

# Valószínűségszámítás és Statisztika

10. előadás  
2023. május 4.

# Likelihood függvény

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

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

ahol  $f_\theta$   $\xi_i$  sűrűségfüggvénye.

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

# Konfidenciaintervallum

- Olyan intervallum, mely legalább  $1-\alpha$  valószínűséggel tartalmazza a keresett paramétert:

$$P_{\theta}(T_1(X) < \theta < T_2(X)) \geq 1 - \alpha$$

# 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)

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

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

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

$\alpha$	$u_{1-\alpha/2}$	$\sqrt{\frac{1}{\alpha}}$
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

$\xi_1, \dots, \xi_n, D^2 \xi_i = \sigma^2$  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.

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

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

Esetünkben 0,9973 valószínűséggel  $0,5132 \leq p \leq 0,5150$



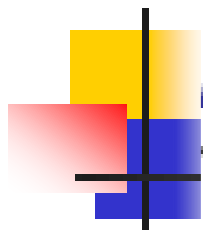
# Konfidencia intervallum normális eloszlás várható értékére (ismeretlen szórás esetén)

- 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:

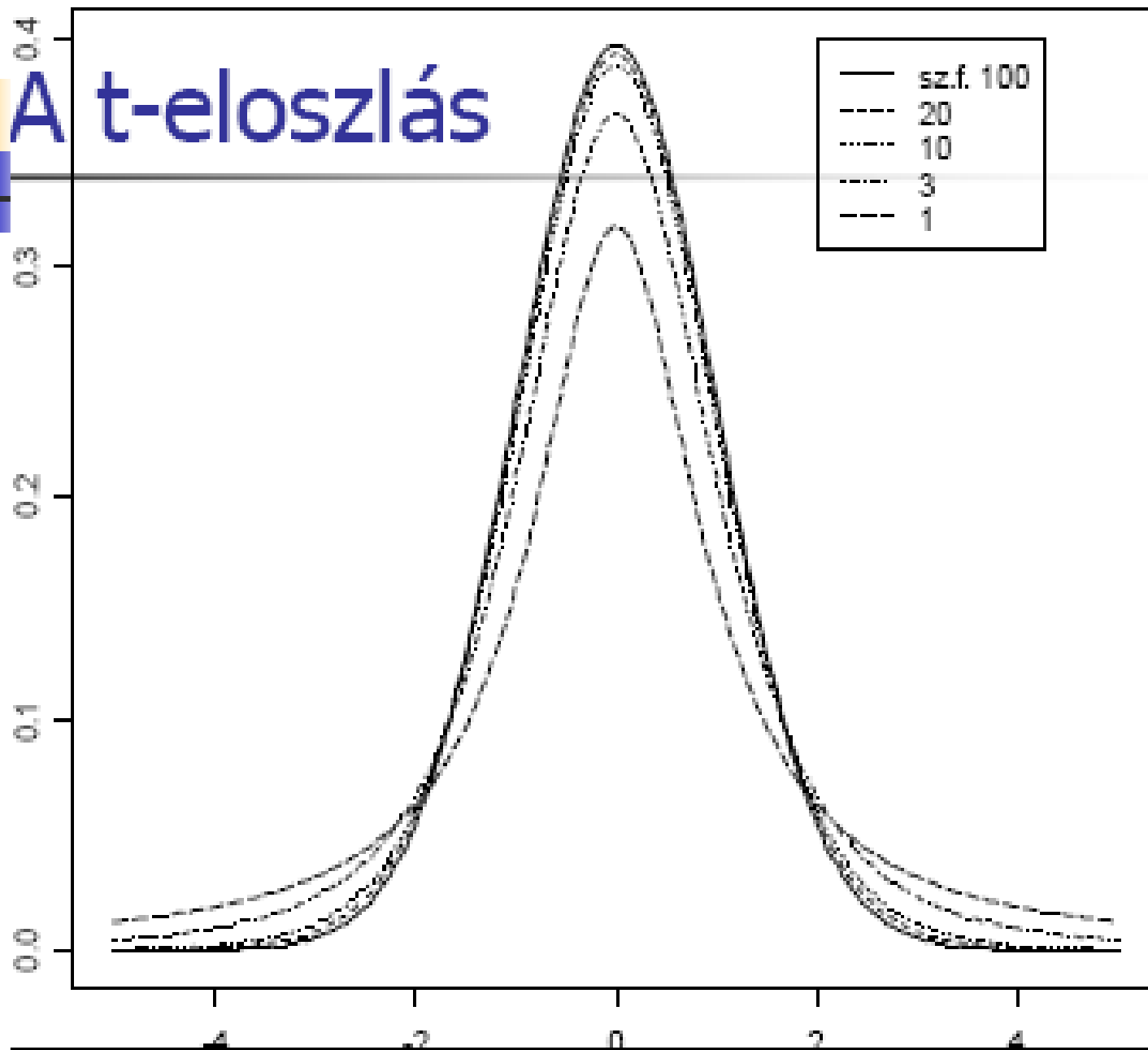
$X_0, X_1, \dots, X_n$  *független*  $N(0,1)$

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

# sűrűségfüggvény



## A t-eloszlás



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

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

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

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

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

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

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

# 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

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

$n$	$t_{n-1,1-5\%}$
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

$\alpha$  a próba terjedelme, ha minden  $\mathcal{G} \in \Theta_0$ -ra

$$P_{\mathcal{G}}(\xi \in X_k) \leq \alpha$$

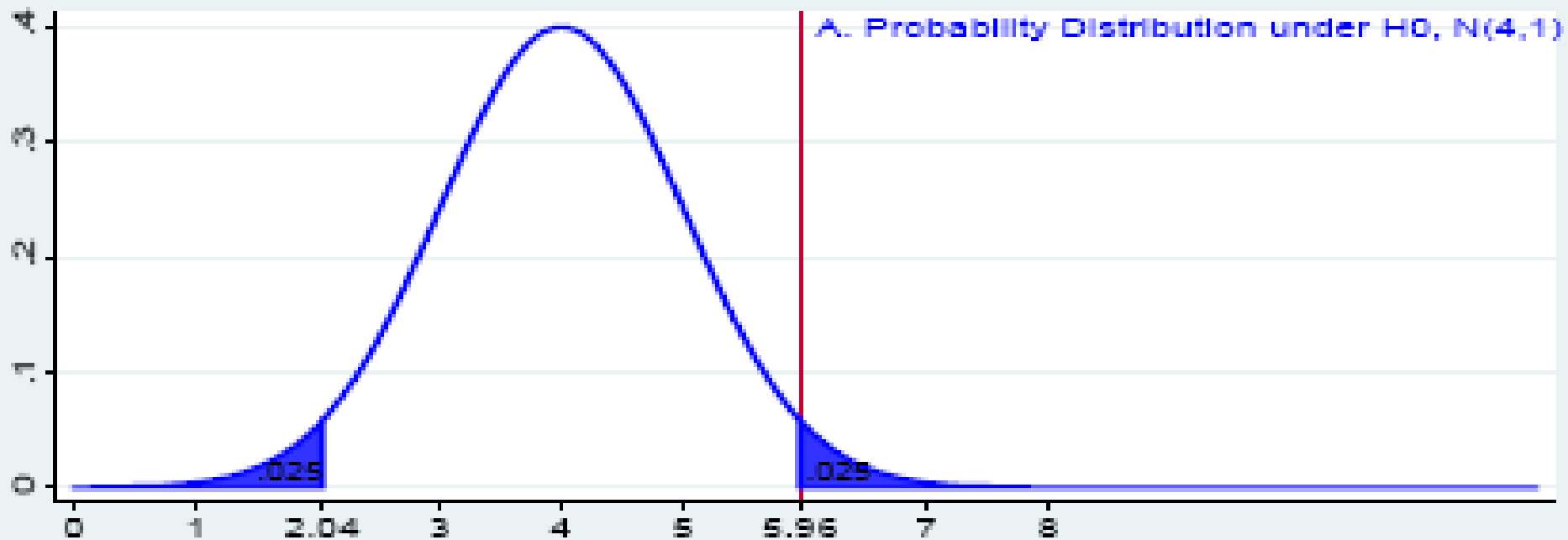
$\alpha$  a próba szignifikanciaszintje

(másképp: a próba pontos terjedelme),

$$\sup_{\mathcal{G} \in \Theta_0} P_{\mathcal{G}}(\xi \in X_k) = \alpha$$

# Példa (egyetlen megfigyelés)

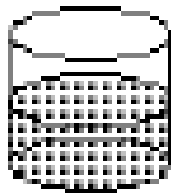
$H_0$ : a megfigyelés  $N(4,1)$  eloszlású



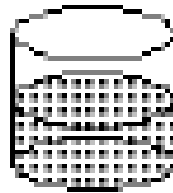
# Példa (sörök megkülönböztetése)

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

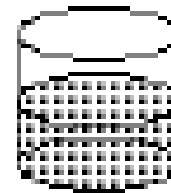
$$H_0 : p = \frac{1}{3}, \quad H_1 : p > \frac{1}{3}$$



**Lowe nbra u**



**Miller**



**Miller**

# Az eloszlás $H_0$ esetén

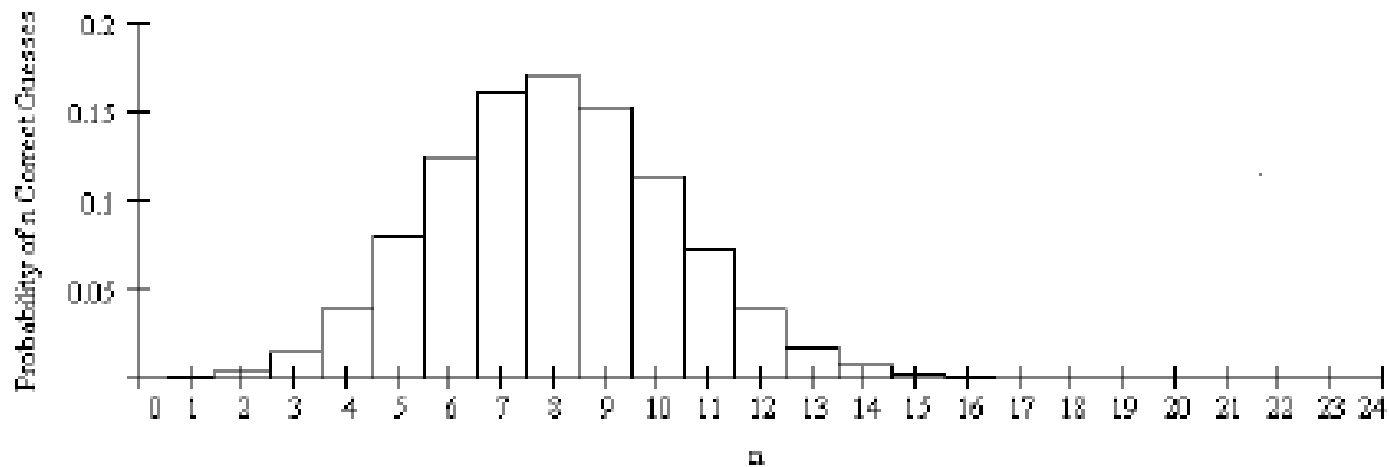


Figure 6: Distribution of Number of Correct Guesses with  $p = \frac{1}{3}$

# Kritikus tartomány megválasztása



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

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

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

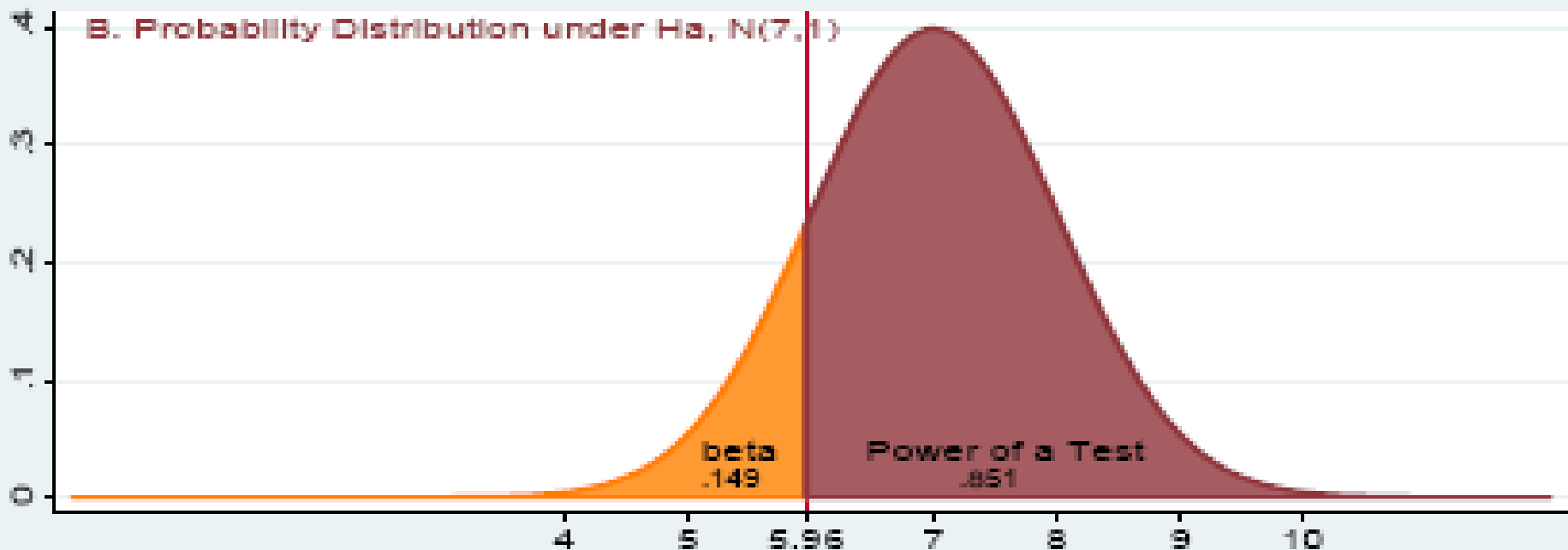
$$p\text{-value} = \sum_{y=11}^{24} \binom{24}{y} \left(\frac{1}{3}\right)^y \left(\frac{2}{3}\right)^{24-y} = 0.14$$

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

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

# Másodfajú hiba valószínűsége

$$P_{\mathcal{G}}(\xi \in X_e), \mathcal{G} \in \Theta_1$$



# Példa (sörös)

- $p=0.5$  esetén a másodfajú hiba valószínűsége

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

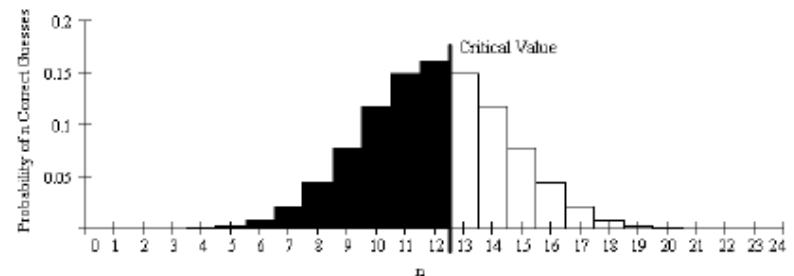
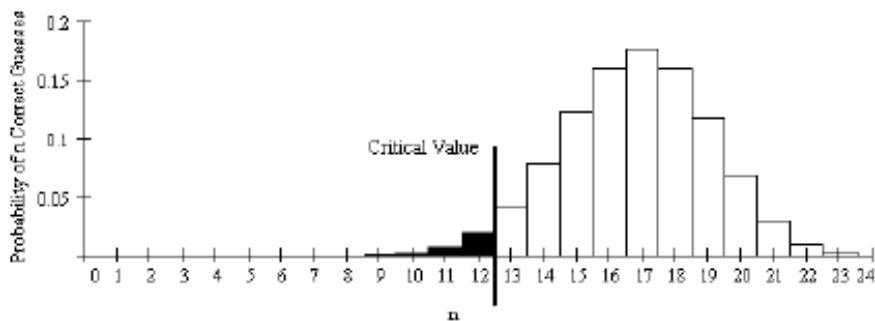
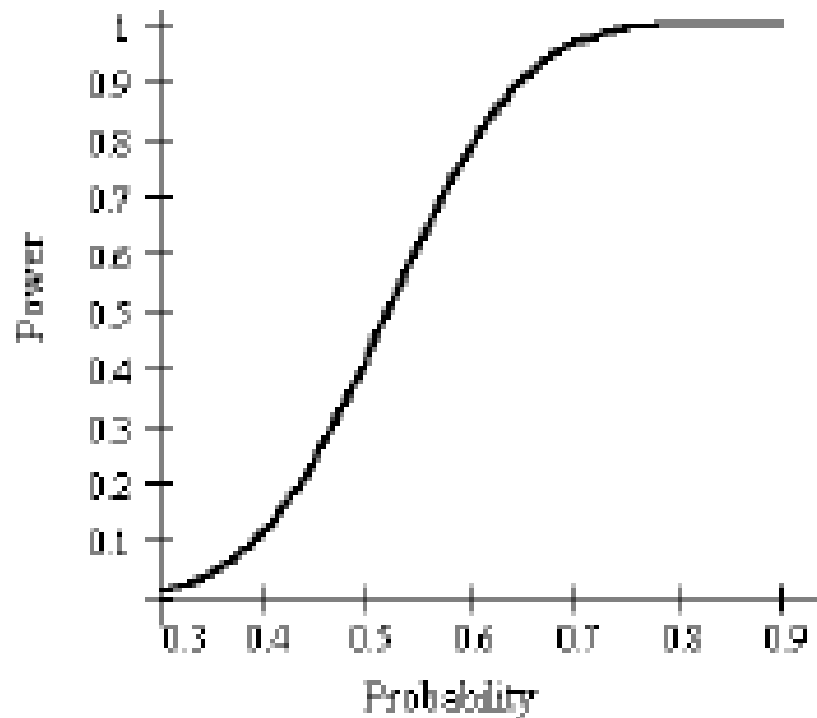


Figure 8: Distribution of Number of Correct Guesses with  $p = \frac{1}{2}$

# Erőfüggvény

A próba erőfüggvénye

$$\beta(\vartheta) = P_{\vartheta}(\xi \in X_k) = 1 - P_{\vartheta}(\xi \in X_e), \vartheta \in \Theta_1$$



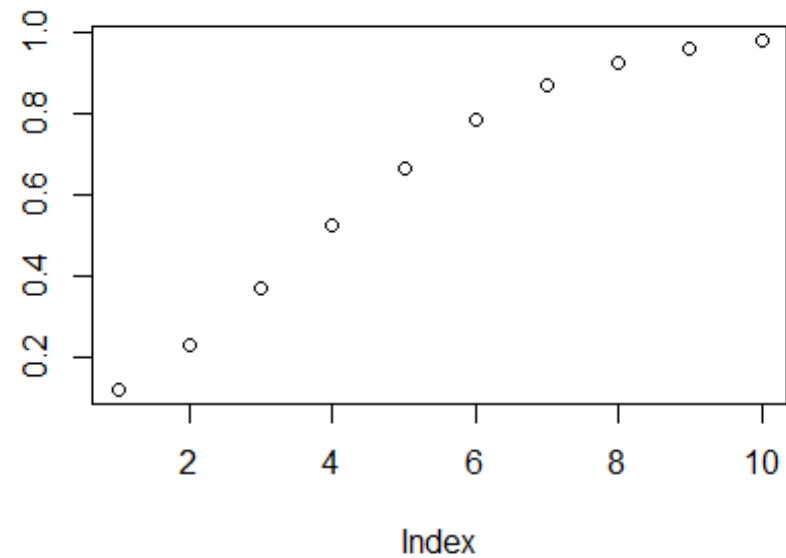


# 10 ezer ember koronavírusos tesztelése

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

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

# Próba ereje $c=28$ esetén



<b><math>p</math></b>	<b>erő</b>
$2 \cdot 2\text{‰}$	0.1222548
$3 \cdot 2\text{‰}$	0.2320277
$4 \cdot 2\text{‰}$	0.3728980
$5 \cdot 2\text{‰}$	0.5252410
$6 \cdot 2\text{‰}$	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

$\mathcal{G} \in \Theta_0$ -ra az elsőfajú hiba valószínűsége:

$$P_{\mathcal{G}}(T(\xi) > c) + \gamma P_{\mathcal{G}}(T(\xi) = c) = E_{\mathcal{G}}(\psi(\xi))$$

$\alpha$  a próba terjedelme, ha minden  $\mathcal{G} \in \Theta_0$ -ra

$$E_{\mathcal{G}}(\psi(\xi)) \leq \alpha$$

$\alpha$  a próba szignifikanciaszintje

(másképp: a próba pontos terjedelme),

$$\sup_{\mathcal{G} \in \Theta_0} E_{\mathcal{G}}(\psi(\xi)) = \alpha$$

## Legerősebb próba egyszerű hipotézis esetében

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

$\psi$  a legerősebb  $\alpha$ -terjedelmű próba, ha:

$$P_{g_0}(T(\xi) > c) + \gamma P_{g_0}(T(\xi) = c) = E_{g_0}(\psi(\xi)) \leq \alpha,$$

továbbá minden más  $\alpha$ -terjedelmű  $\psi'$  próbára, annak másodfajú hibavalószínűsége nagyobb:

$$E_{g_1}(1 - \psi(\xi)) \leq E_{g_1}(1 - \psi'(\xi)).$$

# A legerősebb próba

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

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

# Paraméteres próbák

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

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

$$H_0 : m = m_0$$

$$H_1 : m \neq m_0 \text{ (kétoldali ellenhipotézis)}$$

$$H_1' : m < m_0 \text{ (egyoldali ellenhipotézis)}$$

$$H_1'' : m > m_0 \text{ (egyoldali ellenhipotézis)}$$

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

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

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

$U$  – próba (kétoldali ellenhipotézis)

$$\Phi(u_y) = y$$

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

$$\begin{aligned} P_{m_0}(\xi \in X_k) &= P_{m_0}(|U| \geq u_{1-\alpha/2}) = 1 - \Phi(u_{1-\alpha/2}) + \Phi(-u_{1-\alpha/2}) = \\ &= 1 - (1 - \alpha/2) + 1 - (1 - \alpha/2) = \alpha. \end{aligned}$$

$$\beta(m) = P_m(\xi \in X_k) = P_m(|U| \geq u_{1-\alpha/2}) =$$

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

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

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