Valószínűségszámítás és Statisztika

11. előadás

2017. december 4.

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

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

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?

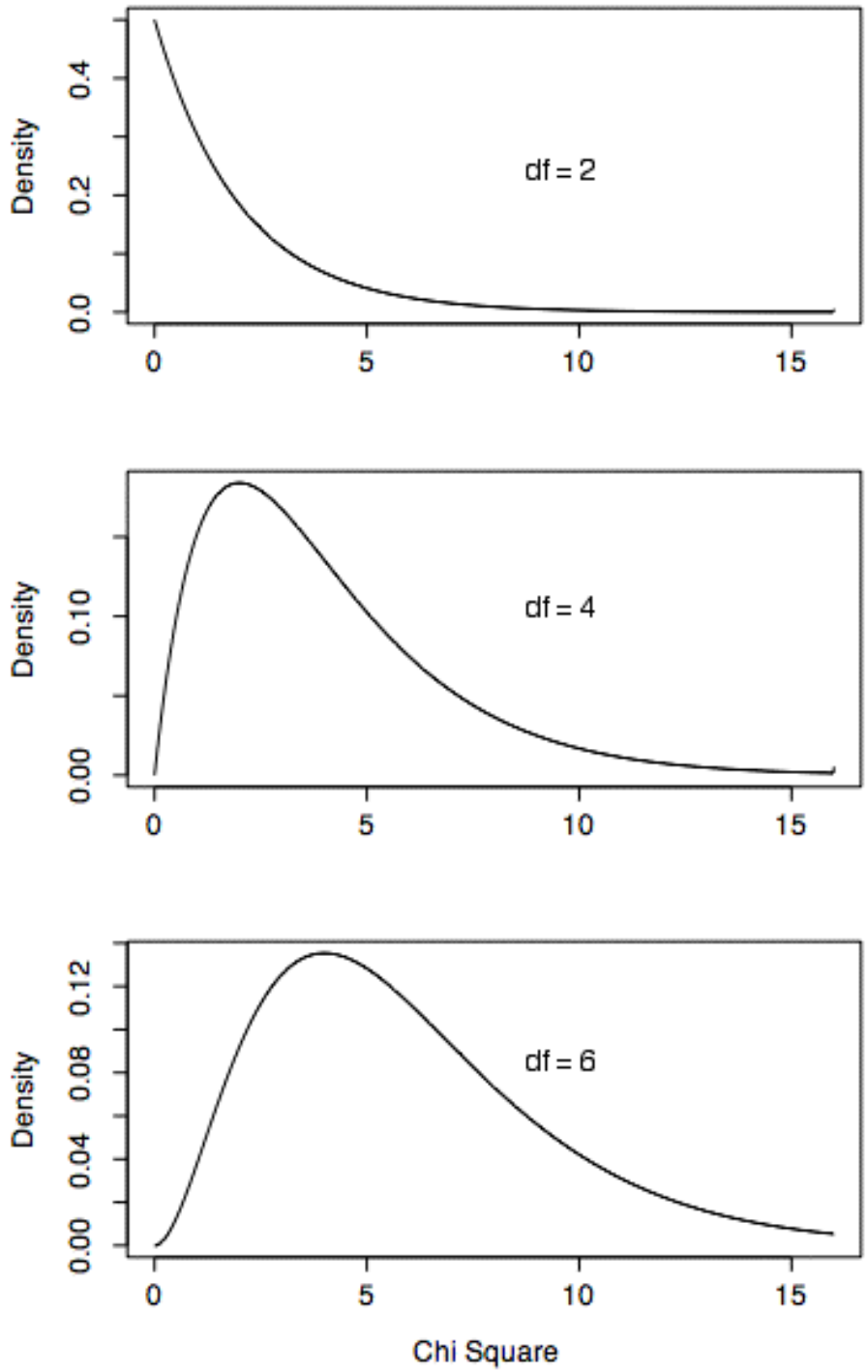
$r = 2,\ H_0 : P(A) = p,\ \nu : A$ gyakorisága n kísérletből

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

$\xi_i = 1$, 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 (*n*)
Veze-tők száma	129524	16267	1966	211	31	5	1	1	0	148006
Poisson	128 433	18 218	1 292	61	2,2	0,06	0,001	3E-05	5E-07	
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

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

Szívbetegek diétája

- http://onlinestatbook.com/case_studies/diet.html
- The subjects, 605 survivors of a heart attack, were randomly assigned follow either (1) a diet close to the "prudent diet step 1" of the American Heart Association (control group) or (2) a Mediterranean-type diet consisting of more bread and cereals, more fresh fruit and vegetables, more grains, more fish, fewer delicatessen foods, less meat. An experimental canola-oil-based margarine was used instead of butter or cream. The oils recommended for salad and food preparation were canola and olive oils exclusively. Moderate red wine consumption was allowed.
- Over a four-year period, patients in the experimental condition were initially seen by the dietician, two months later, and then once a year. Compliance with the dietary intervention was checked by a dietary survey and analyses of plasma fatty acids. Patients in the control group were expected to follow the dietary advice given by their physician.

	Cancers	Deaths	Nonfatal illness	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

	Cancers	Deaths	Nonfatal illness	Healthy	Total
AHA	15 (11.02)	24 (19.03)	25 (16.53)	239 (256.42)	303
Mediterranean	7 (10.98)	14 (18.97)	8 (16.47)	273 (255.58)	302
Total	22	38	33	512	605

-
- $\chi^2 = n \sum_{i,j} \frac{\left(\nu_{i,j} - \frac{\nu_{i,\cdot}\nu_{\cdot,j}}{n}\right)^2}{\nu_{i,\cdot}\nu_{\cdot,j}} = 16{,}55$
- $P(\chi_3^2 > 16{,}55\ ) = 0{,}0009 \Longrightarrow$

elutasítjuk a hipotézist