Valószínűségszámítás és Statisztika

11. előadás

2023. május 18.

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

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

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:

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

Elsőfajú hiba valószínűsége

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

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

α a próba szignifikanciaszintje

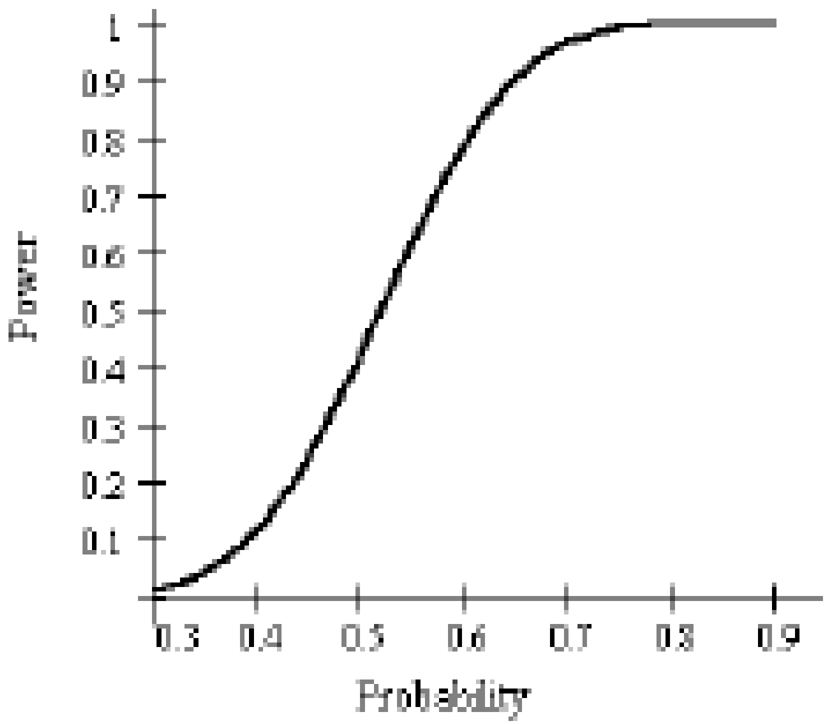
(másképp: a próba pontos terjedelme),

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

Erőfüggvény

A próba erőfüggvénye

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



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:

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

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

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

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:

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

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

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

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

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

$$\Phi(u_y)=y$$

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

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

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

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

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

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

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

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.

Megjegyzések

- Ha egyoldali az ellenhipotézis, akkor a kritikus tartomány

$$t > t_{1-\alpha.n-1}(m > m_0),$$

- illetve

$$t < -t_{1-\alpha.n-1}\text{alakú } (m < m_0).$$

Megjegyzések

- Ezek is legerősebb próbák!
- 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).

Példa

- A butitizmus elleni védőoltás hatásos, ha az oltás után 2 héttel az antitestek száma meghaladja az 5 purc szintet.
- 37 beteget oltottak be. 2 héttel az oltás után az átlagos antitest szintjük 5,7 purc volt, a mérések korrigált tapasztalati szórása pedig 1,3 volt.
- Hatásos-e az oltás?

Példa (folyt.)

- Ez így nem jó! Kellenének az egyedi adatok!
- Feltételezzük, hogy $X_1, X_2, \dots, X_{37} \sim N(m, \sigma^2)$
- $H_0: m = 5, H_1: m > 5$
- $t = \sqrt{37}\frac{5{,}7-5}{1{,}3} = 3{,}275334$
- $t_{1-5\%.37-1} = 1{,}688298$
- $P(t_{36} < 3{,}275334) = 0{,}9988308$

Kétoldali próbák és konfidencia intervallumok

- A normális eloszlásnál a várható értékre vonatkozó α terjedelmű próbánál láttuk, hogy a

$$H_0\colon m = m_0$$

hipotézist a

$$H_1\colon m \neq m_0$$

hipotézissel szemben pontosan akkor fogadjuk el, ha m_0 benne van az $1 - \alpha$ megbízhatóságú konfidencia intervallumban.

Kétmintás eset: párosított megfigyelések

- Példa: Van-e különbség Budapest és Cegléd napi átlaghőmérséklete között?

$$H_0\colon m_1 = m_2$$ 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.

$$H_0^*\colon m = 0\,,\ H_1^*\colon m \neq 0$$ az új hipotézisek.

Kétmintás eset: független minták

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

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

Kritikus tartomány: mint az egymintás esetben

Ha ismeretlenek, de azonosak a szórások:

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

A szórás vizsgálata kétmintás esetben: F-próba

- H_0: $\sigma_1 = \sigma_2$
- Két független, n, illetve k 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 $n-1, k-1$ szabadságfokú F eloszlás $1-\alpha/2$ kvantilise

(n a számlálóbeli, k 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 $A_1, A_2, \dots, A_r$ teljes eseményrendszerre teljesül $P(A_1) = p_1, P(A_2) = p_2, \dots, P(A_r) = p_r$
- 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 $1-\alpha$ kvantilise, elutasítjuk a nullhipotézist.

χ-négyzet próba (folytatás)

- Miért is ez a határeloszlás?

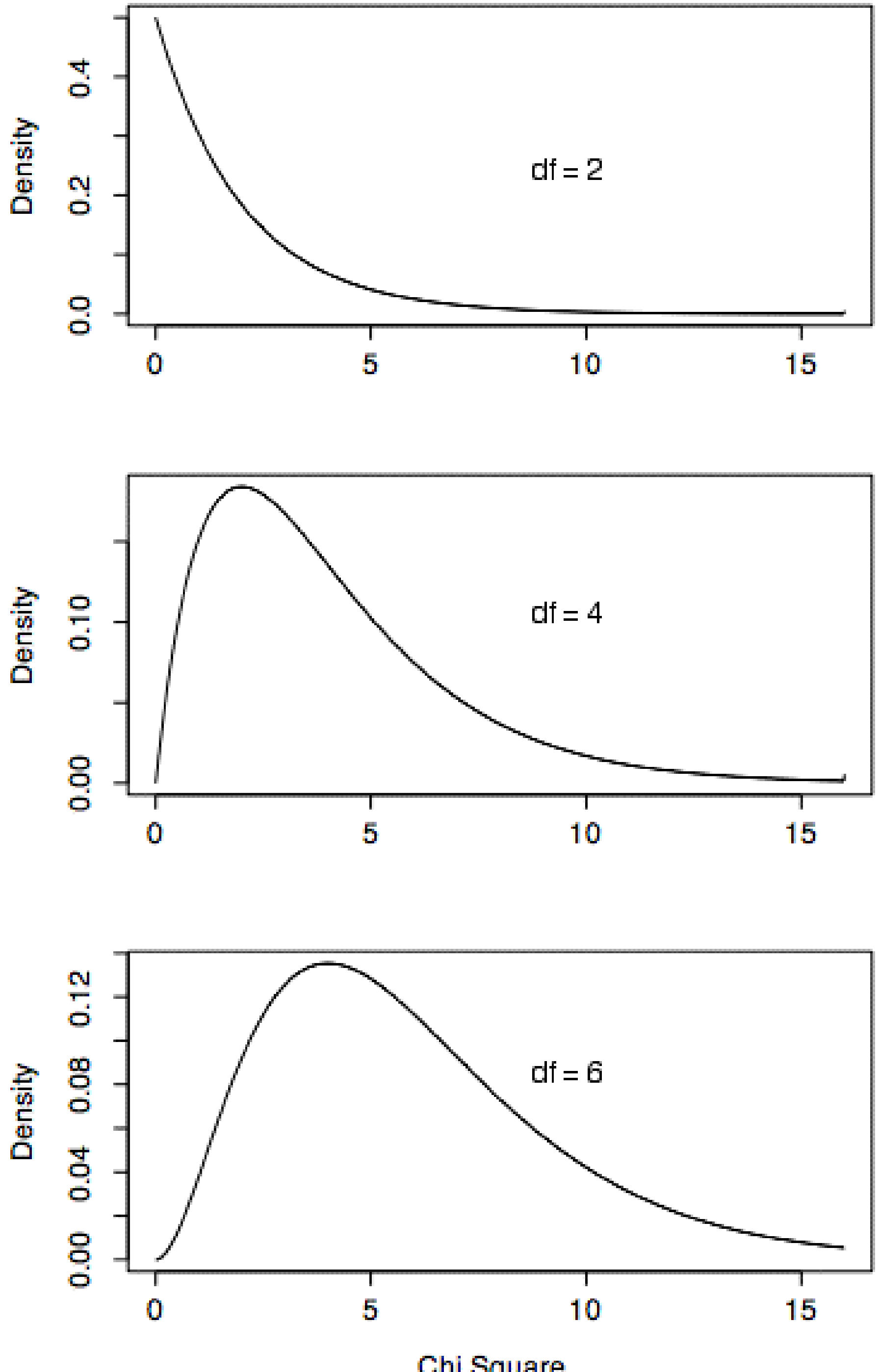
$r = 2,\ \ H_0 : P(A) = p,\ \nu : A$ gyakorisága n kísérletből

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

$\xi_i = 1,$ 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:

$H_0 : \xi_1, \dots, \xi_n \;\; F$ eloszlásfüggvényűek

- Visszavezetjük az előző esetre

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

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

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:

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

$\vartheta_1, \ldots, \vartheta_s$ 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

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

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:

$\hat{\lambda} = 0.709$

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

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

$P\left(\chi^2_3 > 17.7\right) = 0.05\% \Rightarrow$

Elutasítjuk Poisson eloszlás hipotézisét!