10. NEPARAMETARSKI TESTOVI (SI, MAST)

Profesor Milan Merkle
emerkle@etf.rs milanmerkle.etf.rs

Matematička Statistika-jesen 2019

Empirijske funkcije raspodele

(Odeljak 6.4. iz udžbenika, str. 148-151)

Kako se može oceniti funkcija raspodele iz uzorka $X_1, X_2, \ldots, X_n$?
Funkcija

$$F_n(x) = \frac{\textit{Broj elemenata uzorka koji su} \leq x}{n}$$

zove se empirijska funkcija raspodele.
Preko varijacionog niza:

$$F_n(x) = \begin{cases} 0, & \text{ako je } x < X_{(1)}\ , \\ k/n, & \text{ako je } X_{(k)} \leq x < X_{(k+1)},\ 1 \leq k \leq n-1 \\ 1, & \text{ako je } x \geq X_{(n)}. \end{cases}$$

Ovo je slučajna funkcija jer zavisi od slučajnog uzorka. √

Primer 131

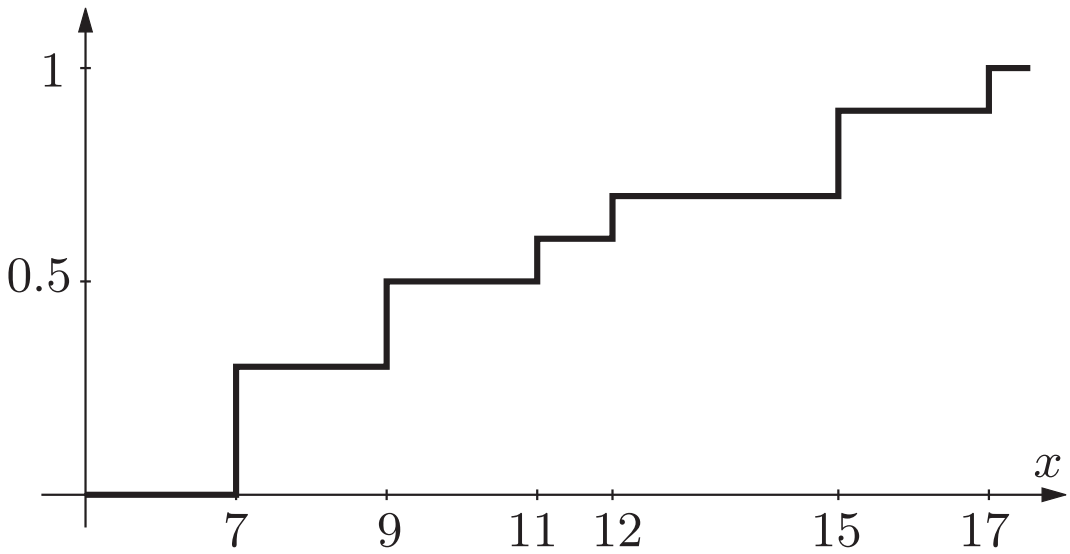
Uzorak obima 10:

$$9, 15, 7, 11, 17, 9, 7, 12, 7, 15$$

Varijacioni niz je:

$$7, 7, 7, 9, 9, 11, 12, 15, 15, 17$$

Empirijska funkcija raspodele:



Konvergencija empirijskih funkcija

- *Za fiksirano x, $F_n(x) = \frac{k}{n}$ - relativna frekvencija događaja $X \leq x$.*
- *Prema ZVB, $\lim F_n(x) = F(x)$ za svako $x \in \mathbb{R}$.*

Teorema 6.13 $\lim_{n \to +\infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0$ *sa verovatnoćom 1.*

Uniformna konvergencija - znači da za svako $\varepsilon > 0$ postoji n_0 tako da za svako $n \geq n_0$ važi da je $|F_n(x) - F(x)| < \varepsilon$ za svako $x \in \mathbb{R}$.

Konvergencija empirijskih funkcija

- *Za fiksirano* x, $F_n(x) = \frac{k}{n}$ *- relativna frekvencija događaja* $X \leq x$.
- *Prema ZVB,* $\lim F_n(x) = F(x)$ *za svako* $x \in \mathbb{R}$.

Teorema 6.13 $\lim_{n \to +\infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0$ *sa verovatnoćom 1.*

Uniformna konvergencija - znači da za svako $\varepsilon > 0$ *postoji* n_0 *tako da za svako* $n \geq n_0$ *važi da je* $|F_n(x) - F(x)| < \varepsilon$ *za svako* $x \in \mathbb{R}$.

Konvergencija empirijskih funkcija

- *Za fiksirano* x, $F_n(x) = \frac{k}{n}$ *- relativna frekvencija događaja* $X \leq x$.
- *Prema ZVB,* $\lim F_n(x) = F(x)$ *za svako* $x \in \mathbb{R}$.

Teorema 6.13 $\lim_{n \to +\infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0$ *sa verovatnoćom 1.*

Uniformna konvergencija - znači da za svako $\varepsilon > 0$ *postoji* n_0 *tako da za svako* $n \geq n_0$ *važi da je* $|F_n(x) - F(x)| < \varepsilon$ *za svako* $x \in \mathbb{R}$.

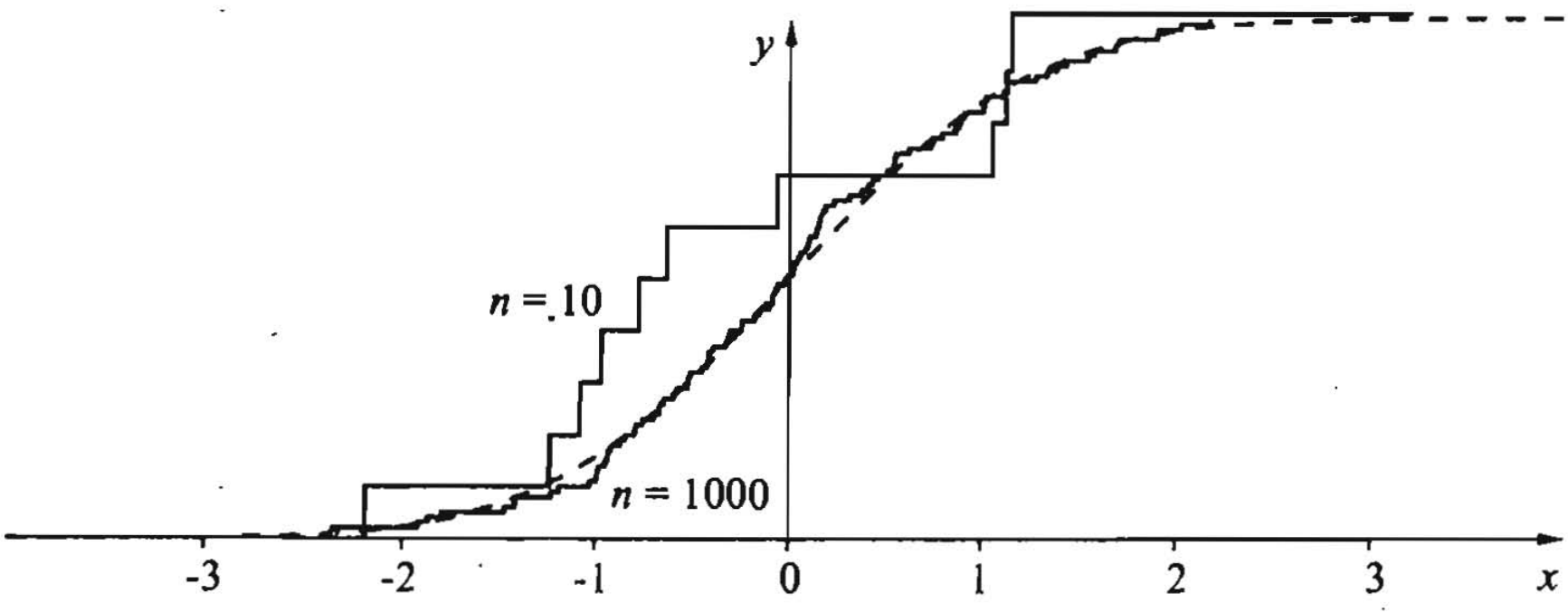
Konvergencija empirijskih funkcija

- *Za fiksirano* x, $F_n(x) = \frac{k}{n}$ *- relativna frekvencija događaja* $X \leq x$.
- *Prema ZVB,* $\lim F_n(x) = F(x)$ *za svako* $x \in \mathbb{R}$.

Teorema 6.13 $\lim_{n \to +\infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0$ *sa verovatnoćom 1.*

Uniformna konvergencija - znači da za svako $\varepsilon > 0$ *postoji* n_0 *tako da za svako* $n \geq n_0$ *važi da je* $|F_n(x) - F(x)| < \varepsilon$ *za svako* $x \in \mathbb{R}$.

Primer 133



Slika 32. Empirijske funkcije raspodele F_n za standardnu normalnu raspodelu iz uzoraka obima n, za $n = 10$ i $n = 1000$ (računarska simulacija) i funkcija raspodele Φ standardne normalne raspodele (isprekidana kriva).

Ocenjivanje kvantila preko empirijske funkcije raspodele

$X_{(1)}, \ldots, X_{(n)}$ varijacioni niz uzorka. Za ocenu kvantila x_p reda p $(0 < p < 1)$ uzima se najmanji kvantil empirijske raspodele uzorka:

$$\hat{x}_p = \min_{x \in \mathbf{R}} \{x \mid F_n(x) \geq p\},$$

gde je F_n empirijska funkcija raspodele dobijena na osnovu datog uzorka.

Ocena kvantila reda p je $X_{(k)}$, gde je k određeno sa:

$$\frac{k-1}{n} < p, \quad \text{i} \quad \frac{k}{n} \geq p, \quad \text{tj.} \quad k - 1 < np \leq k.$$

Medijana: kvantil reda $1/2$.

- *Uzorak:* $11, 11, 13, 14, 15, 16 \quad \hat{x}_{1/2} = 13$
- *Uzorak:* $10, 13, 14, 16, 17 \quad \hat{x}_{1/2} = 14$

Ocenjivanje kvantila preko empirijske funkcije raspodele

$X_{(1)}, \ldots, X_{(n)}$ varijacioni niz uzorka. Za ocenu kvantila x_p reda p $(0 < p < 1)$ uzima se najmanji kvantil empirijske raspodele uzorka:

$$\hat{x}_p = \min_{x \in \mathbf{R}} \{x \mid F_n(x) \geq p\},$$

gde je F_n empirijska funkcija raspodele dobijena na osnovu datog uzorka.

Ocena kvantila reda p je $X_{(k)}$, gde je k određeno sa:

$$\frac{k-1}{n} < p, \quad \text{i} \quad \frac{k}{n} \geq p, \quad \text{tj.} \quad k - 1 < np \leq k.$$

Medijana: kvantil reda $1/2$.

- *Uzorak:* $11, 11, 13, 14, 15, 16 \quad \hat{x}_{1/2} = 13$
- *Uzorak:* $10, 13, 14, 16, 17 \quad \hat{x}_{1/2} = 14$

Ocenjivanje kvantila preko empirijske funkcije raspodele

$X_{(1)}, \ldots, X_{(n)}$ varijacioni niz uzorka. Za ocenu kvantila x_p reda p $(0 < p < 1)$ uzima se najmanji kvantil empirijske raspodele uzorka:

$$\hat{x}_p = \min_{x \in \mathbf{R}} \{x \mid F_n(x) \geq p\},$$

gde je F_n empirijska funkcija raspodele dobijena na osnovu datog uzorka.

Ocena kvantila reda p je $X_{(k)}$, gde je k određeno sa:

$$\frac{k-1}{n} < p, \quad \text{i} \quad \frac{k}{n} \geq p, \quad \text{tj.} \quad k - 1 < np \leq k.$$

Medijana: kvantil reda $1/2$.

- *Uzorak:* $11, 11, 13, 14, 15, 16 \quad \hat{x}_{1/2} = 13$
- *Uzorak:* $10, 13, 14, 16, 17 \quad \hat{x}_{1/2} = 14$

Ocenjivanje kvantila preko empirijske funkcije raspodele

$X_{(1)}, \ldots, X_{(n)}$ varijacioni niz uzorka. Za ocenu kvantila x_p reda p $(0 < p < 1)$ uzima se najmanji kvantil empirijske raspodele uzorka:

$$\hat{x}_p = \min_{x \in \mathbf{R}} \{x \mid F_n(x) \geq p\},$$

gde je F_n empirijska funkcija raspodele dobijena na osnovu datog uzorka.

Ocena kvantila reda p je $X_{(k)}$, gde je k određeno sa:

$$\frac{k-1}{n} < p, \quad \text{i} \quad \frac{k}{n} \geq p, \quad \text{tj.} \quad k - 1 < np \leq k.$$

Medijana: kvantil reda $1/2$.

- *Uzorak:* $11, 11, 13, 14, 15, 16 \quad \hat{x}_{1/2} = 13$
- *Uzorak:* $10, 13, 14, 16, 17 \quad \hat{x}_{1/2} = 14$

Ocenjivanje kvantila preko empirijske funkcije raspodele

$X_{(1)}, \ldots, X_{(n)}$ varijacioni niz uzorka. Za ocenu kvantila x_p reda p $(0 < p < 1)$ uzima se najmanji kvantil empirijske raspodele uzorka:

$$\hat{x}_p = \min_{x \in \mathbf{R}} \{x \mid F_n(x) \geq p\},$$

gde je F_n empirijska funkcija raspodele dobijena na osnovu datog uzorka.

Ocena kvantila reda p je $X_{(k)}$, gde je k određeno sa:

$$\frac{k-1}{n} < p, \quad \text{i} \quad \frac{k}{n} \geq p, \quad \text{tj.} \quad k - 1 < np \leq k.$$

Medijana: kvantil reda $1/2$.

- *Uzorak:* $11, 11, 13, 14, 15, 16 \quad \hat{x}_{1/2} = 13$
- *Uzorak:* $10, 13, 14, 16, 17 \quad \hat{x}_{1/2} = 14$

Intervali poverenja za kvantile

Teorema 8.10 *Neka je $X_{(1)}, \ldots, X_{(n)}$ varijacioni niz uzorka iz neprekidne raspodele čiji je kvantil reda p jednak x_p. Tada je, za $1 \leq k_1 < k_2 \leq n$,*

$$P(X_{(k_1)} \leq x_p \leq X_{(k_2)}) = \sum_{k=k_1}^{k_2-1} \binom{n}{k} p^k (1-p)^{n-k}. \qquad (1)$$

Dokaz. Neka je Y broj onih X_i iz uzorka koji su $\leq x_p$.
$X_{k_1} \leq x_p \leq X_{k_2}$ ako i samo ako je $k_1 \leq Y < k_2$

Primer 146. *Medijana: $n = 13$, $k_1 = 4$, $k_2 = 10$*
Ocena: $\hat{x}_{1/2} = X_{(7)}$,
$P(X_{(4)} \leq x_{1/2} \leq X_{(9)}) = 0.908$
Ne zavisi od raspodele, intervali su egzaktni.

Intervali poverenja za kvantile

Teorema 8.10 *Neka je $X_{(1)}, \ldots, X_{(n)}$ varijacioni niz uzorka iz neprekidne raspodele čiji je kvantil reda p jednak x_p. Tada je, za $1 \leq k_1 < k_2 \leq n$,*

$$P(X_{(k_1)} \leq x_p \leq X_{(k_2)}) = \sum_{k=k_1}^{k_2-1} \binom{n}{k} p^k (1-p)^{n-k}. \qquad (1)$$

Dokaz. Neka je Y broj onih X_i iz uzorka koji su $\leq x_p$.
$X_{k_1} \leq x_p \leq X_{k_2}$ ako i samo ako je $k_1 \leq Y < k_2$

Primer 146. *Medijana: $n = 13$, $k_1 = 4$, $k_2 = 10$*
Ocena: $\hat{x}_{1/2} = X_{(7)}$,
$P(X_{(4)} \leq x_{1/2} \leq X_{(9)}) = 0.908$
Ne zavisi od raspodele, intervali su egzaktni.

Intervali poverenja za kvantile

Teorema 8.10 *Neka je $X_{(1)}, \ldots, X_{(n)}$ varijacioni niz uzorka iz neprekidne raspodele čiji je kvantil reda p jednak x_p. Tada je, za $1 \leq k_1 < k_2 \leq n$,*

$$P(X_{(k_1)} \leq x_p \leq X_{(k_2)}) = \sum_{k=k_1}^{k_2-1} \binom{n}{k} p^k (1-p)^{n-k}. \qquad (1)$$

Dokaz. Neka je Y broj onih X_i iz uzorka koji su $\leq x_p$.
$X_{k_1} \leq x_p \leq X_{k_2}$ ako i samo ako je $k_1 \leq Y < k_2$

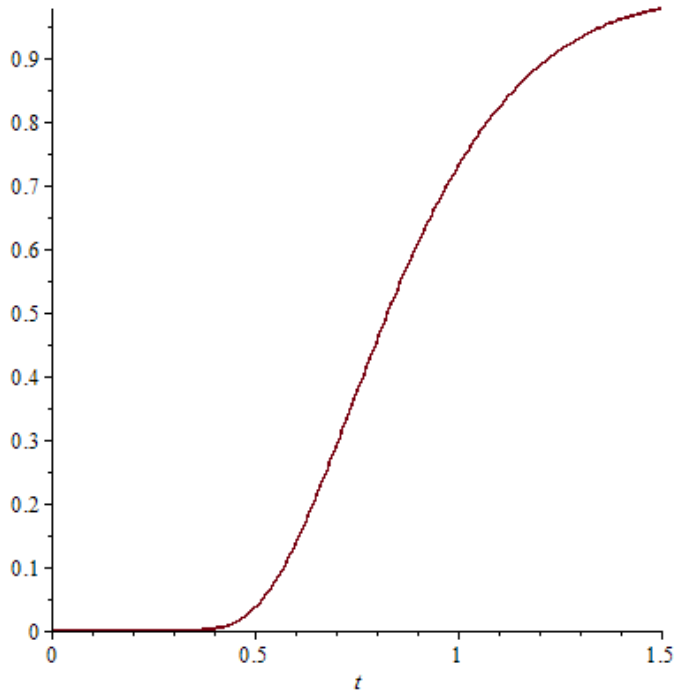
Primer 146. *Medijana: $n = 13$, $k_1 = 4$, $k_2 = 10$*
Ocena: $\hat{x}_{1/2} = X_{(7)}$,
$P(X_{(4)} \leq x_{1/2} \leq X_{(9)}) = 0.908$
Ne zavisi od raspodele, intervali su egzaktni.

K-raspodela (Kolmogorov) $X \sim K$

Funkcija raspodele Kolmogorova:

$$K(t) = 1 - 2\sum_{k=1}^{+\infty}(-1)^{k-1}e^{-2k^2t^2}, \quad t > 0$$

Vrednosti se računaju numeričkim metodima.



Statistika Kolmogorova i njena raspodela

Teorema 6.14 *Sa uzorkom obima n iz neprekidne raspodele sa funkcijom raspodele F, važi*

$$\lim_{n\to+\infty} P(\sqrt{n}\sup_{x\in\mathbf{R}}|F_n(x)-F(x)|\le t)=K(t)$$

Statistika $\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|$ zove se statistika Kolmogorova.

Aproksimacija za $n\ge 30$:

$$\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\sim K$$

Primer 134 *Naći najmanji obim uzorka n tako da je $|F_n(t)-F(t)|<0.01$ sa verovatnoćom 0.99.*

Iz tablice: $K(1.63)=0.99$, $n\ge 3969$.

Statistika Kolmogorova i njena raspodela

Teorema 6.14 *Sa uzorkom obima n iz neprekidne raspodele sa funkcijom raspodele F, važi*

$$\lim_{n\to+\infty} P(\sqrt{n}\sup_{x\in\mathbf{R}}|F_n(x)-F(x)|\le t)=K(t)$$

Statistika $\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|$ zove se statistika Kolmogorova.

Aproksimacija za $n\ge 30$:

$$\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\sim K$$

Primer 134 *Naći najmanji obim uzorka n tako da je $|F_n(t)-F(t)|<0.01$ sa verovatnoćom 0.99.*

Iz tablice: $K(1.63)=0.99$, $n\ge 3969$.

Statistika Kolmogorova i njena raspodela

Teorema 6.14 *Sa uzorkom obima n iz neprekidne raspodele sa funkcijom raspodele F, važi*

$$\lim_{n\to+\infty} P(\sqrt{n}\sup_{x\in\mathbf{R}}|F_n(x)-F(x)|\le t)=K(t)$$

Statistika $\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|$ zove se statistika Kolmogorova.

Aproksimacija za $n\ge 30$:

$$\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\sim K$$

Primer 134 *Naći najmanji obim uzorka n tako da je $|F_n(t)-F(t)|<0.01$ sa verovatnoćom 0.99.*

Iz tablice: $K(1.63)=0.99$, $n\ge 3969$.

Test Kolmogorova i Smirnova

F_0 - funkcija raspodele neprekidne slučajne promenljive.

Test hipoteze $H_0: \ F = F_0$ protiv komplementarne hipoteze $H_1: \ F \neq F_0$ sa sa nivoom značajnosti α (veliki uzorak):

Ako je

$$\lambda = \sqrt{n} \sup_{x \in \mathbf{R}} |F_n(x) - F_0(x)| > \varepsilon_{1-\alpha} \, ,$$

hipotezu H_0 odbacujemo.

Na isti način se testira pripadnost familiji neprekidnih raspodela (normalna, eksponencijalna...)- parametri se najpre ocene.

Pri testiranju neparametarskih hipoteza obično želimo da dokažemo H_0 !

Test Kolmogorova i Smirnova

F_0 - funkcija raspodele neprekidne slučajne promenljive.

Test hipoteze $H_0: \ F = F_0$ protiv komplementarne hipoteze $H_1: \ F \neq F_0$ sa sa nivoom značajnosti α (veliki uzorak):

Ako je

$$\lambda = \sqrt{n} \sup_{x \in \mathbf{R}} |F_n(x) - F_0(x)| > \varepsilon_{1-\alpha} \, ,$$

hipotezu H_0 odbacujemo.

Na isti način se testira pripadnost familiji neprekidnih raspodela (normalna, eksponencijalna...)- parametri se najpre ocene.

Pri testiranju neparametarskih hipoteza obično želimo da dokažemo H_0 !

Primer 174

Vek trajanja jedne komponente na uzorku od $n = 60$:

Vek trajanja (god.)	0.5	1	1.5	2	2.5	3
Broj komponenti	11	24	13	2	6	4

Testira se hipoteza da vreme trajanja ima eksponencijalnu raspodelu.
Ocena parametra: $\hat{\lambda} = 1/1.33 = 0.75$.
$H_0 : \ F \sim \mathrm{Exp}\,(0.75)$, $H_1 : \ F \not\sim \mathrm{Exp}\,(0.75)$.
Tražimo maksimalno odstupanje:

x	0.5	1	1.5	2	2.5	3
$F_n(x)$	0.183	0.583	0.800	0.833	0.889	1.000
$F(x)$	0.313	0.528	0.675	0.777	0.847	0.895
$\lvert F_n(x) - F(x)\rvert$	0.130	0.055	0.125	0.056	0.042	0.105

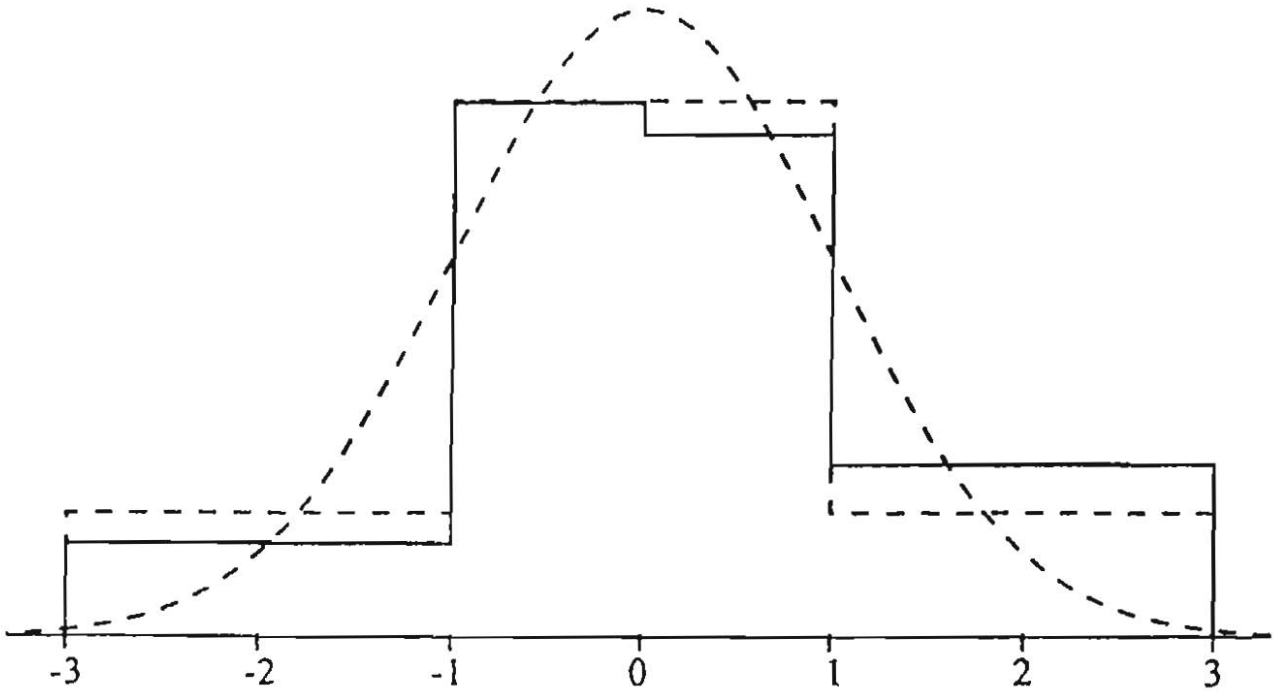
$K = \sqrt{60} \cdot 0.130 \approx 1$. p-vrednost: $P(K > 1) = 1 - K(1) = 0.27$. Ne odbacujemo H_0.

Poređenje histograma

Primer 170 Uzorak obima 50 - da li podaci dolaze iz $\mathcal{N}(0,1)$ raspodele?

Interval	Frekvencija	Relativna frekvencija	Verovatnoća za $\mathcal{N}(0,1)$
$(-\infty,-1)$	6	0.12	0.1587
$(-1,0)$	17	0.34	0.3413
$(0,1)$	16	0.32	0.3413
$(1,+\infty)$	11	0.22	0.1587

Poređenje histograma



Slika 36. Histogram na bazi podataka (puna linija) u poređenju sa histogramom standardne normalne raspodele i gustinom raspodele $\mathcal{N}(0,1)$

Hi kvadrat test

$(X_1, \ldots, X_n)$ nezavisan uzorak iz nepoznate raspodele sa funkcijom raspodele F. Da li je $F = F_0$? (hipoteza H_0).

1. *Podelimo uzorak na r intervala:*

$$(-\infty, a_1], (a_1, a_2], \ldots, (a_{r-2}, a_{r-1}], (a_{r-1}, +\infty)$$

2. *Očekivani broj podataka u i-tom intervalu pri hipotezi H_0 je np_{i0}, gde je $p_{i0} = F(a_i) - F(a_{i-1})$.*
3. *Stvarni broj podataka u i-tom intervalu: N_i.*
4. *Statistika testa-Pearsonova Hi kvadrat statistika:*

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

ima $\chi^2(r-1)$ raspodelu kad $n \to +\infty$ (pod hipotezom).

5. *Oblast odbacivanja: $\chi^2 > c$, c se bira prema nivou značajnosti.*

Hi kvadrat test

$(X_1, \ldots, X_n)$ nezavisan uzorak iz nepoznate raspodele sa funkcijom raspodele F. Da li je $F = F_0$? (hipoteza H_0).

1. *Podelimo uzorak na r intervala:*

$$(-\infty, a_1], (a_1, a_2], \ldots, (a_{r-2}, a_{r-1}], (a_{r-1}, +\infty)$$

2. *Očekivani broj podataka u i-tom intervalu pri hipotezi H_0 je np_{i0}, gde je $p_{i0} = F(a_i) - F(a_{i-1})$.*
3. *Stvarni broj podataka u i-tom intervalu: N_i.*
4. *Statistika testa-Pearsonova Hi kvadrat statistika:*

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

ima $\chi^2(r-1)$ raspodelu kad $n \to +\infty$ (pod hipotezom).

5. *Oblast odbacivanja: $\chi^2 > c$, c se bira prema nivou značajnosti.*

Hi kvadrat test

$(X_1, \ldots, X_n)$ nezavisan uzorak iz nepoznate raspodele sa funkcijom raspodele F. Da li je $F = F_0$? (hipoteza H_0).

1. *Podelimo uzorak na r intervala:*

$$(-\infty, a_1], (a_1, a_2], \ldots, (a_{r-2}, a_{r-1}], (a_{r-1}, +\infty)$$

2. *Očekivani broj podataka u i-tom intervalu pri hipotezi H_0 je np_{i0}, gde je $p_{i0} = F(a_i) - F(a_{i-1})$.*
3. *Stvarni broj podataka u i-tom intervalu: N_i.*
4. *Statistika testa-Pearsonova Hi kvadrat statistika:*

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

ima $\chi^2(r-1)$ raspodelu kad $n \to +\infty$ (pod hipotezom).

5. *Oblast odbacivanja: $\chi^2 > c$, c se bira prema nivou značajnosti.*

Hi kvadrat test

$(X_1, \ldots, X_n)$ nezavisan uzorak iz nepoznate raspodele sa funkcijom raspodele F. Da li je $F = F_0$? (hipoteza H_0).

1. *Podelimo uzorak na r intervala:*

$$(-\infty, a_1], (a_1, a_2], \ldots, (a_{r-2}, a_{r-1}], (a_{r-1}, +\infty)$$

2. *Očekivani broj podataka u i-tom intervalu pri hipotezi H_0 je np_{i0}, gde je $p_{i0} = F(a_i) - F(a_{i-1})$.*
3. *Stvarni broj podataka u i-tom intervalu: N_i.*
4. *Statistika testa-Pearsonova Hi kvadrat statistika:*

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

ima $\chi^2(r-1)$ raspodelu kad $n \to +\infty$ (pod hipotezom).

5. *Oblast odbacivanja: $\chi^2 > c$, c se bira prema nivou značajnosti.*

Hi kvadrat test

$(X_1, \ldots, X_n)$ nezavisan uzorak iz nepoznate raspodele sa funkcijom raspodele F. Da li je $F = F_0$? (hipoteza H_0).

1. *Podelimo uzorak na r intervala:*

$$(-\infty, a_1], (a_1, a_2], \ldots, (a_{r-2}, a_{r-1}], (a_{r-1}, +\infty)$$

2. *Očekivani broj podataka u i-tom intervalu pri hipotezi H_0 je np_{i0}, gde je $p_{i0} = F(a_i) - F(a_{i-1})$.*
3. *Stvarni broj podataka u i-tom intervalu: N_i.*
4. *Statistika testa-Pearsonova Hi kvadrat statistika:*

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

ima $\chi^2(r-1)$ raspodelu kad $n \to +\infty$ (pod hipotezom).

5. *Oblast odbacivanja: $\chi^2 > c$, c se bira prema nivou značajnosti.*

Hi kvadrat test

$(X_1, \ldots, X_n)$ nezavisan uzorak iz nepoznate raspodele sa funkcijom raspodele F. Da li je $F = F_0$? (hipoteza H_0).

1. *Podelimo uzorak na r intervala:*

$$(-\infty, a_1], (a_1, a_2], \ldots, (a_{r-2}, a_{r-1}], (a_{r-1}, +\infty)$$

2. *Očekivani broj podataka u i-tom intervalu pri hipotezi H_0 je np_{i0}, gde je $p_{i0} = F(a_i) - F(a_{i-1})$.*
3. *Stvarni broj podataka u i-tom intervalu: N_i.*
4. *Statistika testa-Pearsonova Hi kvadrat statistika:*

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

ima $\chi^2(r-1)$ raspodelu kad $n \to +\infty$ (pod hipotezom).

5. *Oblast odbacivanja: $\chi^2 > c$, c se bira prema nivou značajnosti.*

Kvantili ε_u hi kvadrat raspodele $\chi^2(n)$

	u							
n	0.005	0.01	0.025	0.05	0.95	0.975	0.99	0.995
1	0.00004	0.00016	0.00098	0.00393	3.841	5.024	6.635	7.879
2	0.010	0.0201	0.0506	0.103	5.991	7.378	9.210	10.59
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.83
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.86
5	0.412	0.554	0.831	1.145	11.070	12.832	13.086	16.75
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.54
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.27
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.95
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.58
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.18
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.75
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.30
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.81
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.31
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.80

Primer 170-nastavak

Imamo 4 klase $\Longrightarrow$ $\chi^2(3)$. Sa nivoom značajnosti $\alpha = 0.05$, nalazimo c iz $P(\chi^2 > c) = 0.05$, tj $c = \varepsilon_{0.95} = 7.815$ (iz tablice χ^2 kvantila na strani 326, sa $n = 3$).
Iz tablice u primeru 170 nalazimo:

$$N_1 = 6, N_2 = 17, N_3 = 16, N_4 = 11, \quad p_{10} = p_{40} = 0.1587, p_{20} = p_{30} = 0.341$$

$$\chi^2 = \frac{(6 - 50 \cdot 0.1587)^2}{50 \cdot 0.1587} + \cdots = 1.723$$

Dobijena vrednost je manja od kritične, tako da hipotezu H_0 ne odbacujemo. (to ne znači da smo je dokazali $\surd$)
Značajnost (izračunata pomoću softvera) je 0.631, što znači da sa datim uzorkom ne bismo odbacili nultu hipotezu ni na jednom od razumnih nivoa značajnosti!

Formulacija Hi kvadrat testa u parametarskom obliku

Test hipoteze $F = F_0$ *svodi se na testiranja hipoteze*

$$H_0 : \ p_1 = p_{10}, \ldots, p_r = p_{r0} \, ,$$

protiv komplementarne alternativne hipoteze

$$H_1 : \ (p_1, \ldots, p_r) \neq (p_{10}, \ldots, p_{r0})$$

Broj stepeni slobode je $r - 1$ *jer je* $\sum p_i = \sum p_{i0} = 1$

Tri pitanja

Koliki broj klasa treba uzeti?

$$r = 1 + \log_2 n = 1 + 3.3 \log_{10} n$$

(Sturges 1926, za normalnu raspodelu, neobavezno pravilo)

Koliko velike klase treba da budu ?

Klasa po pravilu treba biti takva da je očekivani broj podataka u klasi ≥ 5 *. Klase sa* $np_{j0} < 5$ *se spajaju sa susednom klasom ili više njih.*

Da li je oblast odbacivanja uvek oblika $\chi^2 > c$ *?*

Ako su podaci dobijeni iz generatora slučajnih brojeva, za oblast odbacivanja uzima se $\chi^2 < c_1 \vee \chi^2 > c_2$*. Zašto?*

Tri pitanja

Koliki broj klasa treba uzeti?

$$r = 1 + \log_2 n = 1 + 3.3 \log_{10} n$$

(Sturges 1926, za normalnu raspodelu, neobavezno pravilo)

Koliko velike klase treba da budu ?

Klasa po pravilu treba biti takva da je očekivani broj podataka u klasi ≥ 5 *. Klase sa* $np_{j0} < 5$ *se spajaju sa susednom klasom ili više njih.*

Da li je oblast odbacivanja uvek oblika $\chi^2 > c$ *?*

Ako su podaci dobijeni iz generatora slučajnih brojeva, za oblast odbacivanja uzima se $\chi^2 < c_1 \vee \chi^2 > c_2$*. Zašto?*

Tri pitanja

Koliki broj klasa treba uzeti?

$$r = 1 + \log_2 n = 1 + 3.3 \log_{10} n$$

(Sturges 1926, za normalnu raspodelu, neobavezno pravilo)

Koliko velike klase treba da budu ?

Klasa po pravilu treba biti takva da je očekivani broj podataka u klasi ≥ 5 *. Klase sa* $np_{j0} < 5$ *se spajaju sa susednom klasom ili više njih.*

Da li je oblast odbacivanja uvek oblika $\chi^2 > c$ *?*

Ako su podaci dobijeni iz generatora slučajnih brojeva, za oblast odbacivanja uzima se $\chi^2 < c_1 \vee \chi^2 > c_2$*. Zašto?*

Hi kvadrat sa neodređenim parametrima

H_0: podaci se mogu modelovati normalnom raspodelom (bez specifikacije μ i σ^2).

- Ako je nepoznat parametar θ dimenzije k, statistika testa je

$$\chi^2 = \chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})},$$

- Pod hipotezom H_0, statistika χ^2 ima $\chi^2(r-1-k)$ raspodelu za veliko n.
- Teoretski važi samo ako se parametri ocene metodom maksimalne verodostojnosti na osnovu grupisanih podataka, $\hat{\theta} = \theta$ za koje je

$$p_1(\theta)p_2(\theta)\dots p_{r-1}(\theta) \to \max,$$

Težak problem! - obično radimo sa ocenama iz originalnih podataka, razlike nisu velike. Videti primere 173 i 218 za poređenje.

Hi kvadrat sa neodređenim parametrima

H_0: podaci se mogu modelovati normalnom raspodelom (bez specifikacije μ i σ^2).

- Ako je nepoznat parametar θ dimenzije k, statistika testa je

$$\chi^2 = \chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})},$$

- Pod hipotezom H_0, statistika χ^2 ima $\chi^2(r-1-k)$ raspodelu za veliko n.
- Teoretski važi samo ako se parametri ocene metodom maksimalne verodostojnosti na osnovu grupisanih podataka, $\hat{\theta} = \theta$ za koje je

$$p_1(\theta)p_2(\theta)\dots p_{r-1}(\theta) \to \max,$$

Težak problem! - obično radimo sa ocenama iz originalnih podataka, razlike nisu velike. Videti primere 173 i 218 za poređenje.

Hi kvadrat sa neodređenim parametrima

H_0: podaci se mogu modelovati normalnom raspodelom (bez specifikacije μ i σ^2).

- Ako je nepoznat parametar θ dimenzije k, statistika testa je

$$\chi^2 = \chi^2 = \sum_{i=1}^{r} \frac{(N_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})},$$

- Pod hipotezom H_0, statistika χ^2 ima $\chi^2(r-1-k)$ raspodelu za veliko n.
- Teoretski važi samo ako se parametri ocene metodom maksimalne verodostojnosti na osnovu grupisanih podataka, $\hat{\theta} = \theta$ za koje je

$$p_1(\theta)p_2(\theta)\dots p_{r-1}(\theta) \to \max,$$

Težak problem! - obično radimo sa ocenama iz originalnih podataka, razlike nisu velike. Videti primere 173 i 218 za poređenje.

Hi kvadrat - univerzalni test saglasnosti

$$\chi^2 = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

- *Testovi saglasnosti (goodness of fit).*
- *Imamo rezultate n nezavisnih eksperimenata koji su izvedeni pod istim uslovima*
- *Neparametarski testovi u kojima se ne testira raspodela, već neke osobine događaja.*

Primer 175 (Dr.Arbuthnot)

	A	B
N_j	82	0
np_{j0}	41	41

p-vrednost $1.36 \cdot 10^{-19}$

Hi kvadrat - univerzalni test saglasnosti

$$\chi^2 = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

- *Testovi saglasnosti (goodness of fit).*
- *Imamo rezultate n nezavisnih eksperimenata koji su izvedeni pod istim uslovima*
- *Neparametarski testovi u kojima se ne testira raspodela, već neke osobine događaja.*

Primer 175 (Dr.Arbuthnot)

	A	B
N_j	82	0
np_{j0}	41	41

p-vrednost $1.36 \cdot 10^{-19}$

Hi kvadrat - univerzalni test saglasnosti

$$\chi^2 = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

- *Testovi saglasnosti (goodness of fit).*
- *Imamo rezultate n nezavisnih eksperimenata koji su izvedeni pod istim uslovima*
- *Neparametarski testovi u kojima se ne testira raspodela, već neke osobine događaja.*

Primer 175 (Dr.Arbuthnot)

	A	B
N_j	82	0
np_{j0}	41	41

p-vrednost $1.36 \cdot 10^{-19}$

Hi kvadrat - univerzalni test saglasnosti

$$\chi^2 = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

- *Testovi saglasnosti (goodness of fit).*
- *Imamo rezultate n nezavisnih eksperimenata koji su izvedeni pod istim uslovima*
- *Neparametarski testovi u kojima se ne testira raspodela, već neke osobine događaja.*

Primer 175 (Dr.Arbuthnot)

	A	B
N_j	82	0
np_{j0}	41	41

p-vrednost $1.36 \cdot 10^{-19}$

Hi kvadrat - univerzalni test saglasnosti

$$\chi^2 = \sum_{i=1}^{r} \frac{(\text{stvarno} - \text{očekivano})^2}{\text{očekivano}}$$

- *Testovi saglasnosti (goodness of fit).*
- *Imamo rezultate n nezavisnih eksperimenata koji su izvedeni pod istim uslovima*
- *Neparametarski testovi u kojima se ne testira raspodela, već neke osobine događaja.*

Primer 175 (Dr.Arbuthnot)

	A	B
N_j	82	0
np_{j0}	41	41

p-vrednost $1.36 \cdot 10^{-19}$

Testiranje nezavisnosti

Postavka: n eksperimenata, u svakom se dogodi po jedan i samo jedan od događaja A_i i isto za B_j. Tablica kontingencije:

	B_1	B_2	$\cdots$	B_k	Ukupno
A_1	f_{11}	f_{12}	$\cdots$	f_{1k}	a_1
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_v	f_{v1}	f_{v2}	$\cdots$	f_{vk}	a_v
Ukupno	b_1	b_2	$\cdots$	b_k	n

Hi kvadrat test nezavisnosti

Oznake: $p_{ij} = P(A_i B_j)$, $\alpha_i = P(A_i)$, $\beta_j = P(B_j)$.

$$H_0 : \ p_{ij} = \alpha_i \cdot \beta_j, \qquad i = 1, \dots v; \ j = 1 \dots k.$$

- Nepoznati parametri koje treba oceniti: α_i, β_j (ukupno $(v-1)+(k-1)$).
- $\hat{\alpha}_i = \frac{a_i}{n}$, $\hat{\beta}_j = \frac{b_j}{n}$
- Očekivan broj u klasi (i,j): $n \cdot \hat{\alpha}_i \cdot \hat{\beta}_j = \frac{a_i b_j}{n}$.
- Stvarni broj u klasi (i,j): f_{ij}.
- Broj klasa: $r = vk$.
- Stepeni slobode $= vk - 1 - (v-1) - (k-1) = (v-1)(k-1)$.

Hi kvadrat test nezavisnosti

Oznake: $p_{ij} = P(A_i B_j)$, $\alpha_i = P(A_i)$, $\beta_j = P(B_j)$.

$$H_0 : \ p_{ij} = \alpha_i \cdot \beta_j, \qquad i = 1, \dots v; \ j = 1 \dots k.$$

- Nepoznati parametri koje treba oceniti: α_i, β_j (ukupno $(v-1)+(k-1)$).
- $\hat{\alpha}_i = \frac{a_i}{n}$, $\hat{\beta}_j = \frac{b_j}{n}$
- Očekivan broj u klasi (i,j): $n \cdot \hat{\alpha}_i \cdot \hat{\beta}_j = \frac{a_i b_j}{n}$.
- Stvarni broj u klasi (i,j): f_{ij}.
- Broj klasa: $r = vk$.
- Stepeni slobode $= vk - 1 - (v-1) - (k-1) = (v-1)(k-1)$.

Hi kvadrat test nezavisnosti

Oznake: $p_{ij} = P(A_i B_j)$, $\alpha_i = P(A_i)$, $\beta_j = P(B_j)$.

$$H_0 : \ p_{ij} = \alpha_i \cdot \beta_j, \qquad i = 1, \dots v; \ j = 1 \dots k.$$

- Nepoznati parametri koje treba oceniti: α_i, β_j (ukupno $(v-1)+(k-1)$).
- $\hat{\alpha}_i = \frac{a_i}{n}$, $\hat{\beta}_j = \frac{b_j}{n}$
- Očekivan broj u klasi (i,j): $n \cdot \hat{\alpha}_i \cdot \hat{\beta}_j = \frac{a_i b_j}{n}$.
- Stvarni broj u klasi (i,j): f_{ij}.
- Broj klasa: $r = vk$.
- Stepeni slobode $= vk - 1 - (v-1) - (k-1) = (v-1)(k-1)$.

Hi kvadrat test nezavisnosti

Oznake: $p_{ij} = P(A_i B_j)$, $\alpha_i = P(A_i)$, $\beta_j = P(B_j)$.

$$H_0 : \ p_{ij} = \alpha_i \cdot \beta_j, \qquad i = 1, \dots v; \ j = 1 \dots k.$$

- Nepoznati parametri koje treba oceniti: α_i, β_j (ukupno $(v-1)+(k-1)$).
- $\hat{\alpha}_i = \frac{a_i}{n}$, $\hat{\beta}_j = \frac{b_j}{n}$
- Očekivan broj u klasi (i,j): $n \cdot \hat{\alpha}_i \cdot \hat{\beta}_j = \frac{a_i b_j}{n}$.
- Stvarni broj u klasi (i,j): f_{ij}.
- Broj klasa: $r = vk$.
- Stepeni slobode $= vk - 1 - (v-1) - (k-1) = (v-1)(k-1)$.

Hi kvadrat test nezavisnosti

Oznake: $p_{ij} = P(A_i B_j)$, $\alpha_i = P(A_i)$, $\beta_j = P(B_j)$.

$$H_0 : \ p_{ij} = \alpha_i \cdot \beta_j, \qquad i = 1, \dots v; \ j = 1 \dots k.$$

- Nepoznati parametri koje treba oceniti: α_i, β_j (ukupno $(v-1)+(k-1)$).
- $\hat{\alpha}_i = \frac{a_i}{n}$, $\hat{\beta}_j = \frac{b_j}{n}$
- Očekivan broj u klasi (i,j): $n \cdot \hat{\alpha}_i \cdot \hat{\beta}_j = \frac{a_i b_j}{n}$.
- Stvarni broj u klasi (i,j): f_{ij}.
- Broj klasa: $r = vk$.
- Stepeni slobode $= vk - 1 - (v-1) - (k-1) = (v-1)(k-1)$.

Hi kvadrat test nezavisnosti

Oznake: $p_{ij} = P(A_i B_j)$, $\alpha_i = P(A_i)$, $\beta_j = P(B_j)$.

$$H_0 : \ p_{ij} = \alpha_i \cdot \beta_j, \qquad i = 1, \dots v; \ j = 1 \dots k.$$

- Nepoznati parametri koje treba oceniti: α_i, β_j (ukupno $(v-1)+(k-1)$).
- $\hat{\alpha}_i = \frac{a_i}{n}$, $\hat{\beta}_j = \frac{b_j}{n}$
- Očekivan broj u klasi (i,j): $n \cdot \hat{\alpha}_i \cdot \hat{\beta}_j = \frac{a_i b_j}{n}$.
- Stvarni broj u klasi (i,j): f_{ij}.
- Broj klasa: $r = vk$.
- Stepeni slobode $= vk - 1 - (v-1) - (k-1) = (v-1)(k-1)$.

Hi kvadrat test nezavisnosti - nastavak

Test sa nivoom značajnosti α hipoteze H_0 o nezavisnosti događaja A_i i B_j $(i = 1, \ldots, v; j = 1, \ldots, k)$:
Ako je vrednost statistike

$$\chi^2 = \sum_{i=1}^{v} \sum_{j=1}^{k} \frac{\left(f_{ij} - \frac{a_i b_j}{n}\right)^2}{\frac{a_i b_j}{n}} = \sum_{i=1}^{v} \sum_{j=1}^{k} \frac{(n f_{ij} - a_i b_j)^2}{n a_i b_j}$$

veća od kvantila reda $1 - \alpha$ raspodele $\chi^2((v-1)(k-1))$, hipoteza H_0 se odbacuje.

Za vežbu: Primeri: 171, 176 Zadaci: 175-183