Statisztika

Előadások letölthetők a
http://www.cs.elte.hu/~arato/InfBC_stat*.pdf
címről

Becslési módszerek

- Példa: Egy tóban *N* hal van, számukat nem ismerjük. Első héten kihalásznak 1000 halat és megjelölik őket. A következő héten kihalásznak 5000-et és megszámolják a megjelölteket. 50-et találnak. Becsüljük meg *N*-et!

Természetes eljárás

Jelölje ξ a másodjára kihúzott halak számát.

Tudjuk, hogy ez hipergeometrikus eloszlású, így

$$L(50,N)=P_N(\xi=50)=\frac{\binom{1000}{50}\binom{N-1000}{4950}}{\binom{N}{5000}}.$$

Becslés

$$\hat{N}:\ L(50,\hat{N})=\max_N L(50,N)\Rightarrow \hat{N}=100000$$

Maximum likelihood becslés

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

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

Példa (indikátor)

$$L(\mathbf{x}, p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i},$$

$$l(\mathbf{x}, \theta) = \ln L(\mathbf{x}, p) = \left(\sum_{i=1}^{n} x_i \right) \ln p + \left(n - \sum_{i=1}^{n} x_i \right) \ln(1-p).$$

Likelihood egyenlet

$$\frac{\partial l(\mathbf{x}, \theta)}{\partial \theta} = \left(\sum_{i=1}^{n} x_i \right) \frac{1}{p} - \left(n - \sum_{i=1}^{n} x_i \right) \frac{1}{1-p} = 0$$

Ennek megoldása

$$p = \frac{\sum_{i=1}^{n} x_i}{n}.$$

És ez valóban maximumhely!

Így a ML becslés

$$\hat{p} = \frac{\sum_{i=1}^{n} \xi_i}{n}.$$

Konfidencia intervallum

- Def.: 1-α megbízhatóságú konfidencia intervallum: Olyan intervallum, mely legalább 1-α valószínűséggel tartalmazza a keresett paramétert.

$$P_{\vartheta}(T_1(\xi) < \vartheta < T_2(\xi)) \geq 1-\alpha,\ \forall \vartheta \in \Theta$$

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)

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

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

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

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

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

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

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

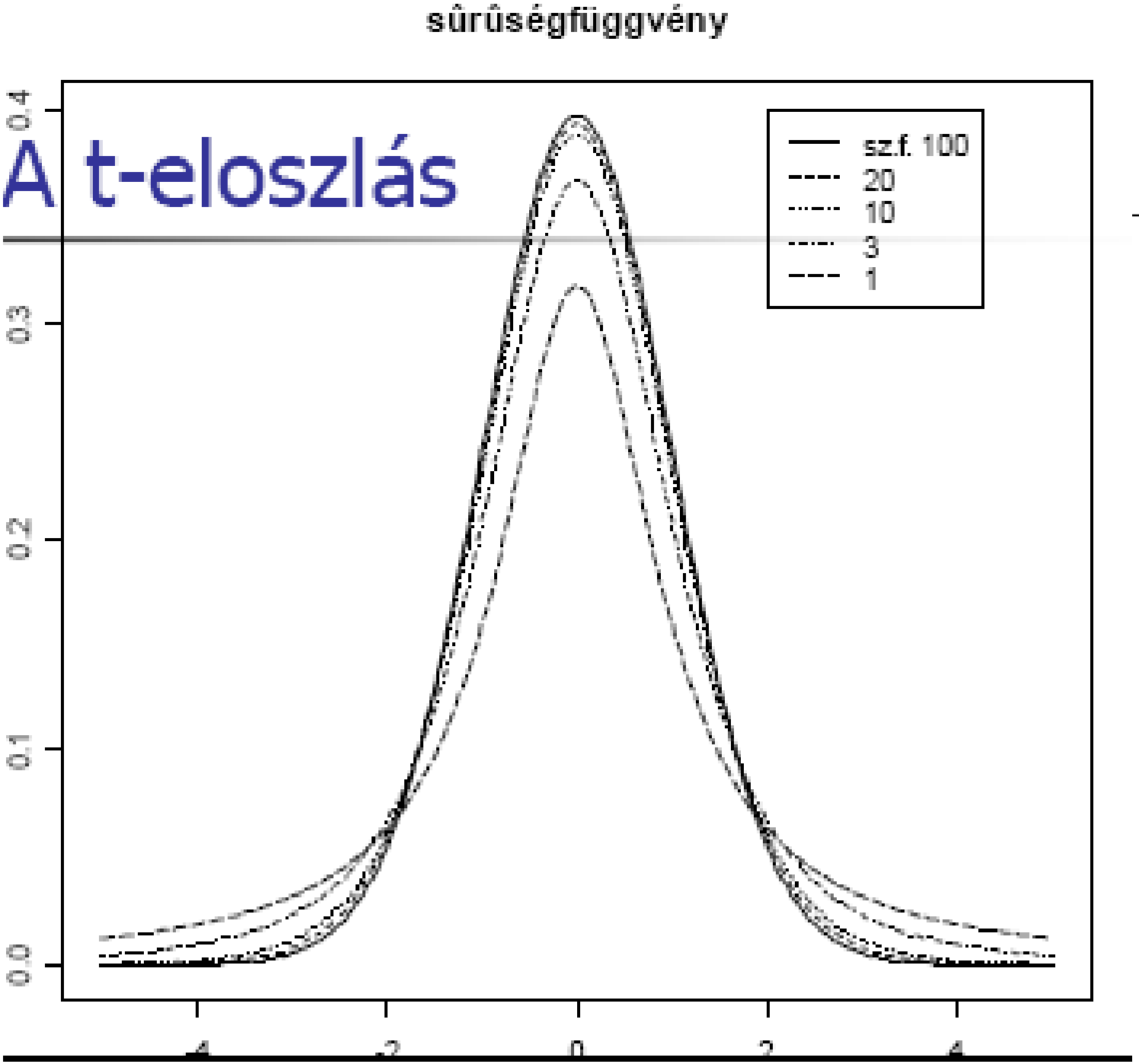
Konfidencia intervallum normális eloszlás várható értékére (ismeretlen szórás esetén)

- 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:

$$X_0, X_1, X_2, ..., X_n \sim N(0,1), \text{ függetlenek}$$

$$\frac{X_0}{\sqrt{\left(X_1^2 + X_2^2 + ... + X_n^2\right)/(n-1)}} \sim t_{n-1}$$

A t-eloszlás



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

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

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

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

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

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

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

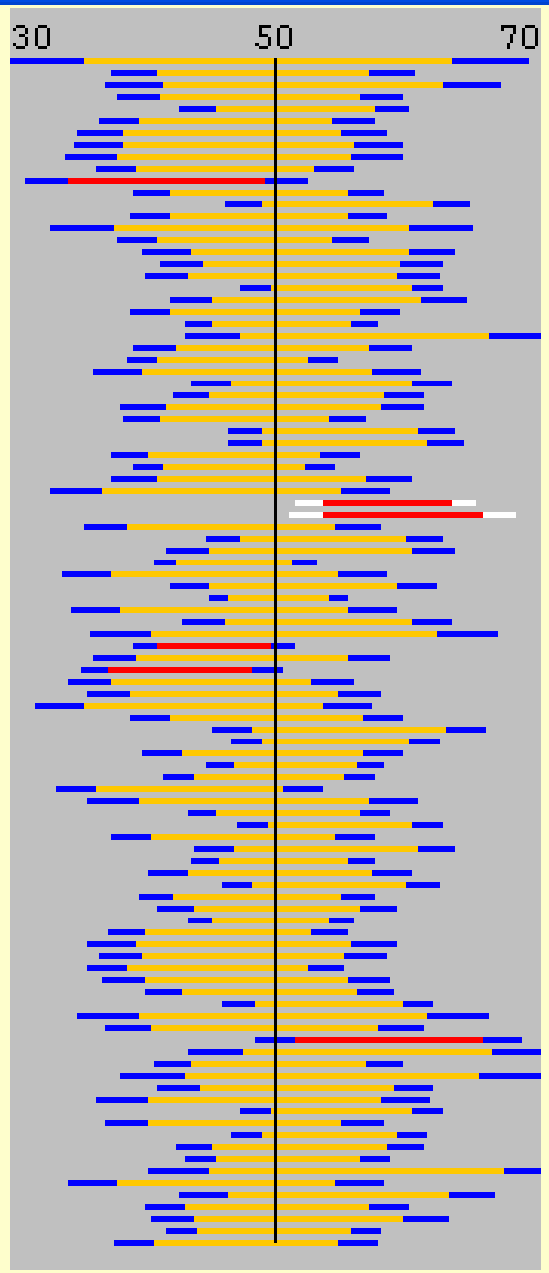
Példa (kenyér. folyt.)

- 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.

Szimulációs példa

- http://onlinestatbook.com/stat_sim/conf_interval/index.html

Confidence Interval Simulation



Sample size: 10 Sample Clear

When you click the sample button, 100 samples of the specified sample size (10, 15, or 20) will be taken from a population with a mean of 50 and a standard deviation of 10. The confidence interval on the mean will be computed for each. If the 95% confidence interval contains the population mean of 50 then a line will show the 95% confidence interval in orange and the 99% confidence interval in blue. If the 95% confidence interval does not contain the population mean then it will be shown in red. If the 99% interval does not contain the population mean it will be shown in white.

Cumulative Results:

	99% Conf. Int	95% Conf. Int
Contained 50	198	192
Did Not Contain 50	2	8
Proportion Contained	0.990	0.960

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

Mindenhol azt olvasni, hogy a napi/heti/havi bad beat nem „számít”, sokkal fontosabb a hosszútáv, amikor a matematikai esély érvényesül. Jelenlegi eredményeink mennyire reálisak? Mikor ésszerűbb inkább felhagyni a pókerrel? Mikor lehetünk optimisták? Mekkora játékszám szükséges ennek megállapítására?

Most bemutatok egy idevágó táblázatot, amely 95% pontossággal megadja a jelenlegi játékszámod és nyerési %-od alapján, hogy a jelen eredményeid mennyire lehetnek valósak. (Pl. Ha 60%-os vagy 100 játék után, akkor a valós nyerési százalékod igen nagy (95%-os) valószínűséggel 50,4-69,6% között van.) Igaz, csak három - 50-55-60 %-os - mutatóval dolgozik, de attól még igencsak hasznos a jövőbeni tervek megalapozottságához.

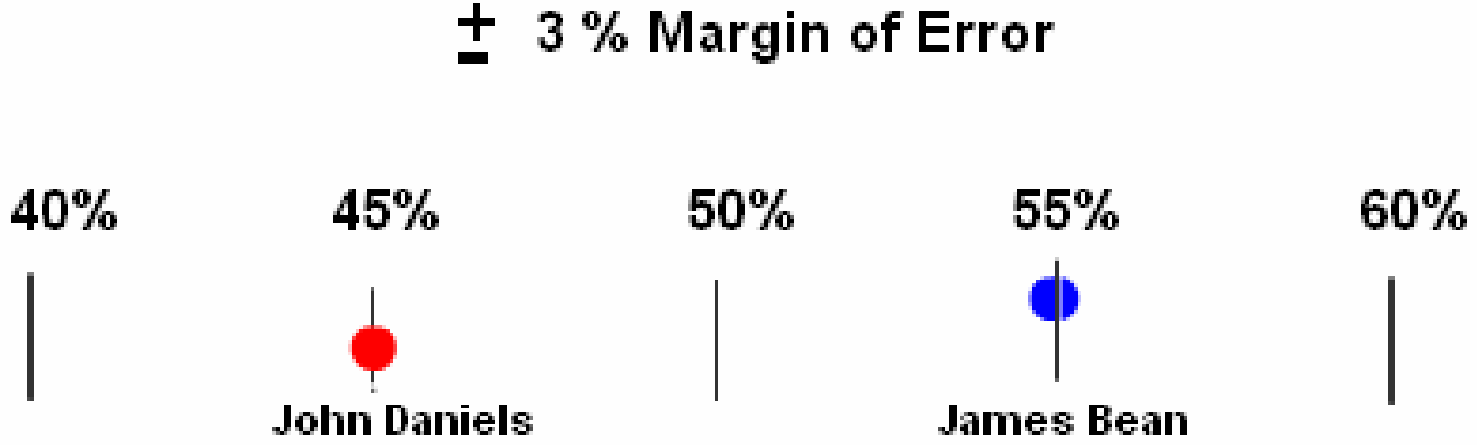
http://www.pokerakademia.com/poker_blogok/hideyoshi/mi_szamit_hosszutavnak_a_hu_sng_ban/

Ninety-five percent confidence interval (+/-) on HU SnG win percentage

Number of SnGs played	Win percentage experienced: 50.0%	55.0%	60.0%
10	31.0%	30.3%	30.4%
50	13.9%	13.3%	13.6%
100	9.3%	9.8%	9.6%
250	6.2%	6.2%	6.1%
500	4.4%	4.4%	4.3%
1,000	3.1%	3.1%	3.0%
2,000	2.2%	2.2%	2.1%
5,000	1.4%	1.4%	1.4%
10,000	1.0%	1.0%	1.0%
20,000	0.7%	0.7%	0.7%
50,000	0.4%	0.4%	0.4%

A táblázat helyességét nem ellenőriztem! /AM/

After polling 1000 eligible voters, the Star-Tribune Newspaper reported that 55% of Americans would vote for James Bean and 45% for John F Daniels +/- 3%.



Hipotézisvizsgálat

- H_0 nullhipotézis (jelezni akarjuk, ha nem igaz)
- H_1 ellenhipotézis
- Gyakran paraméterekkel fogalmazzuk meg:

$$H_0 : \vartheta \in \Theta_0$$

$$H_1 : \vartheta \in \Theta_1$$

$$\Theta_0 \bigcup \Theta_1 = \Theta$$

Példák

- Igaz-e, hogy 0,5 valószínűséggel születik fiúgyermek?
- H_0 : 0,5 valószínűséggel születik fiúgyermek
- H_1 : nem 0,5 valószínűséggel születik fiúgyermek

$$H_0 : p = 0,5$$

$$H_1 : p \neq 0,5$$

Példák (folyt.)

- Mi lehet egy vezető által okozott károk számának eloszlása?
- H_0: a kárszám Poisson eloszlású
- H_1: a kárszám nem Poisson eloszlású

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

Ki tanul jobban?

2009. január 5-ei vizsga

H_0: A nők jobban tanulnak

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

Lehetséges hibák

- Elsőfajú hiba: H_0 igaz, de elutasítjuk
- Másodfajú hiba: H_0 hamis, de elfogadjuk
- Döntésünknek a megfigyelésektől kell függnie.
- Mintateret 2 részre osztjuk: elfogadási és elutasítási tartományra.

Lehetséges hibák

- Elsőfajú hiba: H_0 igaz, de elutasítjuk
- Másodfajú hiba: H_0 hamis, de elfogadjuk

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

Alapfogalmak

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

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

Elsőfajú hiba valószínűsége

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

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

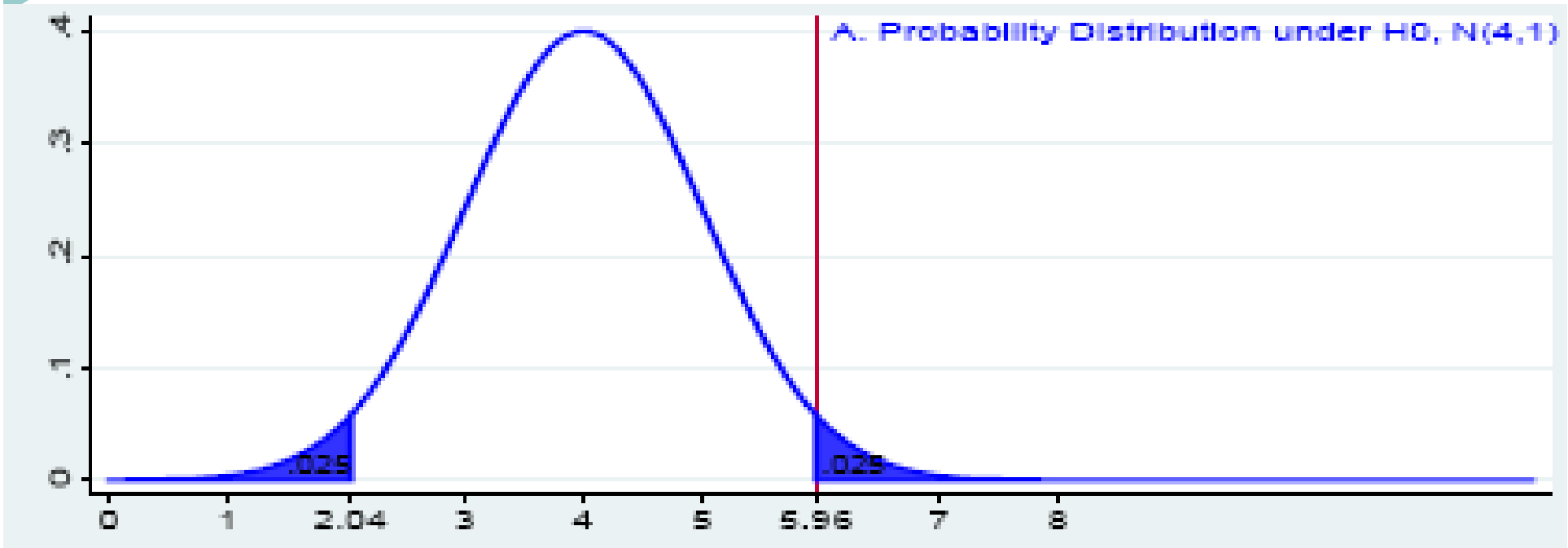
α a próba szignifikanciaszintje

(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

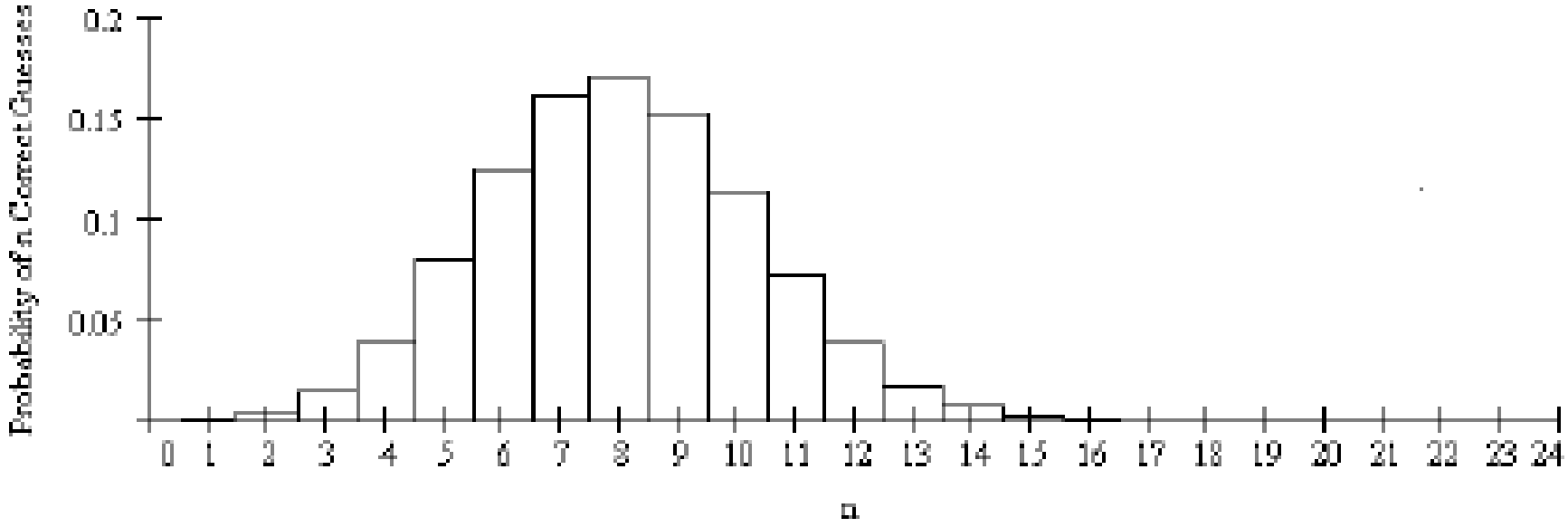


Figure 6: Distribution of Number of Correct Guesses with $p = \frac{1}{3}$

Kritikus tartomány megválasztása

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

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

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

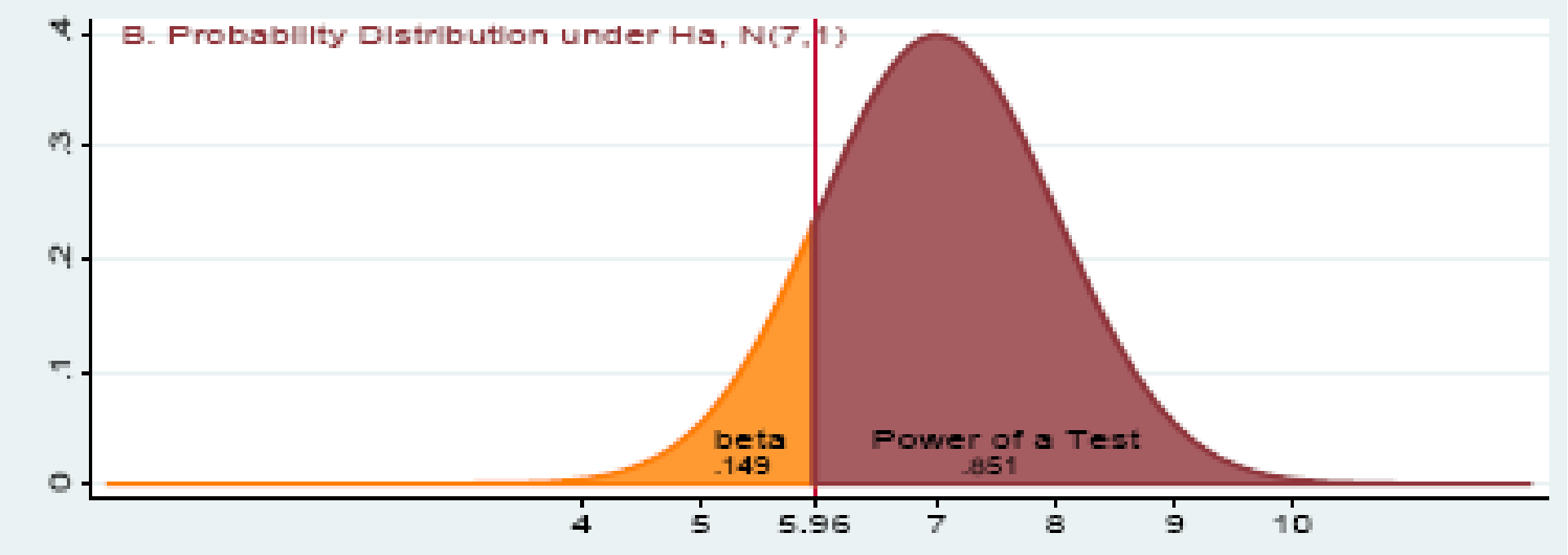
$$p\text{-value} = \sum_{y=11}^{24} \binom{24}{y} \left(\frac{1}{3}\right)^y \left(\frac{2}{3}\right)^{24-y} = 0.14$$

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

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

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

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

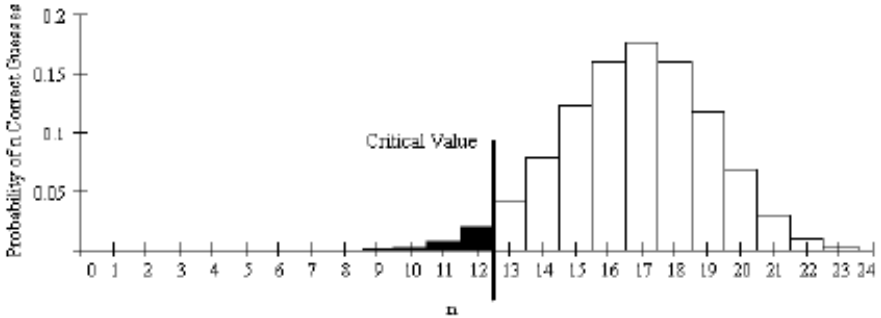


Figure 9: Distribution of Number of Correct Guesses with $p = 0.7$

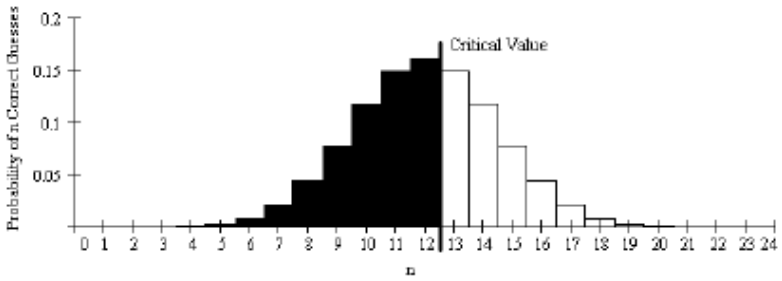
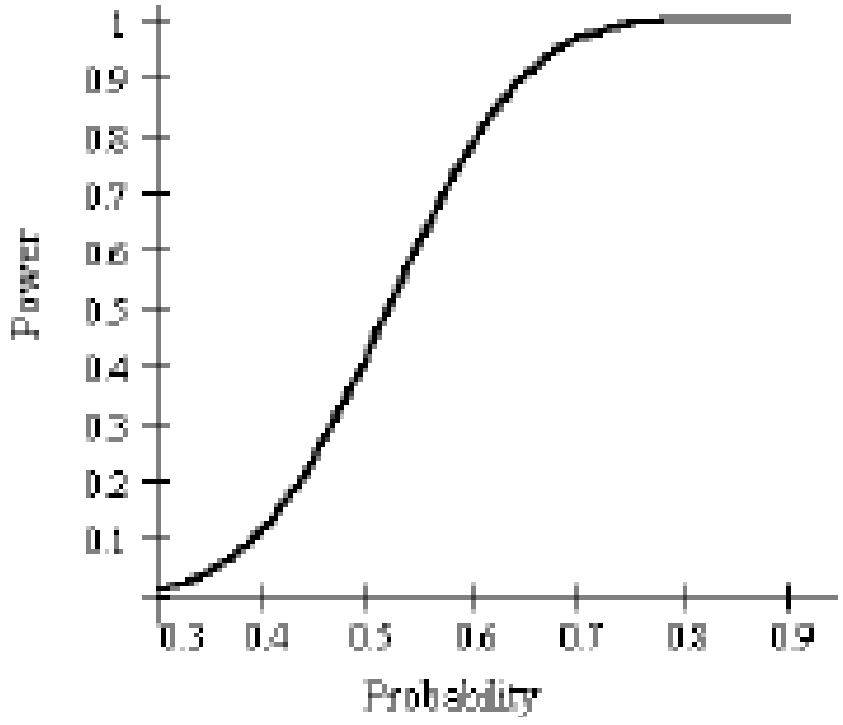


Figure 8: Distribution of Number of Correct Guesses with $p = \frac{1}{2}$

Erőfüggvény

A próba erőfüggvénye

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



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)

$$\Phi(u_y) = y$$

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

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

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

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

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

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

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

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 : $\left|\Theta_0\right| = \left|\Theta_1\right| = 1$.

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.

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^*:$m\neq0$ 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:

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