadjuk el, ha m_0 benne van az 1- α megbízhatóságú konfidencia intervallumban.

Kétmintás eset: párosított megfigyelések

- Példa: Van-e különbség Budapest és Cegléd napi átlaghőmérséklete között? H_0: $m_1=m_2$ a nullhipotézis.
- Ha ugyanazon napokról van megfigyelésünk mindkét helyen: nem függetlenek a minták. Ekkor a párok tagjai közötti különbséget vizsgálva, az előző egymintás esetre vezethető vissza a feladat. H_0^*: $m=0$, H_1^*:$m\neq 0$ az új hipotézisek.

Kétmintás eset: független minták

Ha ismert a szórás: (X n elemű, σ_1 szórású,
Y m elemű, σ_2 szórású), alkalmazható a kétmintás u-próba

$$u = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2 / n + \sigma_2^2 / m}}$$

Kritikus tartomány: mint az egymintás esetben
Ha ismeretlenek, de azonosak a szórások:

$$t_{n+m-2} = \sqrt{\frac{nm(n+m-2)}{n+m}} \frac{\bar{X} - \bar{Y}}{\sqrt{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2}}$$

A szórás vizsgálata kétmintás esetben: F-próba

- H_0: $\sigma_1=\sigma_2$
- Két független, *n*, illetve *m* elemű normális eloszlású minta alapján a próbastatisztika:

  (a korrigált tapasztalati szórásnégyzetek hányadosa)

$$F = \max\left(\frac{s_1^2}{s_2^2}, \frac{s_2^2}{s_1^2}\right)$$

- Kritikus érték: az *n*-1,*m*-1 szabadságfokú F eloszlás 1-α/2 kvantilise (*n* a számlálóbeli, *m* pedig a nevezőbeli minta elemszáma).

Kétmintás t-próba ismét

- Alkalmazható, ha az F-próba elfogadja a szórások azonosságát.
- Ha nem, akkor Welch-próba:

$$t' = \frac{\overline{X} - \overline{Y}}{\sqrt{s_1^{\,2}/n_1 + s_2^{\,2}/n_2}}$$

- H_0 esetén közelítőleg t eloszlású f szabadságfokkal, ahol

$$\frac{1}{f} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1} \qquad c = \frac{s_1^{\,2}/n}{s_1^{\,2}/n + s_2^{\,2}/m}$$

χ-négyzet próba

- H_0 hipotézis: az A_1, A_2 , ..., A_r teljes eseményrendszerre teljesül $P(A_1)=p_1$, $P(A_2)=p_2$, ..., $P(A_r)=p_r$
- A tesztstatisztika: $$\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$$

  ami aszimptotikusan r-1 szabadságfokú χ-négyzet eloszlású, ha igaz a nullhipotézis.
- Kritikus tartomány: ha a statisztika értéke nagyobb, mint az r-1 szabadságfokú χ-négyzet eloszlás 1- α kvantilise, elutasítjuk a nullhipotézist.

χ-négyzet próba (folytatás)

- Miért is ez a határeloszlás?

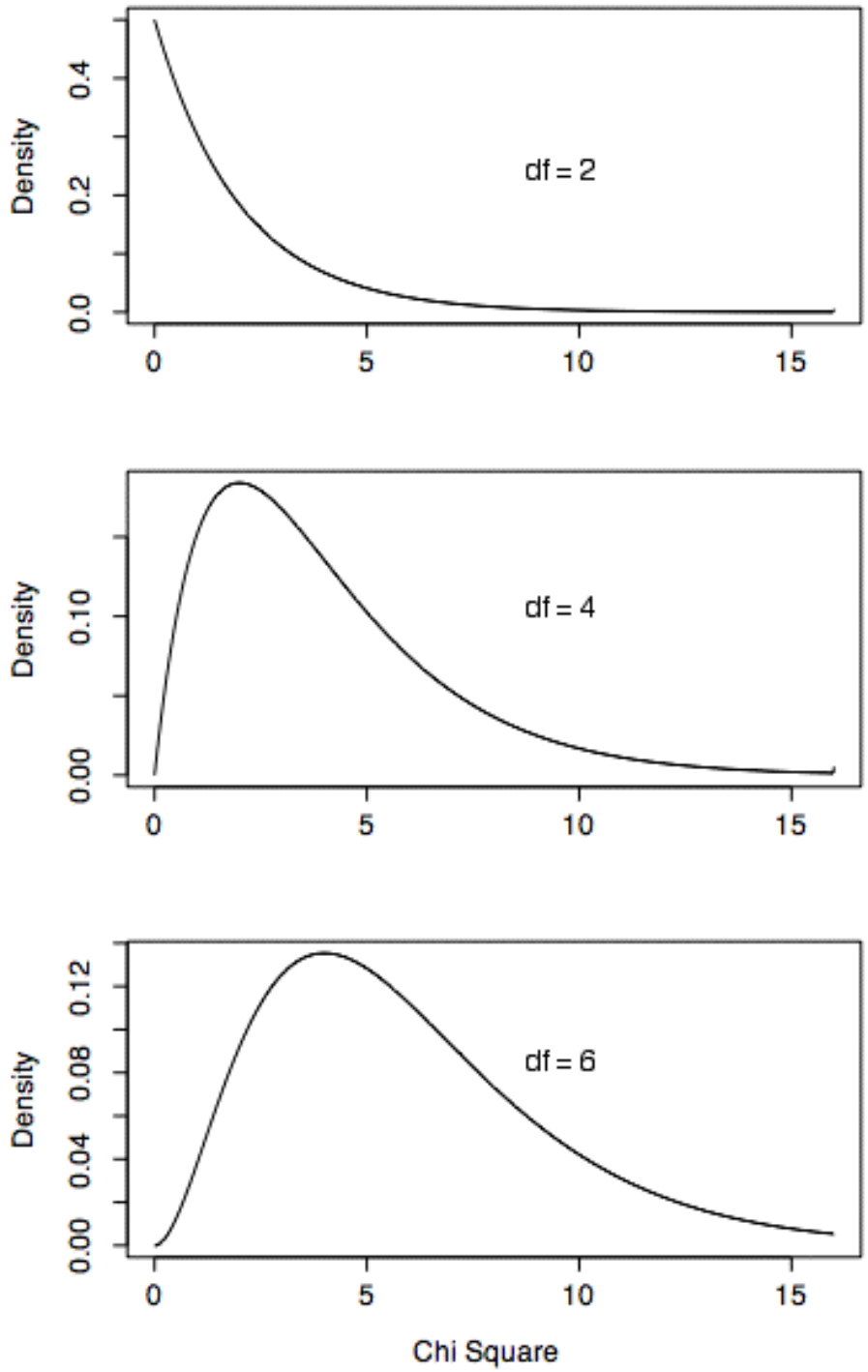
$r = 2,\ H_0 : P(A) = p,\ \nu : A$ gyakorisága n kísérletből

$$\chi^2 = \frac{(\nu - np)^2}{np} + \frac{((n-\nu) - n(1-p))^2}{n(1-p)} = \frac{(\nu - np)^2}{np} + \frac{(\nu - np)^2}{n(1-p)} = \frac{(\nu - np)^2}{np(1-p)}$$

$\xi_i = 1$, ha az $i.$ kísérletnél A bekövetkezik, 0 különben

$$\nu = \sum_{i=1}^{n} \xi_i,\ E\xi_i = p,\ D^2\xi_i = p(1-p),$$

$$\chi^2 = \left(\frac{\sum_{i=1}^{n} \xi_i - nE\xi_1}{\sqrt{n} D\xi_1} \right)^2 \xrightarrow[n \to \infty, \text{eloszlásban}]{} \chi_1^2$$


Density
0.0
0.2
0.4
df = 2
0
5
10
15
Density
0.00
0.10
df = 4
0
5
10
15
Density
0.00
0.04
0.08
0.12
df = 6
0
5
10
15
Chi Square

Példa (kockadobás)

- 36 kockadobás eredménye

Szám	Megfigyelt	np_i	$\frac{(\nu_i - np_i)^2}{np_i}$
1	8	6	0.667
2	5	6	0.167
3	9	6	1.500
4	2	6	2.667
5	7	6	0.167
6	5	6	0.167

$$n = 36,\ r = 6$$

$$\sum_{i=1}^{6} \frac{\left(\nu_i - np_i\right)^2}{np_i} \sim \chi_5^2$$

$$\sum_{i=1}^{6} \frac{\left(\nu_i - np_i\right)^2}{np_i} = 5.333$$

$$P\left(\chi_5^2 > 5.333\right) = 0.377 \Rightarrow$$

Nem tudjuk a szabályosság hipotézisét elutasítani!

Példa (számítógépek népszerűsége)

- 100 amerikai diák

Számí-tógép	Megfigyelt	np_i	$\frac{(\nu_i - np_i)^2}{np_i}$
IBM	47	33.333	5.604
Macintosh	36	33.333	0.213
Egyéb	17	33.333	8.003

$$n = 100,\ r = 3$$

$$\sum_{i=1}^{3} \frac{\left(\nu_i - np_i\right)^2}{np_i} \sim \chi_2^2$$

$$\sum_{i=1}^{3} \frac{\left(\nu_i - np_i\right)^2}{np_i} = 13.820$$

$$P\left(\chi_2^2 > 5.99\right) = 0.05 \Rightarrow$$

Elutasítjuk az egyforma kedveltség hipotézisét!

χ-négyzet próba illeszkedésvizsgálatra

- Illeszkedésvizsgálat:

$H_0 : \xi_1, ..., \xi_n \ F$ eloszlásfüggvényűek

- Visszavezetjük az előző esetre

$$A_i = \{\xi \in C_i\}, i = 1, 2, ..., r, \bigcup_i C_i = \mathbf{R}$$

Diszkrét esetben gyakran: $A_i = \{\xi = x_i\}, i = 1, 2, ..., r$

Példa

- Mi lehet egy vezető által okozott károk számának eloszlása?
- Poisson eloszlású-e?

Kár-szám	0	1	2	3	4	5	6	7	>7	Össze-sen
Veze-tők száma	129524	16267	1966	211	31	5	1	1	0	148006

Becsléses χ-négyzet próba

- H_0 hipotézis: az A_1, A_2 , ..., A_r teljes eseményrendszerre teljesül:

$$P(A_i) = p_i(\vartheta_1, ..., \vartheta_s), i = 1, 2, ..., r$$

$\vartheta_1, ..., \vartheta_s$ ismeretlen paraméterek.

A tesztstatisztika:

$$\chi^2 = \sum_{i=1}^{r} \frac{\left(\nu_i - n\hat{p}_i\right)^2}{n\hat{p}_i} \xrightarrow[n\to\infty]{} \chi^2_{r-s-1},$$

ahol

$$\hat{p}_i = p_i(\hat{\vartheta}_1, ..., \hat{\vartheta}_s).$$

Példa (folyt.)

Kár-szám	0	1	2	3	4	5	6	7	>7	Össze-sen
Veze-tők száma	129524	16267	1966	211	31	5	1	1	0	148006
np_i *Poisson*	128 433	18 218	1 292	61	2,2	0,06	0,001	3E-05	5E-07	
Np_i *Neg. bin.*	129 541	16 237	1 962	234	28	3,3	0,39	0,05	0,006	

$n = 148006,\ r = 5$

$A_i = \{\xi = i\}, i = 0,1,2,3$

$A_4 = \{\xi \geq 4\}$

Poisson eset:

$\hat{\lambda}$=0.709

$$\sum_{i=0}^{4} \frac{\left(\nu_i - n\hat{p}_i\right)^2}{n\hat{p}_i} \sim \chi^2_{5-1-1}$$

$$\sum_{i=0}^{4} \frac{\left(\nu_i - n\hat{p}_i\right)^2}{n\hat{p}_i} > 200$$

$$P\left(\chi_3^2 > 17.7\right) = 0.05\% \Rightarrow$$

Elutasítjuk Poisson eloszlás hipotézisét!

Az illeszkedésvizsgálat alkalmazása folytonos eloszlásokra

- A teljes eseményrendszer a számegyenes felosztása révén jön létre.
- Ügyeljünk arra, hogy minden intervallum közel azonos valószínűségű legyen.
- Ha paraméterbecslés szükséges, ML módszer alkalmazható.

χ-négyzet próba homogenitásvizsgálatra

- Homogenitásvizsgálat:

$H_0 : \xi_1, ..., \xi_n$ és $\eta_1, ..., \eta_m$ ugyanolyan eloszlásúak

- Hasonlóan járunk el, mint korábban

$$\bigcup_{i=1}^{r} C_i = \mathbf{R}$$

$$\nu_i = \left| \left\{ j : \xi_j \in C_i \right\} \right|, \mu_i = \left| \left\{ j : \eta_j \in C_i \right\} \right|, \; i = 1, 2, ..., r,$$

A tesztstatisztika:

$$\chi^2 = nm \sum_{i=1}^{r} \frac{\left(\nu_i / n - \mu_i / m \right)^2}{\nu_i + \mu_i} \xrightarrow[n,m \to \infty]{} \chi^2_{r-1}$$

Ki tanul jobban?

2009. január 5-ei vizsga

Jegy	Férfi	Nő	Összesen
1	47	4	51
2	11	1	12
3	11	2	13
4	9	2	11
5	8	2	10
Összesen	86	11	97
Átlag	2,1	2,7	2,1

$$C_1 = \{1;2\}, C_1 = \{3;4;5\}$$

$$\nu_i = \left|\left\{ j : \xi_j \in C_i \right\}\right|, \mu_i = \left|\left\{ j : \eta_j \in C_i \right\}\right|, \; i = 1,2,$$

$$\nu_1 = 58, \nu_2 = 28, \mu_1 = 5, \mu_2 = 6, n = 86, m = 11$$

A tesztstatisztika:

$$\chi^2 = 86 \cdot 11 \left(\frac{\left(\frac{58}{86} - \frac{5}{11}\right)^2}{58+5} + \frac{\left(\frac{28}{86} - \frac{6}{11}\right)^2}{28+6} \right) = 2.071$$

$$P(\chi_1^2 > 2.71) = 10\% \Rightarrow$$

Nem tudjuk elutasítani az egyforma képesség hipotézisét!

χ-négyzet próba függetlenségvizsgálatra

- H_0 hipotézis: az A_1, A_2 , ..., A_r és B_1, B_2 , ..., B_s teljes eseményrendszerekre teljesül a függetlenség.

$$\sum_{i,j} \frac{\left(\nu_{ij} - np_i q_j\right)^2}{np_i q_j}$$

- Kritikus tartomány: ha a statisztika értéke nagyobb, mint az rs-1 szabadságfokú χ-négyzet eloszlás 1- α kvantilise, elutasítjuk a nullhipotézist.

Becsléses eset

- Általában, ha az illesztendő eloszlást nem ismerjük – csak a családját - becsüljük a paramétereit. Ekkor a próbastatisztika szabadságfoka annyival csökken, ahány paramétert becsültünk.
- Függetlenségvizsgálatnál általában nem ismerjük a teljes eseményrendszer tagjainak valószínűségét, így r-1+s-1 valószínűséget kell becsülnünk. A szabadságfok ekkor tehát

  rs-1-r-s+2=(r-1)(s-1).

ν_{ij} : $A_i B_j$ gyakorisága

$\nu_{i\bullet}$: A_i gyakorisága

$\nu_{\bullet j}$: B_j gyakorisága

A tesztstatisztika

$$n\sum_{i,j}\frac{\left(\nu_{ij}-\frac{\nu_{i\bullet}\nu_{\bullet j}}{n}\right)^2}{\nu_{i\bullet}\nu_{\bullet j}}\xrightarrow[n\to\infty]{}\chi^2_{(r-1)(s-1)}$$

$r=s=1$ esetben

$$n\frac{\left(\nu_{11}\nu_{22}-\nu_{12}\nu_{21}\right)^2}{\nu_{1\bullet}\nu_{2\bullet}\nu_{\bullet 1}\nu_{\bullet 2}}\xrightarrow[n\to\infty]{}\chi^2_1$$