Valószínűségszámítás és Statisztika

8. előadás

2020. április 7.

Likelihood függvény

Def.: A $\xi_1,...,\xi_n$ független, azonos eloszlású minta likelihood függvénye

$$L(\mathbf{x},\theta)=\begin{cases} P_\theta(\boldsymbol{\xi}=\mathbf{x})=\prod_{i=1}^{n} P_\theta(\xi_i=x_i) & \text{diszkrét minta esetén} \\ f_\theta(\mathbf{x})=\prod_{i=1}^{n} f_\theta(x_i) & \text{abszolút folytonos minta esetén} \end{cases}$$

ahol f_θ ξ_i sűrűségfüggvénye.

$l(\mathbf{x},\theta)=\ln L(\mathbf{x},\theta)$ a loglikelihood függvény.

Maximum likelihood becslés

- Definíció heurisztikusan: azt a paraméterértéket keressük, amelyre az adott minta bekövetkezési valószínűsége maximális.

$$\text{Def.: } \theta \text{ maximum likelihood becslése } \hat{\theta} = T(\boldsymbol{\xi}) \in \Theta, \text{ha}$$

$$L(\boldsymbol{\xi}, \hat{\theta}) = \max_{\theta \in \Theta} L(\boldsymbol{\xi}, \theta)$$

Likelihood egyenlet

Gyakran a loglikelihood függvény maximumhelyét keresik a

$$\frac{\partial l(\mathbf{x},\theta)}{\partial\theta} = 0$$ egyenletet (vagy egyenletrendszert) megoldva.

Ez diszkrét minta esetén a

$$\sum_{i=1}^{n} \frac{\partial \ln P_\theta(\xi_i = x_i)}{\partial\theta} = 0$$

egyenletet (vagy egyenletrendszert) jelenti.

Abszolút folytonos minta esetén

$$\sum_{i=1}^{n} \frac{\partial \ln f_\theta(x_i)}{\partial\theta} = 0$$

egyenletet (vagy egyenletrendszert) oldjuk meg.

Koronavírusos halálesetek száma

nap	Ausztria	Csehország	Magyarország	Szlovákia
01.ápr	20	7	1	0
02.ápr	18	8	4	0
03.ápr	12	5	1	0
04.ápr	10	9	11	0
05.ápr	18	6	2	0
06.ápr	18	8	4	0
átlag	16,0	7,2	3,8	0,0

Példa. (Poisson)

Tegyük fel, hogy $\eta_1, \eta_2, \ldots, \eta_n \sim Poisson(\lambda)$. Ekkor

$$\mathrm{L}(\underline{k}, \lambda) = \mathrm{P}_\lambda(\eta_1 = k_1, \ldots, \eta_n = k_n) = \prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = \left(\prod_{i=1}^{n} \frac{1}{k_i!}\right) \lambda^{\sum k_i} e^{-n\lambda}$$

$$\ell(\underline{k}, \lambda) = \ln \mathrm{L}(\underline{k}, \lambda) = \left(\sum_{i=1}^{n} \ln\left(\frac{1}{k_i!}\right)\right) + \left(\sum_{i=1}^{n} k_i\right) \ln \lambda - n\lambda$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{\sum k_i}{\lambda} - n = 0 \iff \lambda = \frac{\sum k_i}{n}$$

így az ML becslés

$$\widehat{\lambda} = \frac{\sum \eta_i}{n}$$

Adottak sorszámozott gömbök (lottóhúzás) 1-től *N*-ig.
Visszatevéses húzás esetén becsüljük meg *N*-t!

$$\mathrm{P}_N(\xi_i = k) = \frac{1}{N}\chi\{k \le N\}$$

$$\mathrm{L}(\underline{k}, \lambda) = \mathrm{P}_N(\xi_1 = k_1, \ldots, \xi_n = k_n) = \frac{1}{N^n}\chi\{\max_i k_i \le N\}$$

$$\widehat{N} = \max_i \xi_i$$

Normális eloszlást feltételezve a paraméterek ML becslése:

$$\hat{m} = 405 \text{ és } \widehat{\sigma^2} = 149\ 284$$

Momentum módszer

- Ha az eloszlást k db paraméter határozza meg, akkor k db egyenletből kaphatunk rájuk becslést. Az egyenletek a tapasztalati és az elméleti momentumok egybevetéséből adódnak:

$$m_i(\underline{\theta}) = E_{\underline{\theta}}(X^i)$$

$$m_i(\underline{\theta}) = \frac{\sum_{j=1}^{n} \left(\xi_j\right)^i}{n}$$

Konfidenciaintervallum

- Olyan intervallum, mely legalább 1-α valószínűséggel tartalmazza a keresett paramétert:

$$P_\theta\left(T_1(X) < \theta < T_2(X)\right) \geq 1-\alpha$$

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)

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

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

α	$\boxed{u_{1-\alpha/2}}$	$\sqrt{\frac{1}{\alpha}}$
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 $0{,}5132 \le p \le 0{,}5150$

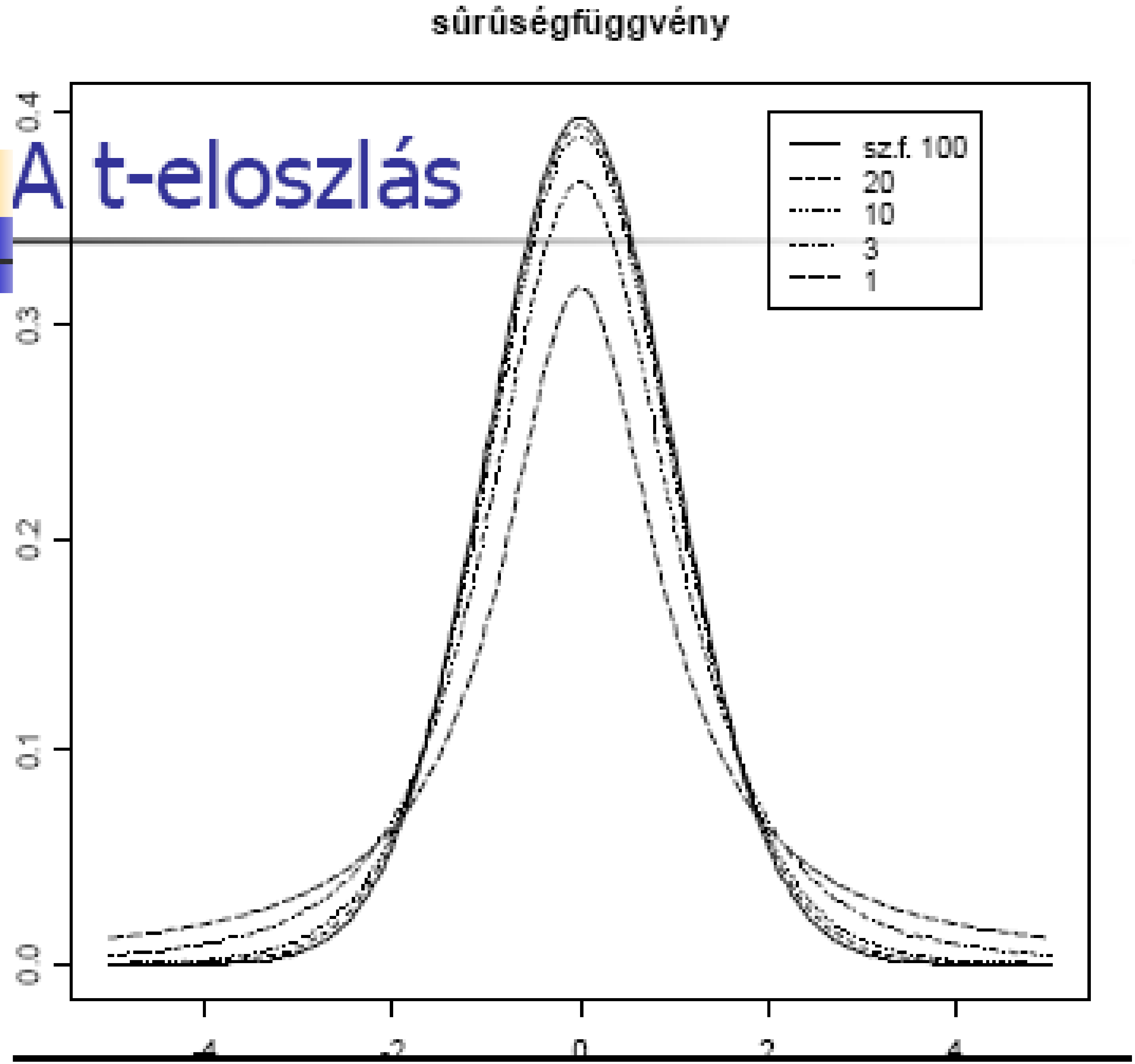
Konfidencia intervallum normális eloszlás várható értékére (ismeretlen szórás esetén)

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

$$X_0, X_1, \ldots, X_n \ \boldsymbol{f\ddot{u}ggetlen} \, N(0,1)$$

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

A t-eloszlás



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

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

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

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

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

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

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

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

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

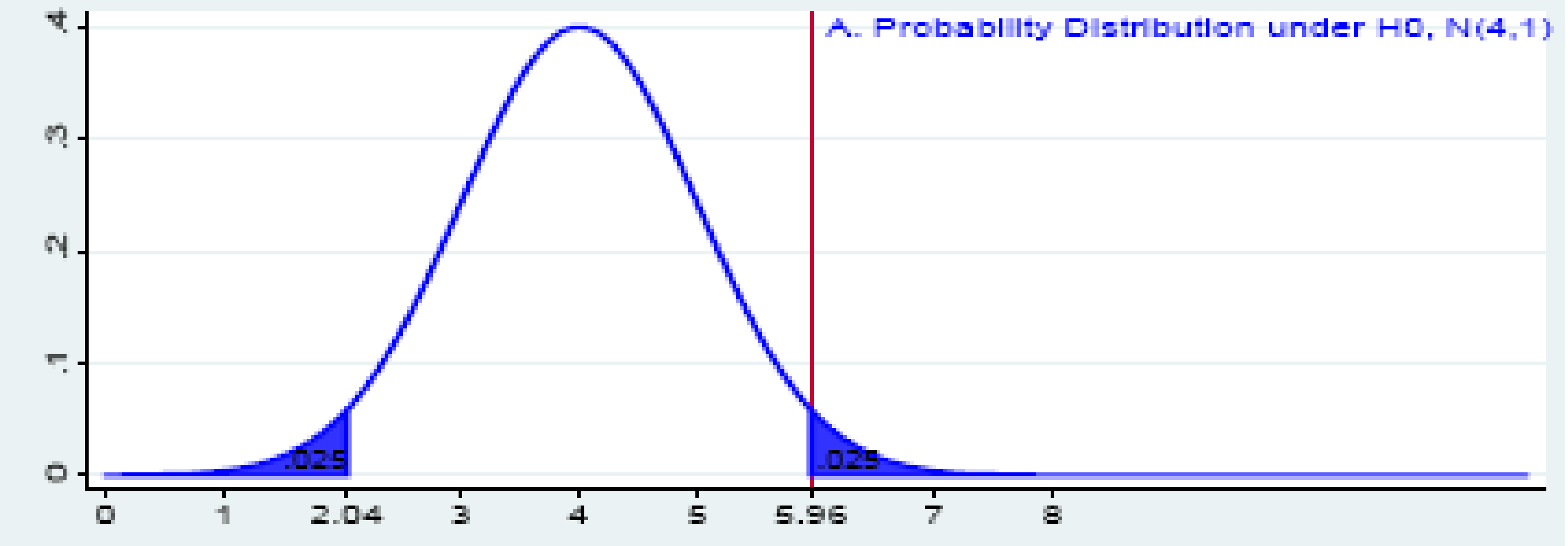
α a próba szignifikanciaszintje

(másképp: a próba pontos terjedelme),

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

Példa (egyetlen megfigyelés)

H_0: a megfigyelés N(4,1) eloszlású



Példa (sörök megkülönböztetése)

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

$$H_0: p = \frac{1}{3},\ H_1: p > \frac{1}{3}$$

Az eloszlás H_0 esetén

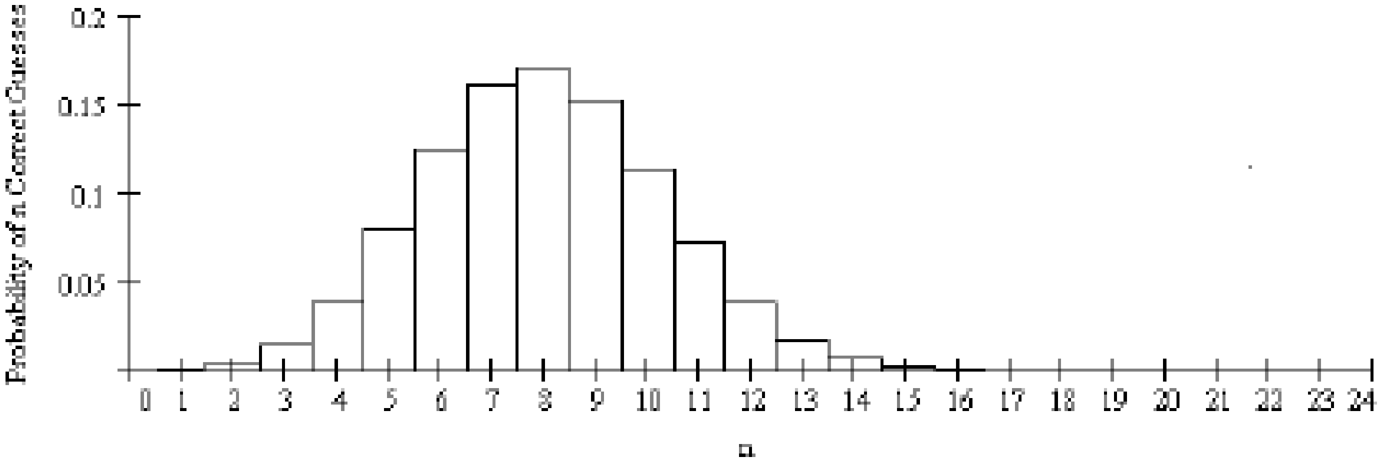


Figure 6: Distribution of Number of Correct Guesses with $p = \frac{1}{3}$

Kritikus tartomány megválasztása

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

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

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

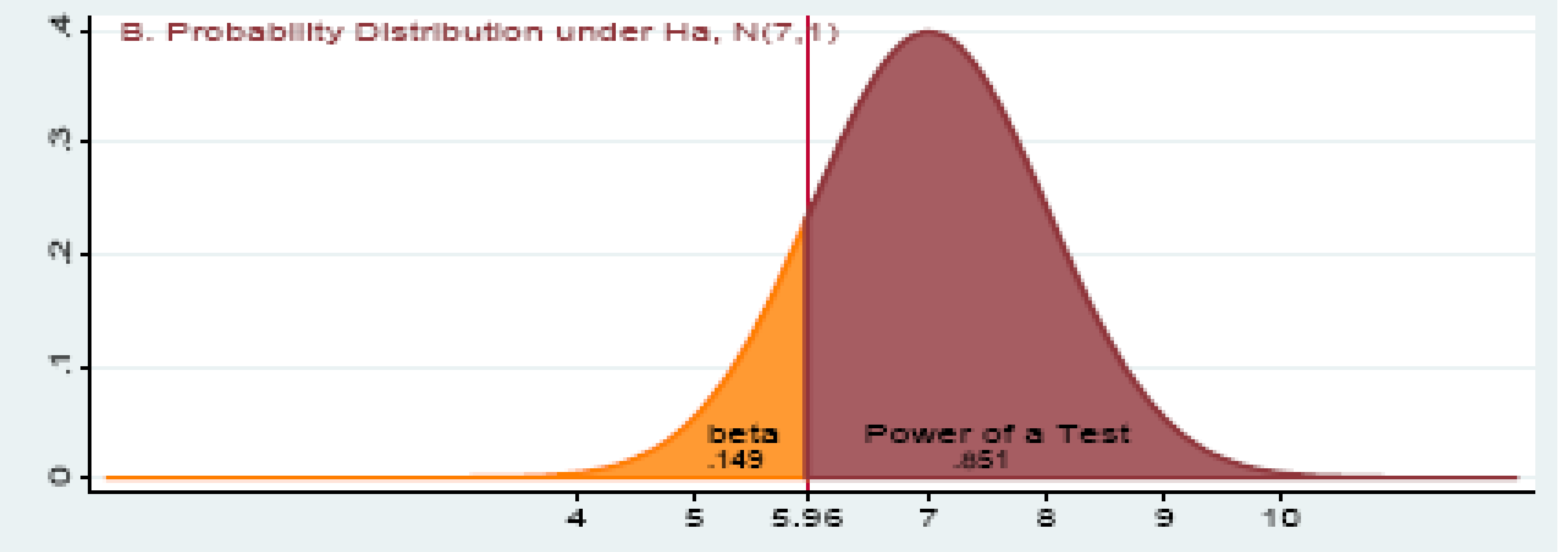
$$p\text{-value} = \sum_{y=11}^{24} \binom{24}{y} \left(\frac{1}{3}\right)^y \left(\frac{2}{3}\right)^{24-y} = 0.14$$

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

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

Másodfajú hiba valószínűsége

$$P_{\vartheta}\left(\boldsymbol{\xi} \in \mathrm{X}_e\right), \vartheta \in \Theta_1$$



Példa (sörös)

- p=0.5 esetén a másodfajú hiba valószínűsége

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

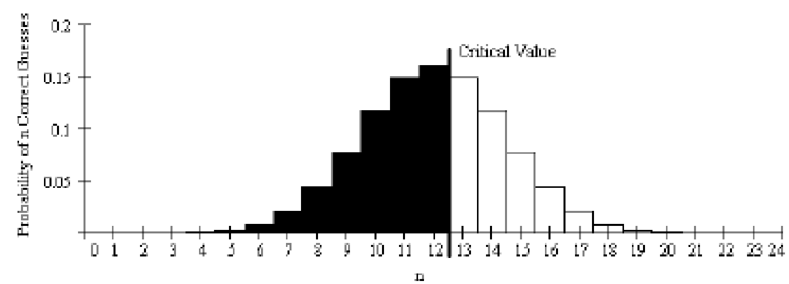


Figure 8: Distribution of Number of Correct Guesses with $p = \frac{1}{2}$

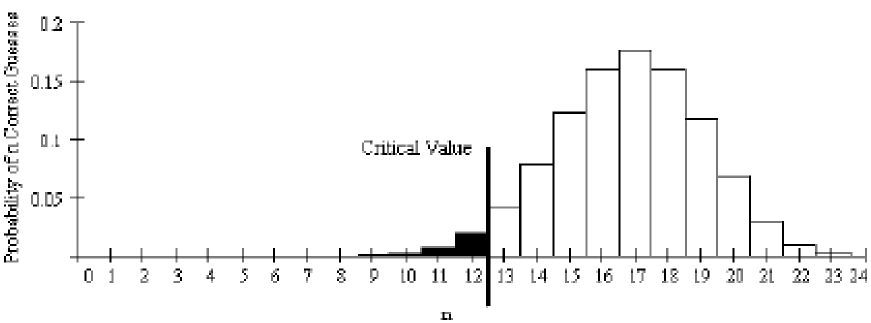
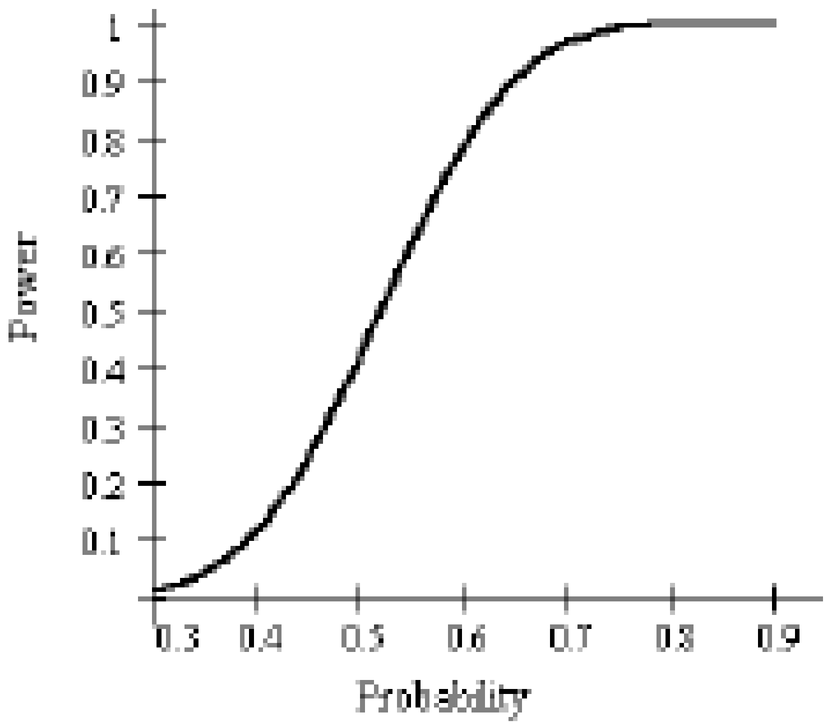


Figure 9: Distribution of Number of Correct Guesses with $p = 0.7$

Erőfüggvény

A próba erőfüggvénye

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

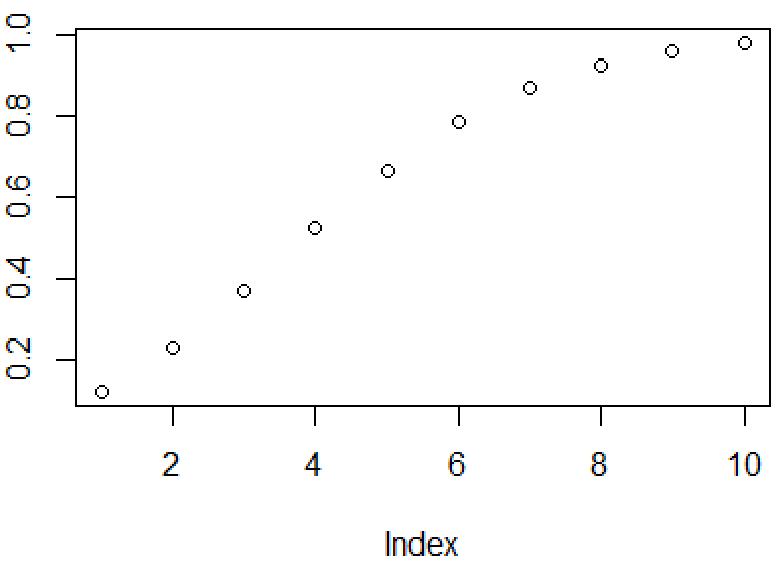


10 ezer ember koronavírusos tesztelése

- $H_0 : p = 0{,}2\%$
- $H_1 : p > 0{,}2\%$
- Teljesen hasonló az előző példához
- $\geq c$ fertőzött esetén elutasítjuk a nullhipotézist

c	**Elsőfajó hiba valószínűsége**
26	0.11196181
27	0.07768084
28	0.05230280
29	0.03418834
30	0.02170562

Próba ereje c=28 esetén



p	erő
2*2‰	0.1222548
3*2‰	0.2320277
4*2‰	0.3728980
5*2‰	0.5252410
6*2‰	0.6674264

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:

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

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

$$E_\vartheta\left(\psi(\boldsymbol{\xi})\right) \leq \alpha$$

α a próba szignifikanciaszintje

(másképp: a próba pontos terjedelme),

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

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):

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

- Kritikus tartomány: $|u| > u_{1-\alpha/2}$. ($u_{1-\alpha/2}$ a standard normális eloszlás $1 - \alpha/2$ kvantilise)
- Ha egyoldali az ellenhipotézis, akkor a kritikus tartomány $u > u_{1-\alpha}(m > m_0)$, illetve

$u < -u_{1-\alpha}$alakú $(m < m_0)$. Ezek legerősebb próbák!

U-próba

$\xi_1, \ldots, \xi_n \sim N(m, \sigma^2)$, 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)

$$\Phi(u_y)=y$$

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

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

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

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

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

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

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