BIOSTATISTIKA (studij Nutricionizam)
(formule za drugi parcijalni ispit)

Poissonova slučajna varijabla

Slučajna varijabla X ima **Poissonovu razdiobu** ili **distribuciju** s parametrom $\lambda > 0$ ako je funkcija gustoće te slučajne varijable zadana s:

$$p_X(k) = P(X = k) = \frac{\lambda^k}{k!}\, e^{-\lambda}, \quad k = 0, 1, 2, 3, \ldots$$

$X \sim P(\lambda)$ i $E[X] = \lambda \quad \text{Var}[X] = \lambda$.

Hipergeometrijska slučajna varijabla

Slučajna varijabla X ima **hipergeometrijsku razdiobu** ili **distribuciju** ako je funkcija gustoće te slučajne varijable zadana s:

$$p_X(k) = P(X = k) = \frac{\binom{m}{k}\binom{n}{s-k}}{\binom{m+n}{s}}, \quad \max(0, s-n) \leq k \leq \min(m, s)$$

Pišemo $X \sim H(n, m, s)$.

- Očekivanje hipergeometrijske razdiobe: $E[X] = \dfrac{s \cdot m}{m+n}$
- Varijanca hipergeometrijske razdiobe: $V[X] = \dfrac{m \cdot n \cdot s \cdot (m+n-s)}{(m+n)^2(m+n-1)}$

Neprekidne slučajne varijable

Za slučajnu varijablu X kažemo da je **neprekidna** ako vrijedi sljedeće:

(i) ImX je interval u $\mathbb{R}$

(ii) postoji nenegativna funkcija $f_X : \mathbb{R} \to \mathbb{R}$ tako da za svaka dva broja a, b $(a < b)$ vrijedi

$$P(a \leq X \leq b) = \int_a^b f_X(t)dt$$

Funkciju f_X zovemo **funkcija gustoće** od X. **Funkcija distribucije** F_X od X definirana je s:

$$F_X(x) := P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

Vrijedi:

$$P(a \leq X \leq b) = F_X(b) - F_X(a)$$

(1) Za svaki broj $a \in \mathbb{R}$ je

$$P(X = a) = \lim_{b \to a} P(a \leq X \leq b) = \lim_{b \to a} \int_a^b f_X(t)dt = \int_a^a f_X(t)dt = 0$$

(2)

$$\int_{-\infty}^{\infty} f_X(t)dt = P(-\infty < X < \infty) = 1$$

Matematičko očekivanje od X definirano je s:

$$E[X] = \int_{-\infty}^{\infty} t \cdot f_X(t)dt,$$

a **varijanca**

$$V[X] = E[X^2] - (E[X])^2$$

gdje je sada

$$E[X^2] = \int_{-\infty}^{\infty} t^2 \cdot f_X(t)dt.$$

Općenito, za $g : \mathbb{R} \to \mathbb{R}$

$$E[g(X)] = \int_{-\infty}^{\infty} g(t) \cdot f_X(t)dt.$$

Normalna slučajna varijabla

Kažemo da neprekidna slučajna varijabla X ima **normalnu razdiobu** s parametrima μ i σ^2 ako joj je funkcija gustoće zadana s:

$$f_X(t) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

Oznaka:

$$X \sim N(\mu, \sigma^2)$$

Vrijedi:

1. $f_X(t) > 0, \ \ \forall t \in \mathbb{R} \ \ \Rightarrow \ \mathrm{Im}X = \mathbb{R}$
2. $E[X] = \mu$
3. $V[X] = \sigma^2$

Standardizirana slučajna varijabla:

$$X^* := \frac{X - E[X]}{\sigma_X} = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Funkciju distribucije jedinične normalne razdiobe $N(0, 1)$:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt, \quad x \in \mathbb{R}$$

Tabelirana funkcija:

$$\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}}\, dt, \quad x > 0$$

$$\Phi(x) = \frac{1}{2} + \Phi_0(x) \text{ i } \Phi_0(x) = -\Phi_0(-x) \quad \text{za} \quad x < 0.$$

Procjena parametara

$$\overline{X}_n = \frac{1}{n}\left(X_1 + \cdots + X_n\right)$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right]$$

$$\Phi_0(z_{\frac{\alpha}{2}}) = \frac{1-\alpha}{2}$$

$(1-\alpha)\cdot 100\%$ pouzdan interval za očekivanje normalne populacije (varijanca poznata)

$$\bar{X}_n - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$(1-\alpha)\cdot 100\%$ pouzdan interval za očekivanje normalne populacije (varijanca nepoznata)

$$\bar{X}_n - t_{\frac{\alpha}{2}}(n-1) \cdot \frac{S_n}{\sqrt{n}} \le \mu \le \bar{X}_n + t_{\frac{\alpha}{2}}(n-1) \cdot \frac{S_n}{\sqrt{n}}$$

$(1-\alpha)\cdot 100\%$ pouzdan interval za očekivanje populacije na osnovi velikih uzoraka

$$\bar{X}_n - z_{\frac{\alpha}{2}} \cdot \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}} \cdot \frac{S_n}{\sqrt{n}}$$

Pouzdan interval za parametar p binomne razdiobe:

$$\hat{P} = \frac{X}{n} = \bar{X}$$

$$\bar{X} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

Testiranje statističkih hipoteza

Test o očekivanju normalno distribuirane populacije

Varijanca poznata:

$$Z = \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

1.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

ako je $Z < -z_{\frac{\alpha}{2}}$ ili $Z > z_{\frac{\alpha}{2}} \Rightarrow$ odbacujemo H_0
Ako je $-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}} \Rightarrow$ ne možemo odbaciti H_0

2.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

H_0 odbacujemo ako je $Z > z_\alpha$

3.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

H_0 odbacujemo ako je $Z < -z_\alpha$

Varijanca nepoznata:

$$T = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$$

1.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Nultu hipotezu H_0 odbacujemo ako je

$$T > t_{\frac{\alpha}{2}}(n-1) \quad \text{ili} \quad T < -t_{\frac{\alpha}{2}}(n-1)$$

2.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

H_0 odbacujemo ako je

$$T > t_\alpha(n-1)$$

3.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

H_0 odbacujemo ako je

$$T < -t_\alpha(n-1)$$

Testovi o očekivanju na osnovi velikih uzoraka

Test o proporciji:

$$Z = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}, \quad \bar{X} = \hat{P}$$

1.

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

Nultu hipotezu H_0 odbacujemo ako je $\quad Z > z_{\frac{\alpha}{2}} \quad$ ili $\quad Z < -z_{\frac{\alpha}{2}}$

2.

$$H_0 : p = p_0$$
$$H_1 : p > p_0$$

H_0 odbacujemo ako je $\quad Z > z_\alpha$

3.

$$H_0 : p = p_0$$
$$H_1 : p < p_0$$

H_0 odbacujemo ako je $\quad Z < -z_\alpha$

Usporedba očekivanja dviju normalno distribuiranih populacija (t-test)

$$T = \frac{\bar{X_1} - \bar{X_2}}{S} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

gdje su

$$\bar{X_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{(1)}, \qquad \bar{X_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i^{(2)},$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left((n_1-1)S_1^2 + (n_2-1)S_2^2\right)$$

za $S_1^2,\ S_2^2$ uzoračke varijance uzoraka 1 i 2.

1.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Nultu hipotezu H_0 odbacujemo ako

$$T > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \quad \text{ili} \quad T < -t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$$

2.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Nultu hipotezu H_0 odbacujemo ako

$$T > t_\alpha(n_1 + n_2 - 2)$$

3.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Nultu hipotezu H_0 odbacujemo ako

$$T < -t_\alpha(n_1 + n_2 - 2)$$

Usporedba proporcija

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$
$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

1.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

Nultu hipotezu H_0 odbacujemo ako

$$Z > z_{\frac{\alpha}{2}} \quad \text{ili} \quad Z < -z_{\frac{\alpha}{2}}$$

2.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 > p_2$$

Nultu hipotezu H_0 odbacujemo ako

$$Z > z_\alpha$$

3.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 < p_2$$

Nultu hipotezu H_0 odbacujemo ako

$$Z < -z_\alpha$$

Usporedba varijanci dviju normalno distribuiranih populacija (F-test)

$$F = \frac{S_1^2}{S_2^2}$$

1.

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Nultu hipotezu H_0 odbacujemo ako

$$F > f_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \quad \text{ili} \quad F < f_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$$

2.

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

Nultu hipotezu H_0 odbacujemo ako

$$F > f_\alpha(n_1 - 1, n_2 - 1)$$

3.

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2$$

Nultu hipotezu H_0 odbacujemo ako

$$F < f_{1-\alpha}(n_1 - 1, n_2 - 1)$$

χ^2 - test o prilagodbi modela podacima

$$H = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'}$$

gdje su f_i eksperimentalne, a $f_i' = np_i$ teorijske frekvencije

k = (konačan) broj razreda u tablici
r = broj nepoznatih parametara

nultu hipotezu da se radi o određenoj razdiobi odbacujemo ako

$$H \geq \chi_\alpha^2(k - r - 1)$$

χ^2 - test nezavisnosti dviju varijabli

Kontingencijska frekvencijska tablica:

$X \backslash Y$	b_1	b_2	$\dots$	b_s	Σ
a_1	f_{11}	f_{12}	$\dots$	f_{1s}	f_1
a_2	f_{21}	f_{22}	$\dots$	f_{2s}	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	f_{r1}	f_{r2}	$\dots$	f_{rs}	f_r
Σ	g_1	g_2	$\dots$	g_s	n

$$p_{ij} = P(X = a_i,\ Y = b_j)$$
$$p_i = P(X = a_i)$$
$$q_j = P(X = b_j)$$

H_0 : $p_{ij} = p_i \cdot q_j, \quad \forall\, i, j$
tj. X i Y su nezavisne slučajne varijable
H_1 : $\exists\, i, j$ takvi da $p_{ij} \neq p_i \cdot q_j$

$$\hat{p_i} = \frac{f_i}{n}, \quad \hat{q_j} = \frac{g_j}{n}$$

$$f_{ij}' = n\, \hat{p_i}\, \hat{q_j} = n \cdot \frac{f_i}{n} \cdot \frac{g_j}{n} = \frac{f_i \cdot g_j}{n}$$

$$H = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - f_{ij}')^2}{f_{ij}'}$$

Hipotezu o nezavisnosti odbacujemo ako

$$H \geq \chi^2_\alpha((r-1)(s-1))$$

χ^2 - test homogenosti populacija

$$H_0 \ : \quad X^{(1)} \stackrel{D}{=} X^{(2)} \stackrel{D}{=} \ldots \stackrel{D}{=} X^{(m)}$$
$$H_1 \ : \quad \exists\, i,j \quad \text{tako da} \quad X^{(i)} \stackrel{D}{\neq} X^{(j)}$$

Frekvencijska tablica:

X	a_1	a_2	$\ldots$	a_k	$\sum$
populacija 1	f_{11}	f_{12}	$\ldots$	f_{1k}	n_1
populacija 2	f_{21}	f_{22}	$\ldots$	f_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
populacija m	f_{m1}	f_{m2}	$\ldots$	f_{mk}	n_m
$\sum$	f_1	f_2	$\ldots$	f_k	n

$$\hat{p}_j = \frac{f_j}{n}, \quad j = 1,\ldots,k, \quad f'_{ij} = n_i \cdot \hat{p}_j = \frac{n_i \cdot f_j}{n}$$

$$H = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

hipotezu o homogenosti populacija odbacujemo ako

$$H \geq \chi^2_\alpha((m-1)(k-1))$$

Usporedba očekivanja više normalno distribuiranih populacija (jednofaktorska analiza varijance ANOVA)

$$H_0 \ : \ \mu_1 = \mu_2 = \ldots = \mu_k,$$

$$\bar{X}_i = \frac{1}{n_i}(X_{i1} + \ldots + X_{in_i}), \ S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_i, \quad n = \sum_{i=1}^{k} n_i$$

$$SST = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{k} n_i \bar{X}_i^2 - n\bar{X}^2, \ SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k} n_i \bar{X}_i^2$$

$$MST = \frac{SST}{k-1}, \quad MSE = \frac{SSE}{n-k}$$

$$F = \frac{MST}{MSE}$$

nultu hipotezu odbacujemo ako

$$F \geq f_\alpha(k-1, n-k)$$

ANOVA tablica:

izvor rasipanja	stupnjevi slobode	suma kvadrata	srednjekvadratno odstupanje	vrijednost test-statistike
zbog razlike među tretmanima	$k-1$	SST	MST	F
zbog greške	$n-k$	SSE	MSE	
$\sum$	$n-1$	SS		

Test koreliranosti dviju varijabli

$$S_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right)$$

$$R = \frac{S_{xy}}{S_x \cdot S_y}, \quad Z = \frac{R}{\sqrt{1-R^2}} \cdot \sqrt{n-2}$$

1.

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Nultu hipotezu H_0 odbacujemo ako je

$$Z > t_{\frac{\alpha}{2}}(n-2) \qquad \text{ili} \qquad Z < -t_{\frac{\alpha}{2}}(n-2)$$

2.

$$H_0 : \rho = 0$$
$$H_1 : \rho > 0$$

H_0 odbacujemo ako je

$$Z > t_{\alpha}(n-2)$$

3.

$$H_0 : \rho = 0$$
$$H_1 : \rho < 0$$

H_0 odbacujemo ako je

$$Z < -t_{\alpha}(n-2)$$

Linearni regresijski model

$$\hat{\alpha} := \frac{S_{xy}}{S_x^2}$$
$$\hat{\beta} := \bar{y} - \hat{\alpha}\,\bar{x}$$

$$\hat{\sigma}^2 = \frac{SSE}{n-2}, \quad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta} - \hat{\alpha}\,x_i)^2 = S_{yy} - \hat{\alpha}^2 S_{xx}$$

$(1-\alpha)\cdot 100\%$ **pouzdan interval za** α:

$$\hat{\alpha} - t_{\frac{\alpha}{2}}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}} \;\leq\; \alpha \;\leq\; \hat{\alpha} + t_{\frac{\alpha}{2}}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}}$$

$(1-\alpha)\cdot 100\%$ **pouzdan interval za** β:

$$\hat{\beta} \;-\; t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)S_x^2}} \;\leq\; \beta \;\leq\; \hat{\beta} \;+\; t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)S_x^2}}$$

1. $H_0 \;:\; \alpha = \alpha_0 \;\; (\alpha_0 \in \mathbb{R})$ (u odnosu na razne alternative):

$$T_{\alpha} = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}}\,\sqrt{(n-1)S_x^2}$$

Ako je H_0 istinita tada je

$$T_{\alpha} \sim t(n-2)$$

2. $H_0 \ : \ \beta = \beta_0 \ \ (\beta_0 \in \mathbb{R})$ (u odnosu na razne alternative):

$$T_\beta = \frac{\hat{\beta} - \beta_0}{\hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}}}$$

Ako je H_0 istinita tada je

$$T_\beta \sim t(n-2)$$

$(1-\alpha)$ 100% pouzdan interval za $E[Y|x = x_0]$:

$$\left[\widehat{E[Y|x = x_0]} \ - \ t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \ , \right.$$
$$\left. \widehat{E[Y|x = x_0]} \ + \ t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \right]$$

$(1-\alpha)$ 100% pouzdan interval za Y u $x = x_0$:

$$\left[\hat{Y} \ - \ t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \ , \right.$$
$$\left. \hat{Y} \ + \ t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \right]$$

koeficijent determinacije

$$R^2 := \frac{(n-1)S_y^2 - SSE}{(n-1)S_y^2} = 1 - \frac{SSE}{(n-1)S_y^2} \ \in \ [0,1]$$