

# Statisztika

Nagy-György Judit

Szegedi Tudományegyetem, Bolyai Intézet

Tekintsünk egy eloszláscsaládot és legyen  $X$  véletlen változó eloszlása ebből az eloszláscsaládból ismeretlen. Paraméteres probléma esetén a paraméteres eloszláscsalád paramétere a  $\Theta \subset \mathbb{R}^k$  **paraméterter** eleme. A  $\theta \in \Theta$  paraméter értékét időnként kihangsúlyozzuk  $P_\theta$ ,  $E_\theta(X)$ ,  $D_\theta^2(X)$  ( $= \text{Var}_\theta(X)$ ) alsó indexében.

## Definíció

Az  $\mathbf{X} = (X_1, \dots, X_n)$  **statisztikai minta** elemei független véletlen változók, eloszlásuk megegyezik az  $X$  **háttérváltozó** eloszlásával. Egy véletlen kísérlet  $\mathbf{x} = \mathbf{X}(\omega)$  kimenetele a minta **realizációja**. A lehetséges realizációk  $\mathcal{X}$  halmaza a **mintatér**.

## A matematikai statisztika alapvető feladatai:

Célok: a háttérváltozó eloszlásának (paraméterének) **becslése**, illetve ezekre vonatkozó **hipotézisek vizsgálata**.

## Definíció

Tekintsük az  $X_1, \dots, X_n$  mintát. Az  $X_1^*, \dots, X_n^*$  az  $n$  elemű **rendezett minta**, ha minden  $\omega \in \Omega$ -ra  $X_1^*(\omega) \leq \dots \leq X_n^*(\omega)$  az  $X_1(\omega), \dots, X_n(\omega)$  realizáció permutációja.

## Definíció

Legyen  $T : \mathbb{R}^n \rightarrow \mathbb{R}^k$  mérhető függvény. A  $T(X_1, \dots, X_n)$  véletlen (vektor)változót **statisztikának** nevezzük.

## Alapstatisztikák

- ▶  $\bar{X} = \frac{X_1 + \dots + X_n}{n}$  mintaátlag,
- ▶  $X_1^*$  legkisebb,  $X_n^*$  legnagyobb mintaelem,
- ▶  $X_n^* - X_1^*$  mintaterjedelem,
- ▶ tapasztalati medián:  $X_{\frac{n+1}{2}}^*$  ha  $n$  ptl,  $\frac{X_{n/2}^* + X_{n/2+1}^*}{2}$  ha  $n$  ps,
- ▶  $(X_1^*, \dots, X_n^*)$  teljes rendezett minta.

Legyen  $\theta$  az  $X$  eloszlásának egy paramétere,  $\hat{\theta}_n := \hat{\theta}(X_1, \dots, X_n)$ .

## Definíció

Az  $\hat{\theta}_n$  statisztika **torzítatlan** becslése  $\theta$ -nak, ha a paraméterhalmaz minden  $\theta$  elemére  $E_{\theta}(\hat{\theta}_n) = \theta$ . A  $\hat{\theta}_n$  és  $\tilde{\theta}_n$  torzítatlan becslések közül  $\hat{\theta}_n$  **hatásosabb**, ha  $D_{\theta}^2(\hat{\theta}_n) < D_{\theta}^2(\tilde{\theta}_n)$  minden  $\theta \in \Theta$ -ra. A  $\hat{\theta}_n$  **hatásos** becslése  $\theta$ -nak, ha nincs nála hatásosabb becslése.

## Definíció

$\hat{\theta}_n$  ( $n = 1, 2, \dots$ ) sorozat **gyengén konzisztens** becslése  $\theta$ -nak, ha

$$\hat{\theta}_n \xrightarrow{\text{sz}} \theta, \quad n \rightarrow \infty,$$

azaz minden  $\varepsilon > 0, \theta \in \Theta$  esetén  $\lim_{n \rightarrow \infty} P_{\theta}(|\hat{\theta}_n - \theta| > \varepsilon) = 0$ .

$\hat{\theta}_n$  ( $n = 1, 2, \dots$ ) sorozat **erősen konzisztens** becslése  $\theta$ -nak, ha

$$\hat{\theta}_n \xrightarrow{\text{mb}} \theta, \quad n \rightarrow \infty,$$

azaz  $P_{\theta}(\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta) = 1$  minden  $\theta \in \Theta$ -ra.

## Definíció

A egy esemény, amire  $P(A) = p$ ,  $X = I(A) \sim \text{Bernoulli}(p) = \mathcal{B}_1(p)$  a hozzá tartozó indikátorváltozó. Végezzünk el függetlenül  $n$  kísérletet, tekintsük az  $X_1, \dots, X_n$  mintát. Ekkor

A **gyakorisága a mintában**  $\sum_{i=1}^n X_i$ , a bekövetkezéseinek száma,  
A **relatív gyakorisága**

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

## Állítás

$\hat{p}$  torzítatlan és erősen konzisztens becslés  $p$ -re.

## Bizonyítás

$$E(\hat{p}) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n E(X_i)}{n} = \frac{np}{n} = p.$$

A nagy számok erős törvénye miatt  $\hat{p} = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{mb} E(X) = p.$

## Definíció (empirikus eloszlásfüggvény)

Legyen  $K_{n,x} = \sum_{i=1}^n I(X_i \leq x)$ , ahol  $I$  indikátorfüggvény.

$$F_n^*(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n} = \begin{cases} 0, & x < X_1^* \\ \frac{k}{n}, & X_k^* \leq x < X_{k+1}^*, k = 1, \dots, n-1 \\ 1, & X_n^* \leq x \end{cases}$$

## Állítás

$E(F_n^*(x)) = F(x)$ ,  $D^2(F_n^*(x)) \rightarrow 0$ , ha  $n \rightarrow \infty$ ,

valamint  $F_n^*(x) \xrightarrow{mb} F(x)$ , ha  $n \rightarrow \infty$  minden  $x \in \mathbb{R}$  esetén.

## Bizonyítás.

$I(X_i \leq x) \sim \mathcal{B}_1(F(x))$ ,  $i = 1, \dots, n$  fgn vv-k,  $K_{n,x} \sim \mathcal{B}_n(F(x))$

$$E(F_n^*(x)) = \frac{1}{n} E(K_{n,x}) = \frac{1}{n} \cdot nF(x).$$

$$D^2(F_n^*(x)) = \frac{1}{n^2} D^2(K_{n,x}) = \frac{n}{n^2} F(x)(1 - F(x)) \rightarrow 0, \text{ ha } n \rightarrow \infty.$$

Az erős konzisztencia a nagy számok erős törvényéből következik.

## Tétel (Glivenko–Cantelli (A statisztika alaptétele))

$$\Delta_n = \sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)| \xrightarrow{mb} 0, \quad n \rightarrow \infty.$$

### Bizonyítás (folytonos eset)

Legyen  $m \in \mathbb{Z}^+$  tetszőleges,  $x_i = \inf\{x : F(x) = \frac{i}{m}\}$ .

( $x_0 = -\infty, x_m = \infty$  is lehet) Ha  $x_{i-1} < x < x_i$ , akkor

$$F_n^*(x) - F(x) \leq F_n^*(x_i) - F(x_{i-1}) = F_n^*(x_i) - F(x_i) + \frac{1}{m},$$

$$F_n^*(x) - F(x) \geq F_n^*(x_{i-1}) - F(x_i) = F_n^*(x_{i-1}) - F(x_{i-1}) - \frac{1}{m},$$

ebből

$$|F_n^*(x) - F(x)| \leq \max_{1 \leq i \leq m-1} |F_n^*(x_i) - F(x_i)| + \frac{1}{m}$$

$$\Rightarrow \Delta_n \leq \max_{1 \leq i \leq m-1} |F_n^*(x_i) - F(x_i)| + \frac{1}{m}$$

A nagy számok erős törvényéből  $\lim_{n \rightarrow \infty} F_n^*(x_i) \xrightarrow{mb} F(x_i)$

$$\Rightarrow P\left(\limsup_{n \rightarrow \infty} \Delta_n > \frac{1}{m}\right) = 0 \quad \forall m \in \mathbb{Z}^+ \Rightarrow P\left(\lim_{n \rightarrow \infty} \Delta_n = 0\right) = 1.$$

# A sűrűségfüggvény becslése

Az empirikus eloszlásfüggvény szakaszonként konstans függvény, deriváltja nem lesz jó becslése az  $f$  sűrűségfüggvénynek.

## Definíció

$I_1, I_2, \dots$  páronként diszjunkt  $h_n$  hosszú intervallumok,  
 $\bigcup_{i=1}^{\infty} I_i = \mathbb{R}$ . Legyen  $\nu_k = \sum_{i=1}^n I(X_i \in I_k)$ ,

$$f_n^*(x) = \frac{\nu_k}{nh_n}, \quad \text{ha } x \in I_k.$$

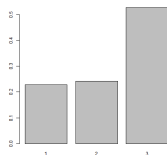
Az  $f_n$  függvényt *sűrűséghisztogram*nak nevezzük.

## Megjegyzések

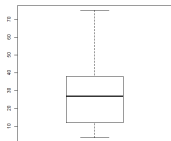
- ▶ Véges sok intervallumban vesz fel nem 0 értéket  $f_n^*(x)$ .
- ▶ Belátható, hogy ha  $x$  az  $f$  sűrűségfüggvény folytonossági pontja,  $\lim_{n \rightarrow \infty} h_n = 0$  és  $\lim_{n \rightarrow \infty} nh_n = \infty$ , akkor  $\lim_{n \rightarrow \infty} f_n^*(x) = f(x)$  1 valószínűséggel.

## Kör- és oszlopdiagram

Diszkrét eloszlás ábrázolására.



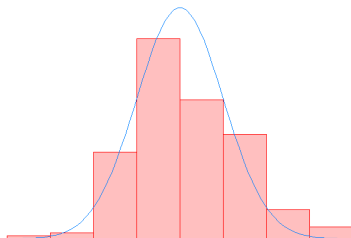
## Boxplot



Minimum,  
maximum, tapasztalati kvartilisek egy ábrán.

## Sűrűséghisztogram

A sűrűségfüggvény  
becslésének ábrázolása.



# A várható érték becslése

## Állítás

*Ha  $E(X)$  létezik,  $\bar{X}$  torzítatlan és erősen konzisztens becslése is  $E(X)$ -nek.*

## Bizonyítás

A várható érték linearitása miatt

$$E(\bar{X}) = E\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{E(X_1) + \dots + E(X_n)}{n} = \frac{1}{n}nE(X).$$

Az állítás második fele éppen a nagy számok erős törvénye.

## Kérdés

Mi lehet a magyarázata, hogy az emberek többségének fizetése az átlagfizetés alatt van?

# A variancia becslése

## Definíció

*Empirikus (tapasztalati) variancia:*  $S^2 = S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

$S = \sqrt{S^2}$  az empirikus (tapasztalati) szórás.

## Tétel (Steiner)

$$x_1, \dots, x_n, c \in \mathbb{R}. \quad \frac{1}{n} \sum_{i=1}^n (x_i - c)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + (\bar{x} - c)^2.$$

## Bizonyítás

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - c)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{2}{n} (\bar{x} - c) \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_0 + (\bar{x} - c)^2.$$

## Következmény

$$S^2 = \overline{X^2} - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - E(X))^2 - (\bar{X} - E(X))^2.$$

## Tétel

Ha  $D(X) < \infty$ , akkor  $E(S^2) = \frac{n-1}{n} D^2(X)$ ,  $S_n^2 \xrightarrow{mb} D^2(X)$ ,  $n \rightarrow \infty$ .

## Bizonyítás

$$\begin{aligned} E(S^2) &= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - E(X))^2 - (\bar{X} - E(X))^2\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i - E(X))^2 - E((\bar{X} - E(X))^2) \\ &= \frac{n}{n} D^2(X) - D^2\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \left(1 - \frac{n}{n^2}\right) D^2(X). \end{aligned}$$

$$\bar{X} \xrightarrow{mb} E(X) \Rightarrow \bar{X}^2 - \bar{X}^2 \xrightarrow{mb} E(X^2) - E^2(X), n \rightarrow \infty.$$

## Definíció

*Korrigált empirikus variancia és szórás:  $S^{*2} = \frac{n}{n-1} S^2$ ,  $S^* = \sqrt{S^{*2}}$ .*

## Következmény

$S^{*2}$  torzítatlan és erősen konzisztens becslése  $D^2(X)$ -nek.

## Definíció

$k \in \mathbb{Z}^+$  esetén az

- ▶  $M_k = \frac{1}{n} \sum_{i=1}^n X_i^k$  a *k. empirikus momentum*,
- ▶  $M_k^c = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$  a *k. empirikus centrális momentum*,
- ▶ az  $\frac{M_3^c}{(M_2^c)^{3/2}}$  statisztikát *ferdeségnek*,
- ▶ az  $\frac{M_4^c}{(M_2^c)^2} - 3$  statisztikát *lapultásnak* (csúcsosságnak) nevezzük.

## Megjegyzés

A nagy számok erős törvényéből következik a fenti becslések erős konzisztenciája.

# Kovariancia, korreláció becslése

Tekintsük  $(X, Y)$  háttérváltozót és  $(X_1, Y_1), \dots, (X_n, Y_n)$  mintát.

## Definíció

*Empirikus kovariancia:*  $\hat{C} = C_n(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ ,

## Állítás

$\hat{C} = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X} \cdot \bar{Y} \xrightarrow{mb} C(X, Y)$ , ha  $n \rightarrow \infty$ .

## Bizonyítás

$$\begin{aligned}\hat{C} &= \frac{1}{n} \left( \sum_{i=1}^n X_i Y_i - \bar{X} \sum_{i=1}^n Y_i - \bar{Y} \sum_{i=1}^n X_i + n\bar{X} \cdot \bar{Y} \right) \\ &= \overline{X \cdot Y} + (-2 + 1)\bar{X} \cdot \bar{Y}.\end{aligned}$$

$$Z = XY, \quad \frac{1}{n} \sum_{i=1}^n Z_i - \bar{X} \cdot \bar{Y} \xrightarrow{mb} E(Z) - E(X)E(Y) = C(X, Y).$$

## Tétel

Ha  $C(X, Y)$  létezik, akkor  $E(\hat{C}) = \frac{n-1}{n}C(X, Y)$ .

## Bizonyítás

$$\begin{aligned} E(\hat{C}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i Y_i - \left(\frac{1}{n} \sum_{i=1}^n X_i\right) \left(\frac{1}{n} \sum_{i=1}^n Y_i\right)\right) \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n E(X_i Y_i)}_{nE(XY)} - \frac{1}{n^2} \left( \underbrace{\sum_{i=1}^n E(X_i Y_i)}_{nE(XY)} + \underbrace{\sum_{i \neq j} E(X_i Y_j)}_{n(n-1)E(X)E(Y)} \right). \end{aligned}$$

## Definíció

*Korrigált empirikus kovariancia:*

$$\hat{C}^* = C_n^*(X, Y) = \frac{n}{n-1} \hat{C} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}).$$

## Következmény

$\hat{C}^*$  torzítatlan és erősen konzisztens becslése  $C(X, Y)$ -nak.

## Definíció

*(Pearson-féle) empirikus korrelációs együttható:*

$$\hat{r} = r_n(X, Y) = \frac{\hat{C}}{S_X S_Y}.$$

## Állítás

$$|\hat{r}| \leq 1.$$

## Bizonyítás

A Cauchy-Bunyakovszkij-Schwarz egyenlőtlenség alapján

$$\begin{aligned} |n\hat{C}| &\leq \sum_{i=1}^n |X_i - \bar{X}| \cdot |Y_i - \bar{Y}| \\ &\leq \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2} = \sqrt{n} S_X \sqrt{n} S_Y. \end{aligned}$$

## Megjegyzés

$$\frac{\hat{C}^*}{S_X^* S_Y^*} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2)^{1/2}} = \frac{\hat{C}}{S_X S_Y} = \hat{r}.$$

# Maximum likelihood becslés

$\theta$  az  $X$  változó ismeretlen paramétere, ezt szeretnénk becsülni.

## Definíció

Legyen  $X$  diszkrét háttérváltozó, tekintsük az  $x_1, \dots, x_n$  realizációt. A hozzá tartozó **likelihood-függvény** a következő:

$$L(\theta) = P_{\theta}(X_1 = x_1, \dots, X_n = x_n) = P_{\theta}(X = x_1) \cdot \dots \cdot P_{\theta}(X = x_n).$$

Legyen  $X$  abszolút folytonos háttérváltozó. Tekintsük a  $x_1, \dots, x_n$  realizációt. A hozzá tartozó **likelihood függvény** a következő:

$$L(\theta) = f_{\theta, x_1, \dots, x_n}(x_1, \dots, x_n) = f_{\theta}(x_1) \cdot \dots \cdot f_{\theta}(x_n).$$

## Definíció

A  $\theta$  paraméter **maximum likelihood (ML) becslése** a  $\hat{\theta}$  statisztika, ha minden  $x_1, \dots, x_n$  realizációra  $L(\hat{\theta}) = \max_{\theta} L(\theta, x_1, \dots, x_n)$ .

## Megjegyzés

Az  $\ell(\theta) = \ln L(\theta)$  **log-likelihood függvény** ugyanott veszi fel szélsőértékeit mint  $L(\theta)$ .

## Normális eloszlás paramétereinek ML becslése

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} = (\sqrt{2\pi}\sigma)^{-n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2},$$

ahol  $\mu \in \mathbb{R}$  és  $\sigma > 0$ .

$$\ell(\mu, \sigma^2) = -n \ln \sqrt{2\pi} - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2,$$

ami kétszer deriválható. Írjuk fel a likelihood egyenleteket:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \quad (1)$$

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu)^2 = 0, \quad (2)$$

(1)-ből kapjuk, hogy  $\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$ , ezt (2)-be helyettesítve pedig

$$\frac{n}{2\hat{\sigma}^2} = \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (x_i - \bar{x})^2$$
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s_n^2.$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{n}{\sigma^2}, \quad \frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} + \frac{-2}{2(\sigma^2)^3} \sum_{i=1}^n (x_i - \mu)^2,$$

$$\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} = \frac{-1}{(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu), \quad \frac{\partial^2 \ell}{\partial \sigma^2 \partial \mu} = \frac{2}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu).$$

Vizsgáljuk a Hesse-mátrixot  $(\hat{\mu}, \hat{\sigma}^2) = (\bar{x}, s_n^2)$  helyen:

$$H = \begin{pmatrix} -\frac{n}{s_n^2} & 0 \\ 0 & -\frac{n}{2(s_n^2)^2} \end{pmatrix}$$

$H$  negatív definit  $\Rightarrow (\hat{\mu}, \hat{\sigma}^2)$  valóban maximumhelye  $L(\mu, \sigma^2)$ -nek tetszőleges realizáció mellett.

## Egyenletes eloszlás paramétereinek ML becslése

$$L(a, b) = \begin{cases} (b - a)^{-n}, & \text{ha } a \leq x_1, \dots, x_n \leq b, \\ 0 & \text{különben} \end{cases}$$

$a \leq x_1, \dots, x_n \leq b \Leftrightarrow a \leq x_1^*$  és  $x_n^* \leq b$  tartományon  $L$  az  $a$  monoton csökkenő,  $b$  monoton növekvő függvénye  $\Rightarrow$  maximumhelye  $(x_1^*, x_n^*)$ , ahol  $L$  nem deriválható!

### Tétel

*Tegyük fel, hogy  $X$  eloszlása egy  $\theta$  paramétertől függ. Bizonyos regularitási feltételek mellett a  $\theta$  paraméter  $\hat{\theta}_n$  ML becslésére a következők teljesülnek:*

- ▶  $\hat{\theta}_n \xrightarrow{mb} \theta$ , ha  $n \rightarrow \infty$ , azaz  $\hat{\theta}_n$  erősen konzisztens,
- ▶  $E(\hat{\theta}_n) \rightarrow \theta$ , ha  $n \rightarrow \infty$ , azaz  $\hat{\theta}_n$  aszimptotikusan torzítatlan,
- ▶ ha  $D^2(\hat{\theta}_n) \rightarrow \sigma^2$ , ha  $n \rightarrow \infty$ , akkor nincs olyan  $\tilde{\theta}$  torzítatlan becslése  $\theta$ -nak, amelyre  $D^2(\tilde{\theta}) < \sigma^2$ , azaz  $\hat{\theta}_n$  aszimptotikusan hatásos becslése  $\theta$ -nak.

# Momentumok módszere

Több paraméter együttes becslésére használják. Tegyük fel, hogy  $X$  eloszlása  $\theta = (\theta_1, \dots, \theta_k)$  paramétertől függ és létezik az első  $k$  momentum:  $E(X^i) = m_i = g_i(\theta)$ ,  $i = 1, \dots, k$ . Ha létezik a

$$g_i(\theta) = M_i \quad i = 1, \dots, k$$

egyenletrendszernek megoldása, az  $\theta$  **momentum becslése**.

## Példa

$$X \sim \mathcal{N}(\mu, \sigma^2), m_1 = \mu, m_2 = \sigma^2 + \mu^2 \Rightarrow \hat{\mu} = M_1, \hat{\sigma}^2 = M_2 - M_1^2.$$

A tapasztalati momentumok erős konzisztenciája miatt igaz a köv.

## Tétel

*Ha  $g_1, \dots, g_k$  függvények folytonosak, léteznek elsőrendű parciális deriváltjaik és a  $J = \det(\partial g_i / \partial \theta_j)$  Jacobi-determináns nem eltűnő függvény  $\Theta$ -n, akkor  $\theta$  momentum becslése erősen konzisztens.*

# Bayes becslés

Tegyük fel, hogy  $\theta$  (egydimenziós) paraméter diszkrét véletlen változó  $P(\theta = t) = q_t$  ( $t \in \Theta$ ) **a priori** eloszlással. Tekintsük az  $\mathbf{x} = (x_1, \dots, x_n)$  mintarealizációt. Ekkor  $L(t) = P(\mathbf{X} = \mathbf{x} | \theta = t)$  az  $\mathbf{X}$  minta feltételes eloszlása a  $\theta = t$  feltétel mellett.

$P(\mathbf{X} = \mathbf{x}, \theta = t) = L(t) \cdot q_t$ . A Bayes-tétel alapján  $\theta$  **a posteriori** feltételes eloszlása  $\mathbf{X} = \mathbf{x}$  feltétel mellett

$$P(\theta = s | \mathbf{X} = \mathbf{x}) = \frac{L(s) \cdot q_s}{P(\mathbf{X} = \mathbf{x})} = \frac{L(s) \cdot q_s}{\sum_{t \in \Theta} L(t) \cdot q_t}.$$

Vegyük  $\theta$  feltételes várható értékét  $\mathbf{X} = \mathbf{x}$  feltétel mellett:

$$E(\theta | \mathbf{X} = \mathbf{x}) = \sum_{t \in \Theta} t \cdot P(\theta = t | \mathbf{X} = \mathbf{x}) = \frac{\sum_{t \in \Theta} t \cdot L(t) \cdot q_t}{\sum_{t \in \Theta} L(t) \cdot q_t} = T(\mathbf{x}),$$

ami csak a realizáció függvénye, becslésünk a  **$T(\mathbf{X})$**  statisztika.  
(Abszolút folytonos esetben az eljárás hasonló.)

## Definíció

$T(\mathbf{X})$  statisztika a  $\theta$  paraméter  $q$  a priori eloszlás alapján konstruált **Bayes-becslése**.

## Megjegyzés

$E((\theta - \hat{\theta})^2) \geq E((\theta - T(\mathbf{X}))^2)$  minden  $\hat{\theta}$  statisztikára (ahol a várható értéket az  $\mathbf{x}$ ,  $t$  együttes eloszlása alapján vesszük), azaz  $T(\mathbf{X})$  minimalizálja a *négyzetes rizikót*.

## Példa

$X_1, \dots, X_n \sim \mathcal{B}_1(p)$ , ahol  $p \sim \mathcal{B}(a, b)$ . Ekkor  $L(t) = t^k(1-t)^{n-k}$ , ahol  $k = \sum_{i=1}^n x_i$ .

$$q(t) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1} \quad (0 \leq t \leq 1)$$

$$q(t|\mathbf{x}) = \frac{\Gamma(n+a+b)}{\Gamma(k+a)\Gamma(n-k+b)} t^{k+a-1}(1-t)^{n-k+b-1} \quad (0 \leq t \leq 1)$$

$$E(t|\mathbf{X} = \mathbf{x}) = \frac{k+a}{n+a+b} \implies T(\mathbf{X}) = \frac{\sum_{i=1}^n X_i + a}{n+a+b}$$

# Normális eloszlásból származtatott eloszlások

## Definíció

Legyenek  $X_0, \dots, X_{n+m} \sim \mathcal{N}(0, 1)$  függetlenek.

$$\chi^2 = X_1^2 + \dots + X_n^2$$

$n$  szabadságfokú  $\chi^2$  eloszlású  $F_{\chi^2, n}$  eloszlás- és  $f_{\chi^2, n}$  sűrűségfüggvénnyel.

$$t = \frac{X_0 \sqrt{n}}{\sqrt{X_1^2 + \dots + X_n^2}}$$

$n$  szabadságfokú  $Student(t)$  eloszlású  $\Phi_n$  eloszlás- és  $\phi_n$  sűrűségfüggvénnyel.

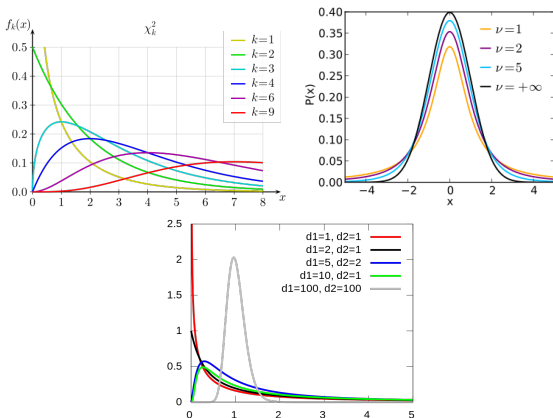
$$F = \frac{n}{m} \cdot \frac{X_1^2 + \dots + X_m^2}{X_{m+1}^2 + \dots + X_{m+n}^2}$$

$(m, n)$  szabadságfokú  $F$  eloszlású  $F_{m, n}$  eloszlás- és  $f_{m, n}$  sűrűségfüggvénnyel.

# Állítás

$$f_{\chi^2, n}(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \quad (x > 0), \quad \phi_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n} \Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

$$f_{m,n}(x) = \frac{n \Gamma(\frac{n+m}{2}) (\frac{n}{m} x)^{\frac{n}{2}-1}}{m \Gamma(\frac{n}{2}) \Gamma(\frac{m}{2}) (1 + \frac{n}{m} x)^{\frac{n+m}{2}}}.$$



## Állítás

- ▶ Ha  $\chi_n^2 \sim \chi^2(n)$ ,  $\chi_m^2 \sim \chi^2(m)$  függetlenek, akkor

$$\chi_n^2 + \chi_m^2 = X_1^2 + \dots + X_n^2 + X_{n+1}^2 + \dots + X_{n+m}^2 \sim \chi^2(n+m).$$

- ▶ Ha  $t_n \sim t(n)$ ,  $n = 1, 2, \dots$ , akkor  $t_n \xrightarrow{\mathcal{D}} X \sim \mathcal{N}(0, 1)$ .
- ▶ Ha  $t \sim t(n)$ , akkor  $t^2 \sim \mathcal{F}(1, n)$ .
- ▶ Ha  $F \sim \mathcal{F}(n, m)$ , akkor  $1/F \sim \mathcal{F}(m, n)$ . Ebből, ha  $x > 0$ ,

$$F_{n,m}(x) = P(F < x) = P(1/F > 1/x) = 1 - F_{m,n}(1/x).$$

- ▶ Ha  $F_n \sim \mathcal{F}(n, m)$ ,  $n = 1, 2, \dots$ ,  $\Rightarrow mF_n \xrightarrow{\mathcal{D}} X \sim \chi_m^2$ ,  $n \rightarrow \infty$ .

## Megjegyzés

$F_{\chi^2, n}(x)$ ,  $\Phi_n(x)$  és  $F_{m,n}(x)$  függvényekre nincs zárt formula.

Némely  $\chi^2$ , Student és F eloszlás néhány kvantilisének közelítő értékét statisztikai táblázatok tartalmazzák.

# Többdimenziós normális eloszlás

## Definíció

A  $d$ -dim.  $\mathbf{Z}$  véletlen vektor **többdimenziós standard normális eloszlású**, ha komponensei független standard normális eloszlásúak.  
Jel:  $\mathbf{Z} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ .

## Észrevétel

Ha  $\mathbf{Z} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ , akkor  $\mathbf{Z}^\top \mathbf{Z} \sim \chi^2(d)$ .

## Definíció

$\mathbf{X} = (X_1, \dots, X_d)^\top$   $d$ -dimenziós véletlen vektor,  $E(X_i^2) < \infty \forall i$ .

Várható érték vektora  $E(\mathbf{X}) = (E(X_1), \dots, E(X_d))^\top$ ,

kovarianciamátrixa  $\text{Cov}(\mathbf{X}) = E((\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^\top)$ , amely komponensei  $\text{Cov}(X_i, X_j) = E((X_i - E(X_i))(X_j - E(X_j)))$ .

## Állítás

Ha  $\mathbf{Z} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ , akkor  $E(\mathbf{Z}) = \mathbf{0}$  és  $\text{Cov}(\mathbf{Z}) = \mathbf{I}_d$ .

## A várható érték vektor és kovarianciamátrix tulajdonságai

Legyen  $\mathbf{X}$  egy  $d$ -dimenziós véletlen vektor, amely kovarianciamátrixa létezik.  $\text{Cov}(\mathbf{X}) = \mathbf{C}$

- ▶ szimmetrikus, azaz  $\mathbf{C} = \mathbf{C}^\top$ ,
- ▶ pozitív szemidefinit, azaz  $\mathbf{x}^\top \mathbf{C} \mathbf{x} \geq 0$  minden  $\mathbf{x} \in \mathbb{R}^d$ -re.
- ▶ Ha  $\mathbf{A} \in \mathbb{R}^{r \times d}$ ,  $\mathbf{b} \in \mathbb{R}^d$ , akkor  $\mathbb{E}(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}\mathbb{E}(\mathbf{X}) + \mathbf{b}$  és  $\text{Cov}(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}\mathbf{C}\mathbf{A}^\top$

### Definíció

Ha  $\mathbf{A}$  egy  $d \times d$ -s (nemszinguláris) mátrix és  $\mathbf{m}$  egy  $d$  dim. vektor,  $\mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{m} \sim \mathcal{N}_d(\mathbf{m}, \mathbf{C})$  (nem elfajult) többdimenziós normális eloszlású, ahol  $\mathbf{C} = \mathbf{A}\mathbf{A}^\top$ .

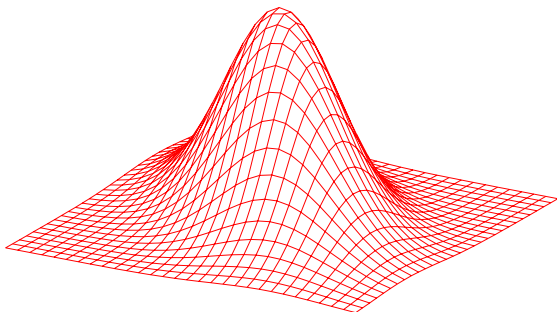
### Állítás

$\mathbb{E}(\mathbf{X}) = \mathbf{m}$ ,  $\text{Cov}(\mathbf{X}) = \mathbf{C}$ .

## Állítás

Ha  $\mathbf{C}$  invertálható,  $\mathbf{X} \sim \mathcal{N}_d(\mathbf{m}, \mathbf{C})$  sűrűségfüggvénye

$$f(\mathbf{x}) = (2\pi)^{-d/2} |\mathbf{C}|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})} \quad \mathbf{x} \in \mathbb{R}^d.$$



## Következmény

Az  $\mathbf{X} \sim \mathcal{N}_d(\mathbf{m}, \mathbf{C})$  véletlen vektor komponensei normális eloszlásúak és pontosan akkor függetlenek, ha  $\mathbf{C}$  diagonális.

## Tétel (Többdimenziós CHT)

Legyenek  $\mathbf{X}_1, \mathbf{X}_2 \dots$  azonos eloszlású független  $d$  dimenziós vektorváltozók,  $E(\mathbf{X}_i) = \mathbf{m}$  és  $\text{Cov}(\mathbf{X}_i) = \mathbf{C}$  (nem feltétlenül invertálható) és  $\mathbf{S}_n = \sum_{i=1}^n \mathbf{X}_i$ . Ekkor

$$\frac{1}{\sqrt{n}}(\mathbf{S}_n - n\mathbf{m}) \xrightarrow{\mathcal{D}} \mathcal{N}_d(\mathbf{0}, \mathbf{C}), \quad \text{ha } n \rightarrow \infty.$$

## Állítás

Legyen  $\mathbf{X} = (X_1, \dots, X_d)^\top \sim \mathcal{N}_d(\mathbf{0}, \mathbf{C})$ . Ekkor

$$\mathbf{x}^\top \mathbf{x} = \sum_{i=1}^d X_i^2 = \sum_{i=1}^d \lambda_i Y_i^2,$$

ahol  $\mathbf{Y} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ ,  $\lambda_1 \geq \dots \geq \lambda_d \geq 0$  a  $\mathbf{C}$  mártix sajátértékei.

## Bizonyítás

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \mathbf{Y} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^\top \mathbf{X} \Rightarrow \|\mathbf{X}\|^2 = \|\mathbf{U}^\top \mathbf{X}\|^2 = \|\mathbf{\Lambda}^{1/2}\mathbf{Y}\|^2.$$

## Tétel (Lukács Jenő)

Legyen  $X_1, \dots, X_n$  egy  $\mathcal{N}(\mu, \sigma^2)$  eloszlásból vett minta.

$$(1) \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

$$(2) \quad \frac{n}{\sigma^2} S^2 \sim \chi^2(n-1),$$

$$(3) \quad \bar{X} \text{ és } S^2 \text{ függetlenek,}$$

$$(4) \quad \frac{\bar{X} - \mu}{S^*/\sqrt{n}} \sim t(n-1).$$

## Bizonyítás

$\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}_n(\mu \mathbf{1}, \sigma^2 \mathbf{I}_n)$ . Legyen  $\mathbf{U}$  egy  $n \times n$ -es ortonormált mátrix, amely első sora  $\mathbf{1}^\top / \sqrt{n}$  és legyen  $\mathbf{Y} = \mathbf{UX}$ .

$$\sum_{i=1}^n Y_i^2 = \mathbf{Y}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{U}^\top \mathbf{U} \mathbf{X} = \mathbf{X}^\top \mathbf{X} = \sum_{i=1}^n X_i^2,$$

amiből  $nS^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2$ ,

mivel  $\mathbf{U}$  definíciója miatt  $Y_1 = \sqrt{n} \cdot \bar{X}$ . Továbbá

$$\mathbf{Y} = \mathbf{U} \mathbf{X} \sim \mathcal{N}_n(\mathbf{U}(\mu \mathbf{1}), \mathbf{U}(\sigma^2 I_n) \mathbf{U}^\top) = \mathcal{N}_n(\sqrt{n} \mu \mathbf{e}_1, \sigma^2 \mathbf{I}) \Rightarrow$$

$$(1) \quad \bar{X} = Y_1 / \sqrt{n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1)$$

$$(2) \quad i > 1 : \frac{Y_i}{\sigma} \sim \mathcal{N}(0, 1) \Rightarrow \frac{nS^2}{\sigma^2} = \sum_{i=2}^n \frac{Y_i^2}{\sigma^2} \sim \chi^2(n-1) \text{ és}$$

(3) ezek függetlensége (a korrelálatlanságból) közvetlenül adódik.

$$(4) \quad \frac{\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}\right) \sqrt{n-1}}{\sqrt{\frac{nS^2}{\sigma^2}}} = \frac{(\bar{X} - \mu) \sqrt{n-1}}{\sqrt{n(n-1)S^2}} = t \sim t(n-1).$$

## Tétel

Legyen  $X_1, \dots, X_{n_1}$  egy  $\mathcal{N}(\mu_1, \sigma_1^2)$ ,  $Y_1, \dots, Y_{n_2}$  egy  $\mathcal{N}(\mu_2, \sigma_2^2)$  eloszlásból vett független minta. A következők teljesülnek:

(1)

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

(2) ha  $\sigma_1 = \sigma_2$ , akkor

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{n_1 S_X^2 + n_2 S_Y^2}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \sim t(n_1 + n_2 - 2)$$

(3)

$$\frac{S_X^{*2} \sigma_2^2}{S_Y^{*2} \sigma_1^2} \sim \mathcal{F}(n_1 - 1, n_2 - 1).$$

## Bizonyítás

Előző tétel alapján  $\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ ,  $\bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$ ,  
 $\frac{n_1}{\sigma_1^2} S_X^2 \sim \chi^2(n_1 - 1)$ ,  $\frac{n_2}{\sigma_2^2} S_Y^2 \sim \chi^2(n_2 - 1)$  függetlenek.

$$(1) \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

$$(2) \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2(n_1 + n_2)}{n_1 n_2}}} \sim \mathcal{N}(0, 1), \quad \frac{n_1 S_X^2 + n_2 S_Y^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

$$\Rightarrow \frac{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2(n_1 + n_2)}{n_1 n_2}}} \cdot \sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{n_1 S_X^2 + n_2 S_Y^2}{\sigma^2}}} \sim t(n_1 + n_2 - 2),$$

$$(3) \frac{n_2 - 1}{n_1 - 1} \cdot \frac{\frac{n_1}{\sigma_1^2} S_X^2}{\frac{n_2}{\sigma_2^2} S_Y^2} = \frac{n_2 - 1}{n_1 - 1} \cdot \frac{\frac{n_1 - 1}{\sigma_1^2} S_X^{*2}}{\frac{n_2 - 1}{\sigma_2^2} S_Y^{*2}} \sim \mathcal{F}(n_1 - 1, n_2 - 1).$$

Legyen  $\theta$  ismeretlen paraméter. Az **intervallumbecslés** lényege olyan intervallum konstruálása statisztikák segítségével, amelybe  $\theta$  nagy valószínűséggel (általában 0,95 vagy 0,99) beleesik.

## Definíció

Legyen  $S_n < T_n$  két statisztika, amelyre  $P_\theta(S_n < \theta < T_n) = 1 - \alpha$ . Ekkor azt mondjuk,  $(S_n, T_n)$  egy  **$1 - \alpha$  megbízhatósági szintű konfidencia intervallum**  $\theta$ -ra.

## A konfidencia-intervallum szerkesztésének sémája

- ▶ Keresünk egy  $Z_n(\theta)$  változót, aminek eloszlása ismert.
- ▶  $Z_n(\theta)$ -ra szerkesztünk intervallumot:  
 $P_\theta(a < Z_n(\theta) < b) = 1 - \alpha$ , ahol  $a, b$  konstansok.
- ▶ A  $Z_n(\theta)$ -ra felírt egyenlőtlenségeket átalakítjuk  $\theta$ -ra felírt egyenlőtlenségekké:  $P_\theta(S_n(a, b) < \theta < T_n(a, b)) = 1 - \alpha$ , ahol  $S_n(a, b), T_n(a, b)$  statisztikák.

## Konfidencia intervallum normális eloszlás várható értékére

Legyen  $X_1, \dots, X_n \mathcal{N}(\mu, \sigma^2)$ -ből vett minta, ahol  $\sigma$  ismert.  $\mu$ -re keresünk  $1 - \alpha$  megbízhatósági szintű konfidencia intervallumot.

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n) \Rightarrow Z_n(\mu) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

$$\begin{aligned} P(-x_\alpha < Z_n(\mu) < x_\alpha) &= P(Z_n(\mu) < x_\alpha) - P(Z_n(\mu) < -x_\alpha) \\ &= \Phi(x_\alpha) - \Phi(-x_\alpha) = 2\Phi(x_\alpha) - 1 = 1 - \alpha, \end{aligned}$$

amiből  $x_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})$ .

$$\begin{aligned} 1 - \alpha &= P\left(-x_\alpha < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < x_\alpha\right) \\ &= P\left(\bar{X} - x_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + x_\alpha \frac{\sigma}{\sqrt{n}}\right). \end{aligned}$$

Az intervallum hossza  $|T_n - S_n| = 2x_\alpha\sigma/\sqrt{n} \rightarrow 0$ , ha  $n \rightarrow \infty$ .

Ha  $\sigma$  ismeretlen, vegyük a becslését. Ekkor

$$Z_n(\mu) = \frac{\bar{X} - \mu}{S^*/\sqrt{n}} \sim t(n-1).$$

A  $t(n-1)$  eloszlás szimmetriáját felhasználva a

$P(-x_\alpha < Z_n(\mu) < x_\alpha) = 2\Phi_{n-1}(x_\alpha) - 1 = 1 - \alpha$  összefüggésből az előzőhöz hasonlóan adódik  $x_\alpha = \Phi_{n-1}^{-1}(1 - \frac{\alpha}{2})$ . Ebből

$$\begin{aligned} 1 - \alpha &= P\left(-x_\alpha < \frac{\bar{X} - \mu}{S^*/\sqrt{n}} < x_\alpha\right) \\ &= P\left(\bar{X} - x_\alpha \frac{S^*}{\sqrt{n}} < \mu < \bar{X} + x_\alpha \frac{S^*}{\sqrt{n}}\right). \end{aligned}$$

Az intervallum hossza  $|T_n - S_n| = 2x_\alpha S^*/\sqrt{n} \xrightarrow{mb} 0$ , ha  $n \rightarrow \infty$ .

## Konfidencia intervallum normális eloszlás szórására

Tudjuk, hogy  $Z_n(\sigma) = nS^2/\sigma^2 \sim \chi^2(n-1)$ .

Legyen  $a_\alpha = F_{\chi^2, n-1}^{-1}(\alpha/2)$  és  $b_\alpha = F_{\chi^2, n-1}^{-1}(1 - \alpha/2)$ ,

$$P(a_\alpha < Z_n(\sigma) < b_\alpha) = F_{\chi^2, n-1}(b_\alpha) - F_{\chi^2, n-1}(a_\alpha) = \left(1 - \frac{\alpha}{2}\right) - \frac{\alpha}{2}.$$

$$\begin{aligned} 1 - \alpha &= P\left(a_\alpha < \frac{nS^2}{\sigma^2} < b_\alpha\right) \\ &= P\left(\sqrt{\frac{nS^2}{b_\alpha}} < \sigma < \sqrt{\frac{nS^2}{a_\alpha}}\right). \end{aligned}$$

## Konf. intervallum normális eloszlások vé. különbségére

Legyenek  $X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$  és  $Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$   
fgn. minták.  $\mu_1 - \mu_2$ -re keresünk konfidencia intervallumot.

$$Z_{n_1, n_2}(\mu_1 - \mu_2) = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

Mivel  $P(-x < Z_{n_1, n_2}(\mu_1 - \mu_2) < x) = 2\Phi(x) - 1$ , az

$x_\alpha = \Phi^{-1}(1 - \alpha/2)$  választással kapjuk:

$$\begin{aligned} 1 - \alpha &= P\left(-x_\alpha < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < x_\alpha\right) \\ &= P\left(\bar{X} - \bar{Y} - x_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + x_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right). \end{aligned}$$

Ha  $\sigma_1 = \sigma_2 = \sigma$  ismeretlen, legyen

$$Z_{n_1, n_2}(\mu_1 - \mu_2) = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{D_*} \sim t(n_1 + n_2 - 2),$$

ahol

$$D_*^2 = (n_1 S_X^2 + n_2 S_Y^2) \frac{n_1 + n_2}{n_1 n_2 (n_1 + n_2 - 2)}.$$

$P(-x < Z_{n_1, n_2}(\mu_1 - \mu_2) < x) = 2\Phi_{n_1+n_2-1}(x) - 1$ , így az

$x_\alpha = \Phi_{n_1+n_2-2}^{-1}(1 - \alpha/2)$  választással kapjuk:

$$\begin{aligned} 1 - \alpha &= P\left(-x_\alpha < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{D_*} < x_\alpha\right) \\ &= P\left(\bar{X} - \bar{Y} - x_\alpha D_* < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + x_\alpha D_*\right). \end{aligned}$$

## Konf. intervallum normális eloszlások szóráshányadosára

$X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$  és  $Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$  fgn. minták.  
 $\sigma_1/\sigma_2$ -re keresünk konfidencia intervallumot. Legyen

$$Z_{n_1, n_2}(\sigma_1/\sigma_2) = \frac{S_X^{*2} \cdot \sigma_2^2}{S_Y^{*2} \cdot \sigma_1^2} \sim \mathcal{F}(n_1 - 1, n_2 - 1).$$

Ha  $a_\alpha = F_{n_1-1, n_2-1}^{-1}(\alpha/2)$  és  $b_\alpha = F_{n_1-1, n_2-1}^{-1}(1 - \alpha/2)$ , akkor

$$\begin{aligned} 1 - \alpha &= F_{n_1-1, n_2-1}(b_\alpha) - F_{n_1-1, n_2-1}(a_\alpha) \\ &= P\left(a_\alpha < \frac{S_X^{*2} \cdot \sigma_2^2}{S_Y^{*2} \cdot \sigma_1^2} < b_\alpha\right) \\ &= P\left(\sqrt{\frac{S_X^{*2}}{S_Y^{*2} b_\alpha}} < \frac{\sigma_1}{\sigma_2} < \sqrt{\frac{S_X^{*2}}{S_Y^{*2} a_\alpha}}\right). \end{aligned}$$

Az ismeretlen  $P \in \mathcal{P}$  eloszlásra (paraméteres esetben  $\theta \in \Theta$ -ra) vonatkozó hipotéziseket vizsgálunk. A fogalmakat paraméteres próbákra vezetjük be, nemparaméteresekre hasonlóan definiálhatók.

Legyen  $\Theta = \Theta_0 \cup \Theta_1$  a paraméterter egy osztályozása. A következő hipotézispárt vizsgáljuk:

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1.$$

$H_0$ -t **nullhipotézis**nek,  $H_1$ -et **ellenhipotézis**nek vagy **alternatív hipotézis**nek nevezzük. Ha  $|\Theta_0| = 1$ ,  $H_0$  egyszerű, különben összetett. A nullhipotézis általában  $H_0 : \theta_i = c$  alakú. Ekkor ha  $H_1 : \theta_i \neq c$ , akkor **kétoldali**, ha  $H_1 : \theta_i < c$  vagy  $H_1 : \theta_i > c$ , akkor **egyoldali ellenhipotézis**, a hipotézispárról döntő próbát is ennek megfelelően nevezzük majd egy- illetve kétoldalinak.

**Cél:** nagy valószínűséggel helyesen dönteni a hipotézisekről.

## Definíció

Legyen  $\mathcal{X} = \mathcal{X}_e \cup \mathcal{X}_k$  felbontás, amelyre  $P_\theta(\mathbf{X} \in \mathcal{X}_k) = \alpha$  minden  $\theta \in \Theta_0$  esetén és tekintsük az  $\mathbf{x}$  realizációt. A fenti hipotézispárt tesztelő,  $\mathcal{X}_k$  **kritikus tartománnyal** értelmezett  **$\alpha$  terjedelmű próba elfogadja  $H_0$ -t**, ha  $\mathbf{x} \in \mathcal{X}_e$  és **elutasítja**, ha  $\mathbf{x} \in \mathcal{X}_k$ .

$\mathcal{X}_e$ -t **elfogadási tartomány**nak nevezzük. Ha  $\theta \in \Theta_0$  és  $\mathbf{x} \in \mathcal{X}_k$ , akkor **elsőfajú**, ha  $\theta \in \Theta_1$  és  $\mathbf{x} \in \mathcal{X}_e$ , **másodfajú hibáról** beszélünk.

## Megjegyzés

$\alpha$ -t nevezik **szignifikancia-szint**nek is és általában 0,05-nak vagy 0,01-nak, ritkán 0,1-nek választják.

## A próbák általános menete

Vesünk egy  $T_n$  **próbatatisztikát**, amely eloszlása ismert  $H_0$  mellett. Legyen  $c_\alpha$  **kritikus érték** ennek az eloszlásnak az  $1 - \alpha$  kvantilise.  $\mathbf{x} \in \mathcal{X}_k \Leftrightarrow T_n \geq c_\alpha$ . Kétoldali próba esetén néha két kritikus értéket adnak meg, az  $a_\alpha$   $\alpha/2$  és a  $b_\alpha$   $1 - \alpha/2$  kvantilist,  $\mathbf{x} \in \mathcal{X}_e \Leftrightarrow a_\alpha < T_n < b_\alpha$ .

## Példa: u próba

Tekintsünk egy  $\mathcal{N}(\mu, \sigma^2)$  eloszlásból vett mintát. Tegyük fel, hogy  $\sigma$  ismert. Rögzített  $\mu_0$  esetén vizsgáljuk a következő hipotéziseket:

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

- ▶ Tekintsük a következő próbastatisztikát:

$$u = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

ami  $H_0$  teljesülése esetén standard normális eloszlású.

- ▶ Ehhez és  $\alpha > 0$ -hoz keressük az  $u_\alpha$  kritikus értéket úgy, hogy

$$P_{H_0}(-u_\alpha < u < u_\alpha) = \Phi(u_\alpha) - \Phi(-u_\alpha) = 2\Phi(u_\alpha) - 1 = 1 - \alpha$$

legyen, így  $u_\alpha = \Phi^{-1}(1 - \alpha/2)$  adódik,  $\mathcal{X}_k = \{\mathbf{x} : |u| \geq u_\alpha\}$ .

- ▶ Döntés: ha  $|u| < u_\alpha$ , akkor  $H_0$ -t elfogadjuk, különben elvetjük.

## Példa

Megmérték 10 ember IQ-ját, a következő adatokat kapták:  
133, 144, 103, 90, 83, 133, 140, 88, 81, 89. Tegyük fel, hogy a  
szórás 25. Teszteljük a  $H_0 : \mu = 100$  vs  $H_1 : \mu \neq 100$  hipotézispárt!

- ▶ A próbastatisztika

$$u = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{108,4 - 100}{25/\sqrt{10}} \approx 1,06.$$

- ▶ Mivel  $\alpha = 0,05$ , a kritikus érték  $u_{0,05} = \Phi^{-1}(0,975) = 1,96$ .
- ▶ Döntés:  $|u| \leq u_{0,05}$ , ezért elfogadjuk a nullhipotézist.
- ▶ A 0,95 megbízhatósági szintű konfidencia intervallum  $\mu$ -re

$$\left( 108,4 - 1,96 \cdot \frac{25}{\sqrt{10}}; 108,4 + 1,96 \cdot \frac{25}{\sqrt{10}} \right) \approx (92,9; 123,89).$$

## Példa: egyoldali u próba („nagyobb” eset)

Most a következő hipotéziseket vizsgáljuk:

$$H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0.$$

- ▶ A próbastatisztika most is

$$u = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

ami  $H_0$  teljesülése esetén standard normális eloszlású.

- ▶ Ehhez és  $\alpha > 0$ -hoz keressük az  $u_\alpha$  kritikus értéket úgy, hogy

$$P_{H_0}(u < u_\alpha) = \Phi(u_\alpha) = 1 - \alpha$$

legyen amiből  $u_\alpha = \Phi^{-1}(1 - \alpha)$  adódik,  $\mathcal{X}_k = \{\mathbf{x} : u \geq u_\alpha\}$ .

- ▶ Döntés: ha  $u < u_\alpha$ , akkor  $H_0$ -t elfogadjuk, különben elutasítjuk.

## Tévedési lehetőségek

- ▶ **Elsőfajú hiba:**  $H_0$  teljesül, de elvetjük. Ennek  $H_0$  melletti feltételes valószínűsége  $\alpha$  terjedelmű próba esetén  $p_1 = P_{H_0}(\mathbf{X} \in \mathcal{X}_\alpha) = P(\mathbf{X} \in \mathcal{X}_k | \theta \in \Theta_0) = \alpha$ .
- ▶ **Másodfajú hiba:**  $H_0$  nem teljesül, de elfogadjuk. Ennek  $H_1$  melletti feltételes valószínűsége  $p_2 = P_{H_1}(\mathbf{X} \notin \mathcal{X}_k)$ .

## Definíció

$e_n(\alpha, \theta) = 1 - p_2 = P_{H_1}(\mathbf{X} \in \mathcal{X}_k)$  a próba **erőfüggvénye (ereje)**.

## Definíció

Egy  $\alpha$  terjedelmű próba **torzítatlan**, ha  $e_n(\alpha, \theta) \geq \alpha$  és **konzisztens**, ha minden  $\theta \in \Theta_1$  esetén  $\lim_{n \rightarrow \infty} e_n(\alpha, \theta) = 1$ .

## Megjegyzés

A konzisztencia azt jelenti, hogy a mintaelemszám növelésével a másodfajú hiba tetszőlegesen kicsivé tehető.

## Állítás

*Az egyoldali  $u$  próba torzítatlan és konzisztens.*

## Bizonyítás

$$\begin{aligned}e_n(\alpha, \mu) = 1 - p_2 &= 1 - P_{H_1} \left( \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq u_\alpha \right) \\&= 1 - P_{H_1} \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) \\&= 1 - \Phi \left( u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) \\&\geq 1 - \Phi(u_\alpha) = 1 - (1 - \alpha).\end{aligned}$$

Továbbá  $\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \rightarrow -\infty$ , ha  $n \rightarrow \infty$ , így

$$e_n(\alpha, \mu) = 1 - p_2 = 1 - \Phi \left( u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) \rightarrow 1, \text{ ha } n \rightarrow \infty.$$

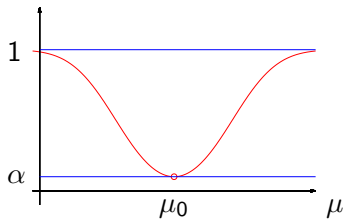
## Állítás

*Az  $u$  próba torzítatlan és konzisztens.*

## Bizonyítás

$$\begin{aligned}e_n(\alpha, \mu) &= 1 - P_{H_1} \left( \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| < u_\alpha \right) \\&= 1 - P_{H_1} \left( -u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) \\&= 1 - \Phi \left( u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) + \Phi \left( -u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) \stackrel{?}{>} \alpha.\end{aligned}$$

Kell:  $e_n^*(\mu) = 1 - \Phi \left( u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) + \Phi \left( -u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right)$   
minimumhelye  $\mu = \mu_0$ -ban van.



$$\begin{aligned}
 (e_n^*)'(\mu) &= \left( 1 - \Phi \left( u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) + \Phi \left( -u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) \right)' \\
 &= -\phi \left( u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) \frac{-\sqrt{n}}{\sigma} + \phi \left( -u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) \frac{-\sqrt{n}}{\sigma},
 \end{aligned}$$

$\phi$  szimmetriája és monotonitási tulajdonágai miatt ez pontosan akkor 0, ha  $\mu = \mu_0$ .

$$(e_n^*)''(\mu_0) = -\phi'(u_\alpha) \left( \frac{-\sqrt{n}}{\sigma} \right)^2 + \phi'(-u_\alpha) \left( \frac{-\sqrt{n}}{\sigma} \right)^2 = 2\phi'(u_\alpha) \frac{-n}{\sigma^2} < 0.$$

Konzisztencia: ha  $\mu \neq \mu_0$ ,

$$1 - \Phi\left(u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) + \Phi\left(-u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \rightarrow \infty \quad n \rightarrow \infty.$$

Továbbá ha  $\mu > \mu_0$ ,  $\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \rightarrow -\infty$ , ha  $n \rightarrow \infty$ , és ha  $\mu < \mu_0$ ,  $\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \rightarrow 1$ , ha  $n \rightarrow \infty$ , amiből következik a konzisztencia.

## Megjegyzések

- ▶ Az  $u$  próba pontosan akkor fogadja el a nullhipotézist, ha  $\mu_0$  belesik a várható értékre vonatkozó  $1 - \alpha$  megbízhatósági szintű konfidencia intervallumba ismert szórás esetén.
- ▶ A CHT miatt az  $u$  próba nagy minták esetén nem csak normális eloszlásból származó mintára alkalmazható.

## Megjegyzések

- ▶ A próbák gyakorlati alkalmazásában a kritikus érték meghatározása helyett azt a  **$p$ -értéket** (más néven **szignifikancia** értéket) határozzák meg, amelyre  $u_p = u$ . Az  $u$  próbára egyoldali esetben

$$\begin{aligned}\Phi^{-1}(1 - p) &= u \\ p &= p = 1 - \Phi(u),\end{aligned}$$

kétoldali esetben

$$\begin{aligned}\Phi^{-1}(1 - p/2) &= |u| \\ p &= p = 2(1 - \Phi(|u|)).\end{aligned}$$

Döntés: Ha  $p > \alpha$ ,  $H_0$ -t elfogadjuk.

## A hibák valószínűségével kapcsolatos problémák

- ▶  $p_2$ -t általában nem ismerjük.
- ▶ Az elsőfajú hiba valószínűségének értelmezésével is vigyázni kell:

$$P(\mathbf{X} \in \mathcal{X}_k | H_0) = p_1 = \alpha$$

$$P(\mathbf{X} \notin \mathcal{X}_k | H_1) = p_2 \Rightarrow P(\mathbf{X} \in \mathcal{X}_k | H_1) = 1 - p_2$$

Legyen  $P(H_0) = q$ , ekkor

$$\begin{aligned} P(H_0 | \mathbf{X} \in \mathcal{X}_k) &= \frac{P(\mathbf{X} \in \mathcal{X}_k | H_0)P(H_0)}{P(\mathbf{X} \in \mathcal{X}_k | H_0)P(H_0) + P(\mathbf{X} \in \mathcal{X}_k | H_1)P(H_1)} \\ &= \frac{\alpha \cdot q}{\alpha \cdot q + (1 - p_2)(1 - q)}. \end{aligned}$$

$\Rightarrow$  ha  $p_2$  és  $q$  nagy, akkor  $P(H_0 | \mathbf{X} \in \mathcal{X}_k)$  is nagy!

- ▶ Publikációs torzítás

Általánosan használható módszer próbák konstrukciójára.  
Tegyük fel, hogy  $\dim(\Theta) = r$  véges és  $\dim(\Theta_0) = k < r$ . A

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1$$

hipotézispárt eldöntő likelihood-hányados próba statisztikája

$$\lambda_n(\mathbf{X}) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)}.$$

## Állítás

*Bizonyos regularitási feltételek mellett,  $H_0$  teljesülése esetén*  
 $-2 \ln \lambda_n(\mathbf{X}) \xrightarrow{\mathcal{D}} \chi^2(r - k),$  ha  $n \rightarrow \infty$ .

Tehát a kritikus érték  $-2 \ln \lambda_\alpha = F_{\chi^2, r-k}^{-1}(1 - \alpha)$ , a kritikus tartomány  $\mathcal{R}_k = \{\mathbf{x} : \lambda_n(\mathbf{X}) \leq \lambda_\alpha\}$ .

## Példa

Legyen  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$ , ahol  $\sigma_0$  ismert.  $H_0 : \mu = \mu_0$ ,  
 $H_1 : \mu \neq \mu_0$ .

$$\begin{aligned}\lambda_n(\mathbf{X}) &= \frac{\sup_{\mu=\mu_0} L(\theta)}{\sup_{\mu \in \mathbb{R}} L(\theta)} \\ &=^* \frac{(\sqrt{2\pi}\sigma_0)^{-n} e^{-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (X_i - \mu_0)^2}}{(\sqrt{2\pi}\sigma_0)^{-n} e^{-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (X_i - \bar{X})^2}} \\ &= e^{-\frac{1}{2\sigma_0^2} (\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \bar{X})^2)} \\ &=^{**} e^{-\frac{n}{2\sigma_0^2} (\bar{X} - \mu_0)^2}.\end{aligned}$$

\* A számlálóban  $\mu$  értéke rögzített, a nevezőben a  $\mu$  paraméter ML becslését írjuk:  $\hat{\mu} = \bar{X}$ . \*\* Steiner egyenlőséget alkalmazzuk.

$-2 \ln \lambda_n(\mathbf{X}) = \frac{n(\bar{X} - \mu_0)^2}{\sigma_0^2} = u^2 \sim \chi^2(1)$ , ahol  $u$  az  $u$  próba statisztikája.

## t próba (Student próba)

Tekintsünk egy  $\mathcal{N}(\mu, \sigma^2)$  eloszlásból vett mintát, ahol  $\sigma$  ismeretlen.

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

- ▶ A próbastatisztika  $t = \frac{\bar{X} - \mu_0}{S^*/\sqrt{n}} \sim t(n-1)$   $H_0$  mellett,
- ▶  $t_\alpha = \Phi_{n-1}^{-1}(1 - \alpha/2)$  és  $\mathcal{X}_k = \{\mathbf{x} : |t| \geq t_\alpha\}$ .

## Egyoldali t próba

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

- ▶ A próbastatisztika ugyanaz mint fent,
- ▶  $t_\alpha = \Phi_{n-1}^{-1}(1 - \alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : t \geq t_\alpha\}$ .

## Példa

Megmérték 10 ember IQ-ját, a következő adatokat kapták:  
133, 144, 103, 90, 83, 133, 140, 88, 81, 89.

$$H_0 : \mu = 100, \quad H_1 : \mu \neq 100$$

$$\bar{X} = 108,4 \text{ és } S^* \approx 25,89, \quad t_{0,05} = \Phi_9^{-1}(0,975) = 2,262.$$

$$t = \frac{\bar{X} - \mu_0}{S^*/\sqrt{n}} = \frac{108,4 - 100}{25,89 \cdot \sqrt{10}} \approx 0,1$$

$|t| \leq t_{0,05}$ , ezért elfogadjuk a nullhipotézist.

Az  $1 - \alpha$  megbízhatósági szintű konfidencia intervallum

$$\left( 108,4 - 2,262 \frac{25,89}{\sqrt{10}}; 108,4 + 2,262 \frac{25,89}{\sqrt{10}} \right) = (89,88; 126,92).$$

## Megjegyzések

- ▶ Belátható, hogy a két- és egyoldali t-próba is konzisztens.
- ▶ Könnyen látható, hogy az  $\alpha$  terjedelmű kétoldali t-próba pontosan akkor fogadja el  $H_0$ -t, ha  $\mu_0$  benne van az ismeretlen  $\sigma$  esetén  $\mu$ -re konstruált  $1 - \alpha$  megbízhatósági szintű konfidencia intervallumban.
- ▶  $t$  aszimptotikus normalitása miatt nagy minta esetén ugyanazt a kritikus értéket használhatjuk, mint az u-próbánál.
- ▶ Nagy mintaelemszám esetén – a centrális határeloszlástétel miatt – mind az u-, mind a t-próba tetszőleges eloszlásból vett minta esetén alkalmazható.
- ▶ Az egyoldali u- és t-próbának van  $H_0 : \mu = \mu_0, H_1 : \mu < \mu_0$  alternatívákat vizsgáló változata is, ekkor a próbastatisztika helyett annak  $-1$ -szeresét használjuk. Általában az egyoldali próbák statisztikája ugyanaz, mint a kétoldali változat, csak az elfogadási és a kritikus tartományt módosítják.

## Páros t próba

Összetartozó (nem független) minta esetén a különbségváltozóra vonatkozó, ún. **önkontrollos vizsgálat** alkalmazható.

Legyen  $(X, Y)$  olyan vektorváltozó, amely komponenseinek létezik a várható értéke:  $E(X) = \mu_1$  és  $E(Y) = \mu_2$ , továbbá  $Z = X - Y$  normális eloszlású. Tekintsük a  $(X_1, Y_1), \dots, (X_n, Y_n)$  mintát. A

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

hipotézispárt vizsgáljuk. Az eljárás a következő:

- ▶ Képezzük a  $Z_i = X_i - Y_i$  változókat, amik függetlenek és normális eloszlásúak  $\mu = \mu_1 - \mu_2$  várható értékkel.
- ▶ A  $Z_1, \dots, Z_n$  minta segítségével **egymintás t**-próbával teszteljük a  $H_0 : \mu = 0$  nullhipotézist. Ha  $H_1 : \mu_1 > \mu_2$ , akkor az egyoldali t próbát alkalmazzuk.

## Kétmintás t-próba

Legyenek  $X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$  és  $Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma^2)$   
fgn. minták.

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

► A próbastatisztika 
$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{n_1 S_X^2 + n_2 S_Y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

$\sim t(n_1 + n_2 - 2)$   $H_0$  mellett,

►  $t_\alpha = \Phi_{n_1+n_2-2}^{-1}(1 - \alpha/2)$  és  $\mathcal{X}_k = \{\mathbf{x} : |t| \geq t_\alpha\}.$

## Egyoldali változat

►  $H_1 : \mu_1 > \mu_2$ , a próbastatisztika ua.

►  $t_\alpha = \Phi_{n_1+n_2-2}^{-1}(1 - \alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : t \geq t_\alpha\}.$

## Welch próba

Tekintsük az  $X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$  és  $Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$  független mintákat.

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

- A próbastatisztika  $t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^{*2}}{n_1} + \frac{S_Y^{*2}}{n_2}}} \sim t(df)$  közelítőleg  $H_0$

mellett,

$$\text{ahol } df \approx \left( \frac{c^2}{n_1-1} + \frac{(1-c)^2}{n_2-1} \right)^{-1} \text{ és } c = \frac{S_X^{*2}/n_1}{S_X^{*2}/n_1 + S_Y^{*2}/n_2}.$$

- $t_\alpha = \Phi_{df}^{-1}(1 - \alpha/2)$  és  $\mathcal{X}_k = \{\mathbf{x} : |t| \geq t_\alpha\}.$

## Egyoldali változat

- $H_1 : \mu_1 > \mu_2$ , a próbastatisztika ua.
- $t_\alpha = \Phi_{df}^{-1}(1 - \alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : t \geq t_\alpha\}.$

## Megjegyzések

- ▶ Belátható, hogy a kétmintás t próba konzisztens, és nagy minta esetén tetszőleges eloszlásra alkalmazható.
- ▶ Könnyen látható, hogy az  $\alpha$  terjedelmű kétmintás t-próba pontosan akkor fogadja el a nullhipotézist, ha 0 benne van az ismeretlen szórás esetén  $\mu_1 - \mu_2$ -re konstruált  $1 - \alpha$  megbízhatósági szintű konfidencia intervallumban.
- ▶ A kétmintás t-próba feltétele a szórások egyezése, ezért a várható értékek egyenlőségének tesztelése a következő eljárás szerint megy:
  - ▶ A szórások egyenlőségét teszteljük a később ismertetendő F-próbával.
  - ▶ Ha elfogadjuk az F próba nullhipotézisét, akkor alkalmazzuk a kétmintás t próbát.
  - ▶ Ha elutasítjuk az F próba nullhipotézisét, akkor a Welch próbát alkalmazzuk.

(Mi a probléma ezzel az eljárással?)

## A (varianciára vonatkozó) $\chi^2$ próba

Tekintsünk egy  $\mathcal{N}(\mu, \sigma)$  eloszlásból vett mintát.

$$H_0 : \sigma = \sigma_0, \quad H_1 : \sigma \neq \sigma_0.$$

- ▶ A próbastatisztika  $\chi^2 = \frac{nS^2}{\sigma_0^2} \sim \chi^2(n-1)$   $H_0$  mellett,
- ▶  $\chi_{1,\alpha}^2 = F_{\chi^2, n-1}^{-1}(\alpha/2)$ ,  $\chi_{2,\alpha}^2 = F_{\chi^2, n-1}^{-1}(1 - \alpha/2)$  és

$$\mathcal{X}_k = \{\mathbf{x} : \chi^2 \leq \chi_{1,\alpha}^2\} \cup \{\mathbf{x} : \chi^2 \geq \chi_{2,\alpha}^2\}.$$

## Egyoldali változat

- ▶  $H_1 : \sigma > \sigma_0$ , a próbastatisztika ua.
- ▶  $\chi_\alpha^2 = F_{\chi^2, n-1}^{-1}(1 - \alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : \chi^2 \geq \chi_\alpha^2\}$ .

## F próba

Tekintsük az  $X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$  és  $Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$  független mintákat.

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

- ▶ A próbastatisztika  $F = \frac{S_X^{*2}}{S_Y^{*2}} \sim \mathcal{F}(n_1 - 1, n_2 - 1)$   $H_0$  esetén.
- ▶  $f_{1,\alpha} = F_{n_1-1, n_2-1}^{-1}(\alpha/2)$  és  $f_{2,\alpha} = F_{n_1-1, n_2-1}^{-1}(1 - \alpha/2)$ .

$$\mathcal{X}_k = \{\mathbf{x} : F \leq f_{1,\alpha}\} \cup \{\mathbf{x} : F \geq f_{2,\alpha}\}.$$

## Megjegyzések

- ▶ A gyakorlatban  $F^* = \max\{F, 1/F\}$  (ha  $1/F > 1$ , akkor tulajdonképpen a két minta szerepét felcseréljük) és elegendő  $F^*$ -hoz a felső kritikus értéket nézni.
- ▶ Az  $\alpha$  terjedelmű F próba pontosan akkor fogadja el a nullhipotézist, ha 1 benne van a  $\sigma_1/\sigma_2$ -re konstruált  $1 - \alpha$  megbízhatósági szintű konfidencia intervallumban.

## Valószínűségek tesztelése $\chi^2$ próbával

Legyen  $A_1, \dots, A_r$  teljes eseményrendszer,  $p_i > 0$  minden  $i$ -re és  $p_1 + \dots + p_r = 1$ . Végezzünk  $n$  független megfigyelést.

$$H_0 : P(A_i) = p_i, \quad i = 1, \dots, r,$$

$H_1$  pedig  $H_0$  tagadása.

► Jelölje  $\nu_i$  az  $A_i$  bekövetkezéseinek számát. A próbastatisztika

$$\chi^2 = \sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i}$$

## Tétel

$H_0$  esetén  $\chi^2 \xrightarrow{\mathcal{D}} \chi^2(r-1)$ , ha  $n \rightarrow \infty$ .

## Bizonyítás

$$\underline{\nu} = (\nu_1, \dots, \nu_r)^\top \sim \text{Poly}_n(p_1, \dots, p_r) \quad (p_i > 0 \ \forall i).$$

$$E(\underline{\nu}) = n\mathbf{p}, \text{Cov}(\underline{\nu}) = \mathbf{P} - \mathbf{p}\mathbf{p}^\top, \text{ ahol } \mathbf{P} = \text{diag}(\mathbf{p}). \text{ CHT:}$$

$$\mathbf{S}_n = \mathbf{P}^{-1/2} \underline{\nu} \Rightarrow \frac{1}{\sqrt{n}} \mathbf{P}^{-1/2} \underline{\nu} \xrightarrow{\mathcal{D}} \mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{P}^{-1/2} \mathbf{C} \mathbf{P}^{-1/2})$$

$$\mathbf{P}^{-1/2} \mathbf{C} \mathbf{P}^{-1/2} = \mathbf{I}_r - \mathbf{q}\mathbf{q}^\top, \text{ ahol } \mathbf{q} = (\sqrt{p_1}, \dots, \sqrt{p_r})^\top$$

$\mathbf{q}\mathbf{q}^\top$  a  $\mathbf{q}$  irányra való 1 rangú vetítés  $\Rightarrow \mathbf{P}^{-1/2} \mathbf{C} \mathbf{P}^{-1/2}$   $r - 1$  rangú vetítés, sajátértékei  $\lambda_1 = \dots = \lambda_{r-1} = 1, \lambda_r = 0$ . Az előző állítás alapján ha  $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{P}^{-1/2} \mathbf{C} \mathbf{P}^{-1/2})$ , akkor  $\mathbf{X}^\top \mathbf{X} \sim \chi^2(r - 1)$ , valamint

$$\frac{1}{n} \mathbf{P}^{-1/2} \underline{\nu} \cdot \underline{\nu}^\top \mathbf{P}^{-1/2} = \sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i}.$$

$$\blacktriangleright \chi_\alpha^2 = F_{\chi^2, r-1}^{-1}(1 - \alpha) \text{ és } \mathcal{X}_k = \{\mathbf{x} : \chi^2 \geq \chi_\alpha^2\}.$$

## Tiszta illeszkedésvizsgálat $\chi^2$ próbával

Legyen  $X_1, \dots, X_n$  minta ismeretlen  $F$  eloszlásfüggvénnyel.

$$H_0 : F(x) = F_0(x) \quad \forall x \in \mathbb{R},$$

ahol  $F_0$  egy adott eloszlásfüggvény.

- ▶ Legyenek  $-\infty = x_0 < x_1 < \dots < x_{r-1} < x_r = \infty$  osztópontok, valamint  $A_i = \{X \in [x_{i-1}, x_i)\}$ ,  $i = 1, \dots, r$ .

$$P(A_i) = P(x_{i-1} \leq X < x_i) = F(x_i) - F(x_{i-1}).$$

- ▶ Legyen  $p_i = F_0(x_i) - F_0(x_{i-1}) > 0$  minden  $i$ -re.
- ▶ A fentiek alapján  $H_0$  maga után vonja az alábbi hipotézist:

$$H'_0 : P(A_i) = p_i, \quad i = 1, \dots, r,$$

tehát a feladatot visszavezettük valószínűségek tesztelésére.

## Megjegyzések

- ▶ Ha  $X$  értékkészlete véges:  $x_1, \dots, x_r$ , akkor nincs szükség osztópontokra, hanem az  $A_i = \{X = x_i\}$  választással a valószínűségek tesztelését közvetlenül alkalmazhatjuk.
- ▶ Jól használható ökölszabály a mintaelemszámra:  
 $n > \max_{1 \leq i \leq r} \{10/p_i\}.$

## Becsléses illeszkedésvizsgálat $\chi^2$ próbával

$$H_0 : F(x) = F_{\theta_1, \dots, \theta_k}(x) \quad \forall x \in \mathbb{R},$$

ahol  $F_{\theta_1, \dots, \theta_k}$  adott eloszlásfüggvény  $\theta_1, \dots, \theta_k$  ismeretlen paraméterekkel.

Vegyük a paraméterek  $\hat{\theta}_1, \dots, \hat{\theta}_k$  becsléseit és legyen  $F_0 = F_{\hat{\theta}_1, \dots, \hat{\theta}_k}$ . Az eljárás ugyanaz, mint a tiszta illeszkedésvizsgálatnál, de itt  $H_0$  mellett  $n \rightarrow \infty$  esetén  $\chi^2 \xrightarrow{\mathcal{D}} \chi^2(r - k - 1).$

## Homogenitásvizsgálat $\chi^2$ próbával

Tekintsünk két mintát:  $X_1, \dots, X_{n_1}$  és  $Y_1, \dots, Y_{n_2}$ .

$$H_0 : P(X < x) = P(Y < x) \quad \forall x \in \mathbb{R}.$$

- ▶ Legyenek most is  $-\infty = x_0 < x_1 < \dots < x_{r-1} < x_r = \infty$  osztópontok és  $\mu_i$  az  $[x_{i-1}, x_i)$  intervallumba eső első,  $\nu_i$  a második mintabeli elemek száma,  $\mu_i + \nu_i > 0$  minden  $i$ -re.

$$\chi^2 = n_1 n_2 \sum_{i=1}^r \frac{1}{\mu_i + \nu_i} \left( \frac{\mu_i}{n_1} - \frac{\nu_i}{n_2} \right)^2.$$

$H_0$  esetén  $n_1 \rightarrow \infty, n_2 \rightarrow \infty$  esetén  $\chi^2 \xrightarrow{\mathcal{D}} \chi^2(r-1)$ .

- ▶ Tehát most is  $\chi_\alpha^2 = F_{\chi^2, r-1}^{-1}(1-\alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : \chi^2 \geq \chi_\alpha^2\}$ .

## Megjegyzés

Diszkrét véges értékkeszletű változók esetén itt sincs szükség az osztópontokra: intervallumokba esés helyett  $x_i$  értéket keresünk.

## Függetlenségvizsgálat $\chi^2$ próbával

Legyen  $X$  értékészlete  $x_1, \dots, x_r$  és  $Y$ -é  $y_1, \dots, y_s$ .

$$H_0 : P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j), \quad 1 \leq i \leq r, 1 \leq j \leq s,$$

azaz  $X$  és  $Y$  függetlenek. Vegyük az  $(X_1, Y_1), \dots, (X_n, Y_n)$  mintát.

- ▶ Legyen  $\nu_{ij}$  az  $(x_i, y_j)$  gyakorisága a mintában (a  $\nu_{ij}$  értékeket az ún. **kontingenciatáblázat** tartalmazza),  $\nu_{i.} = \sum_{j=1}^s \nu_{ij}$  és  $\nu_{.j} = \sum_{i=1}^r \nu_{ij}$ .

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^s \frac{(\nu_{ij} - \frac{\nu_{i.} \nu_{.j}}{n})^2}{\nu_{i.} \nu_{.j}},$$

$H_0$  mellett  $n \rightarrow \infty$  esetén  $\chi^2 \xrightarrow{D} \chi^2((r-1)(s-1))$ .

- ▶ Tehát  $\chi_\alpha^2 = F_{\chi^2, (r-1)(s-1)}^{-1}(1 - \alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : \chi^2 \geq \chi_\alpha^2\}$ .

## Megjegyzés

Ennek speciális esete a homogenitásra vonatkozó  $\chi^2$  próba.

## Példa

Azt szeretnénk tesztelni, hogy a dohányzás nemtől független-e. Megkérdeztük az utcán szembejövő első 50 embert. Köztük 6 dohányos nő, 18 nemdohányos nő, 14 dohányos férfi és 12 nemdohányos férfi volt.

Legyen  $X$  : nem,  $Y$  : dohányzás indikátorváltozója,  $H_0 : X, Y$  fgn.

| megf.  | $X=1$ | $X=2$ | $\sum$ | elvárt | $X=1$ | $X=2$ | $\sum$ |
|--------|-------|-------|--------|--------|-------|-------|--------|
| $Y=0$  | 12    | 18    | 30     | $Y=0$  | 15,6  | 14,4  | 30     |
| $Y=1$  | 14    | 6     | 20     | $Y=1$  | 10,4  | 9,6   | 20     |
| $\sum$ | 26    | 24    | 50     | $\sum$ | 26    | 24    | 50     |

$$\chi^2 = \frac{(12-15,6)^2}{15,6} + \frac{(18-14,4)^2}{14,4} + \frac{(14-10,4)^2}{10,4} + \frac{(6-9,6)^2}{9,6} = 4,326.$$

$$\chi_{0,05}^2 = F_{\chi^2,1}^{-1}(0,95) = 3,841 < \chi^2 \implies \text{elvetjük } H_0\text{-t.}$$

## Kolmogorov-Szmirnov próba (tisztá illeszkedésvizsgálat)

$$H_0 : P(X < x) = F_0(x) \quad \forall x \in \mathbb{R}.$$

- ▶ A próbastatisztika

$$\Delta_n = \sup_{x \in \mathbb{R}} |F_n^*(x) - F_0(x)|.$$

$H_0$  mellett  $\sqrt{n}\Delta_n$  aszimptotikusan Kolmogorov-eloszlású.

- ▶  $k_\alpha = K^{-1}(1 - \alpha)/\sqrt{n}$  és  $\mathcal{X}_k = \{\mathbf{x} : \Delta_n \geq k_\alpha\}$ .

### Megjegyzések

- ▶  $\Delta_n$  meghatározásához elég  $2n$  különbséget meghatározni, mivel  $F_0$  monoton növvő,  $F_n^*$  pedig lépcsős függvény.
- ▶ Egyoldali változatban a próbastatisztika abszolútérték nélküli és közelítőleg Szmirnov-eloszlású.

## Kétmintás Kolmogorov-Szmirnov próba (homogenitásteszt)

$$H_0 : P(X < x) = P(Y < x) \quad \forall x \in \mathbb{R}.$$

Tekintsük az  $X_1, \dots, X_{n_1}$  és  $Y_1, \dots, Y_{n_2}$  mintákat és a hozzájuk tartozó  $F_{n_1}^*$  és  $G_{n_2}^*$  empirikus eloszlásfüggvényeket.

- ▶ A próbastatisztika

$$\Delta_{n_1, n_2} = \sup_{x \in \mathbb{R}} |F_{n_1}^*(x) - G_{n_2}^*(x)|.$$

$H_0$  mellett  $\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \Delta_{n_1, n_2}$  közelítőleg Kolmogorov-eloszlású.

- ▶ Tehát  $k_\alpha = \sqrt{\frac{n_1 + n_2}{n_1 n_2}} K^{-1}(1 - \alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : \Delta_n \geq k_\alpha\}$ .

Az illeszkedésvizsgálat megjegyzései itt is érvényesek.

## Wilcoxon-próba

Legyen  $X$  folytonos, a mediánra szimmetrikus eloszlású vv.

$$H_0 : P(X < m_0) = \frac{1}{2}.$$

- ▶ Rendezzük sorba  $|X_1 - m_0|, \dots, |X_n - m_0|$  mintát és legyen  $R_i$  rangszám  $|X_i - m_0|$  sorszáma ebben a rendezett mintában.
- ▶ Legyen  $T^+ = \sum_{i: X_i > m_0} R_i$ .  $H_0$  mellett
  - ▶ kis  $n$  esetén  $T^+$  eloszlásának  $t_{\alpha, n}$  kvantiliseit táblázat tartalmazza.
  - ▶ Nagy  $n$ -re  $\frac{T^+ - E(T^+)}{D(T^+)} \sim \mathcal{N}(0, 1)$  közelítőleg,  
 $E(T^+) = \frac{1}{2} \sum_{i=1}^n i = \frac{n(n+1)}{4}$  és  $D^2(T^+) = \frac{n(n+1)(2n+1)}{24}$ .
- ▶ A kritikus tartomány
  - ▶  $\mathcal{X}_k = \{\mathbf{x} : t_{\alpha/2, n} \geq T^+\} \cup \{\mathbf{x} : T^+ \geq \frac{n(n+1)}{2} - t_{\alpha/2, n}\},$
  - ▶ nagy  $n$ -re  $\mathcal{X}_k = \left\{ \mathbf{x} : \left| \frac{T^+ - E(T^+)}{D(T^+)} \right| \geq \Phi^{-1}(1 - \alpha/2) \right\}.$

## Páros Wilcoxon-próba

$X, Y$  folytonos eloszlásúak. Önkontrollos vizsgálatot alkalmazunk a következő hipotézis vizsgálatára:

$$H_0 : P(X < Y) = \frac{1}{2}$$

Tekintsük a  $(X_1, Y_1), \dots, (X_n, Y_n)$  mintát. Legyen  $Z_i = X_i - Y_i$ , így

$$H_0 : P(Z < 0) = \frac{1}{2},$$

amit egymintás Wilcoxon-próbával tesztelhetünk, ha  $Z$  eloszlása folytonos és a mediánjára szimmetrikus.

## Megjegyzés

A páros t-próba nemparaméteres alternatívájaként alkalmazzák.  $H_0$  értelmezése okozhat problémát (a mediánnak nincs a várható értékhez hasonló additív tulajdonsága).

## Kétmintás Wilcoxon (Mann-Whitney) próba

$Y_i = X_i, i = 1, \dots, n_1$  és  $Z_j = X_j + \theta, j = n_1 + 1, \dots, n_1 + n_2$   
minták, ahol  $X_1, \dots, X_{n_1+n_2}$  független, azonos folytonos eloszlású  
változók.

$$H_0 : \theta = 0.$$

- ▶ Rendezzük sorba az egyesített mintát és legyen  $R_i$  a  $Y_i$  rangja az egyesített mintában.
- ▶  $R_Y = \sum_{i=1}^{n_1} R_i$ .  $H_0$  mellett
  - ▶ kis elemszám esetén  $R_Y$  eloszlásának  $r_{\alpha, n_1, n_2}$  kvantiliseit táblázat tartalmazza.
  - ▶ Nagy elemszámnál  $\frac{R_Y - E(R_Y)}{D(R_Y)} \sim \mathcal{N}(0, 1)$  közelítőleg, ahol  $E(R_Y) = \frac{n_1(n_1+n_2+1)}{2}$  és  $D^2(R_Y) = \frac{n_1 n_2 (n_1+n_2+1)}{12}$ .
- ▶ A kritikus tartomány
  - ▶  $\mathcal{X}_k = \mathcal{X} - \{ \mathbf{x} : r_{\alpha/2, n_1, n_2} < R_Y < n_1(n_1 + n_2 + 1) - r_{\alpha/2, n_1, n_2} \},$
  - ▶ nagy elemszáma  $\mathcal{X}_k = \left\{ \mathbf{x} : \left| \frac{R_Y - E(R_Y)}{D(R_Y)} \right| \geq \Phi^{-1}(1 - \alpha/2) \right\}.$

## Kruskal-Wallis teszt

Adott  $r$  független minta, az  $i$ . minta elemszáma  $n_i$ , eloszlásfüggvénye  $F_i$  folytonos, az eloszlások legfeljebb eltolási paraméterben térnek el.

$$H_0 : F_1(x) = \dots = F_r(x), \quad \forall x \in \mathbb{R}$$

Vegyük az egyesített mintabeli rangokat legyen  $R_i$  az  $i$ . mintabeli elemek rangösszege és  $N = \sum_{i=1}^r n_i$ .

$$H = \frac{12}{N(N+1)} \sum_{i=1}^r \frac{1}{n_i} \left( R_i - \frac{n_i(N+1)}{2} \right)^2$$

$H_0$  mellett  $H \xrightarrow{\mathcal{D}} \chi^2(r-1)$ , ha  $n_1, \dots, n_r \rightarrow \infty$ , így

$$\mathcal{X}_k = \left\{ \mathbf{x} : H \geq F_{\chi^2, r-1}^{-1}(1 - \alpha) \right\}.$$

## Hotelling-féle $T^2$ próba

Legyen  $\mathbf{X} \sim \mathcal{N}_d(\mathbf{m}, \mathbf{C})$  és tekintsük az  $\mathbf{X}_1, \dots, \mathbf{X}_n$  mintát ( $n > d$ ).

$$H_0 : \mathbf{m} = \mathbf{m}_0, \quad H_1 : \mathbf{m} \neq \mathbf{m}_0$$

►  $T^2 = n(\bar{\mathbf{X}} - \mathbf{m}_0)^\top \hat{\mathbf{C}}^{*-1}(\bar{\mathbf{X}} - \mathbf{m}_0)$ , ahol

$$\hat{\mathbf{C}}^* = \frac{1}{n-1} \mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{X}})(\mathbf{x}_i - \bar{\mathbf{X}})^\top$$

a **korrigált empirikus kovarianciamátrix**.  $H_0$  mellett

$$F = \frac{n-d}{d} \cdot \frac{T^2}{n-1} \sim \mathcal{F}(d, n-d).$$

►  $F_\alpha = F_{d, n-d}^{-1}(1-\alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : F \geq F_\alpha\}$ .

## Kétmintás Hotelling teszt

Legyen  $\mathbf{X} \sim \mathcal{N}_d(\mathbf{m}_1, \mathbf{C})$  és  $\mathbf{Y} \sim \mathcal{N}_d(\mathbf{m}_2, \mathbf{C})$ . Tekintsük az  $\mathbf{X}_1, \dots, \mathbf{X}_{n_1}$  és  $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_2}$  független mintákat ( $n_1, n_2 > d$ ).

$$H_0 : \mathbf{m}_1 = \mathbf{m}_2, \quad H_1 : \mathbf{m}_1 \neq \mathbf{m}_2$$

$$\blacktriangleright T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^\top \hat{\mathbf{C}}^{*-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}), \text{ ahol } \hat{\mathbf{C}}^* = \frac{1}{n_1 + n_2 - 2} \mathbf{S},$$

$$\mathbf{S} = \sum_{i=1}^{n_1} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top + \sum_{i=1}^{n_2} (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^\top.$$

$H_0$  mellett

$$F = \frac{n_1 + n_2 - d - 1}{d} \cdot \frac{T^2}{n_1 + n_2 - 2} \sim \mathcal{F}(d, n_1 + n_2 - d - 1).$$

$$\blacktriangleright F_\alpha = F_{d, n_1 + n_2 - d - 1}^{-1}(1 - \alpha) \text{ és } \mathcal{X}_k = \{\mathbf{x} : F \geq F_\alpha\}.$$

# Egyszempontos varianciaanalízis (ANOVA)

Tekintsük az  $X_{i,j} \sim \mathcal{N}(\mu_i, \sigma^2)$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, n_i$  független mintákat. A várható értékek felírhatók a következő formában:  $\mu_i = \mu + a_i$ , ahol  $\mu$  a várható értékek  $n_i$  értékekkel súlyozott átlaga,  $a_i$  pedig az  $i$ -edik csoporthatás.

$$H_0 : \mu_1 = \dots = \mu_r,$$

azaz minden csoporthatás 0.

Legyen  $n = n_1 + \dots + n_r$ . Képezzük a következő statisztikákat:

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{i,j}}{n_i}, \quad \bar{X} = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} X_{i,j}}{n}, \quad Q_t = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{i,j} - \bar{X})^2,$$
$$Q_a = \sum_{i=1}^r n_i (\bar{X}_i - \bar{X})^2, \quad Q_e = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2,$$

$Q_a$  a csoportok közötti,  $Q_e$  a csoporton belüli eltérések négyzetösszege.

## Tétel (Fisher-Cochran)

Legyen  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$  és  $Q = \mathbf{X}^\top \mathbf{X}$  és  $Q_j = \mathbf{X}^\top \mathbf{A}_j \mathbf{X}$ ,  $j = 1, \dots, k$  kvadratikus alakok, ahol  $\mathbf{A}_j$ -k szimmetrikus mátrixok és tfh.  $Q = Q_1 + \dots + Q_k$  teljesül.  $Q_1 \sim \chi^2(\text{rang}(\mathbf{A}_1)), \dots, Q_k \sim \chi^2(\text{rang}(\mathbf{A}_k))$  és függetlenek pontosan akkor, ha  $\sum_{j=1}^k \text{rang}(\mathbf{A}_j) = n$ .

## Tétel

$Q_t = Q_a + Q_e$ , továbbá  $H_0$  mellett  $Q_e/\sigma^2 \sim \chi^2(n-r)$  és  $Q_a/\sigma^2 \sim \chi^2(r-1)$  függetlenek.

## Bizonyítás

A nevezetes statisztikákra vonatkozó tétel miatt  $H_0$  mellett  $Q_t/\sigma^2 \sim \chi^2(n-1)$  és  $\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2/\sigma^2 \sim \chi^2(n_i-1)$ , amiből következik, hogy  $Q_e/\sigma^2 \sim \chi^2(n-r)$ , ahol  $n = n_1 + \dots + n_r$ .

Továbbá Steiner tétele miatt

$$\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 - n_i(\bar{X}_i - \bar{X})^2,$$

ebből  $Q_t = Q_a + Q_e$ , majd alkalmazzuk a Fisher-Cochran tételt.

## $H_0$ vizsgálata

- ▶ a próbastatisztika  $F = \frac{Q_a}{Q_e} \cdot \frac{n-r}{r-1} \sim \mathcal{F}(r-1, n-r)$   $H_0$  mellett.
- ▶ Tehát  $F_\alpha = F_{r-1, n-r}^{-1}(1 - \alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : F \geq F_\alpha\}$ .

## Megjegyzések

- ▶ ANOVA a kétmintás t-próba általánosítása.
- ▶ A szórások egyenlősége Levene, Hartly vagy Bartlett próbával tesztelhető. Ha a szórások nem egyenlőek, akkor ANOVA helyett a többmintás Welch-próbát alkalmazzák (hasonló problémák adódnak mint a kétmintás t próba esetén).
- ▶ PostHoc tesztek
- ▶ Ha a csoportokat több csoportosító változó segítségével képezzük, többszemponos varianciaanalízist alkalmazunk.
- ▶ Függetlenség tesztelésére is alkalmazzák.

# Lineáris regresszió

## Emlékeztető

Tekintsük a  $(X, Y)$  véletlen vektort. Keressük  $Y$ -t legjobban közelítő lineáris függvényt, azaz  $a$ -t és  $b$ -t, amelyre

$E((Y - (aX + b))^2)$  négyzetes eltérés minimális:

$$a = \frac{\text{Cov}(X, Y)}{D^2(X)} = r(X, Y) \frac{D(Y)}{D(X)},$$

$$b = E(Y) + \frac{\text{Cov}(X, Y)}{D^2(X)} E(X).$$

## Definíció

A fenti paraméterekkel felírt  $y = ax + b$  egyenes a **regressziós egyenes**,  $a, b$  paraméterek a **lineáris regressziós együtthatók**.

## Megjegyzés

$a = 0$  pontosan akkor, ha  $X$  és  $Y$  korrelálatlanok.

## Determinisztikus változók esete

Tekintsük a  $(x_1, Y_1), \dots, (x_n, Y_n)$  mintát, ahol  $x_i$  értékek determinisztikusak és az  $(x_1, y_1), \dots, (x_n, y_n)$  realizációt. Legyen  $Y_i = ax_i + b + \varepsilon_i$ , ahol  $\varepsilon_1, \dots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$  függetlenek, eloszlásuk nem függ sem  $a$ -tól, sem  $b$ -től, sem  $x_i$  értékektől. Tehát  $Y_1, \dots, Y_n$  függetlenek,  $Y_i \sim \mathcal{N}(ax_i + b, \sigma^2)$ .

## Regressziós egyenes paramétereinek ML becslése

Jelölje  $Y_i$  sűrűségfüggvényét  $f_i$

$$L(a, b) = \prod_{i=1}^n f_i(y_i) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - ax_i - b)^2}.$$

$L(a, b)$  akkor maximális, ha  $Q(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$  minimális (legkisebb négyzetes becslés).

$Q(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$  mindkét változó szerint deriválható, a deriváltak zérushelyei  $Q(a, b)$  lehetséges szélsőértékhelyei.

$$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^n x_i (y_i - ax_i - b) = 0$$

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^n (y_i - ax_i - b) = 0$$

A lehetséges szélsőértékhelyek teljesítik a következő egyenlőséget

$$n \frac{\partial Q}{\partial a} - \sum_{i=1}^n x_i \frac{\partial Q}{\partial b} = 0$$

$$n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i - \hat{a} \left( n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \right) = 0.$$

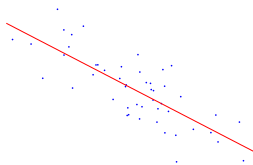
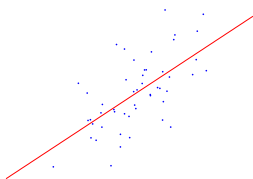
Ebből

$$\begin{aligned}\hat{a} &= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ &= \frac{n^2 \hat{C}}{n^2 S_x^2} = r_n(x, Y) \frac{S_Y}{S_x}.\end{aligned}$$

$\hat{a}$ -t beírva  $\frac{\partial Q}{\partial b} = \sum_{i=1}^n y_i - a \sum_{i=1}^n x_i - nb = 0$  egyenletbe

$$\hat{b} = \frac{\sum_{i=1}^n y_i}{n} - \hat{a} \frac{\sum_{i=1}^n x_i}{n}.$$

Belátható, hogy  $(\hat{a}, \hat{b})$  valóban  $Q$  minimumhelye.



## Megjegyzések

- ▶ A legkisebb négyzetek módszere más modelleken is alkalmazható.
- ▶ A regressziót gyakran  $Y_{n+1}$  érték becslésére használják, ha csak a független változók adottak. (Vigyázat: extrapoláció!)
- ▶ Tipikus hiba: kapcsolat mögött okságot feltételezni nem determinisztikus változók esetén.
- ▶ A regressziós egyenes becslése érzékeny a kiugró adatokra.

## Varianciaanalízis lineáris modellben

Tekintsük a  $Y = ax + b + \varepsilon$ ,  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$  regressziós modellt és az  $(x_1, Y_1), \dots, (x_n, Y_n)$  mintát.

$$H_0 : a = 0 \quad H_1 : a \neq 0.$$

Vegyük a regressziós egyenes paramétereinek  $\hat{a}, \hat{b}$  ML becsléseit és legyen  $\hat{Y}_i = \hat{a}x_i + \hat{b}$ . Képezzük a következő négyzetösszegeket:

$$Q_t = \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad Q_r = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \quad Q_e = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

### Tétel

$Q_t = Q_r + Q_e$ , továbbá  $H_0$  mellett  $Q_e/\sigma^2 \sim \chi^2(n-2)$  és  $Q_r/\sigma^2 \sim \chi^2(1)$  függetlenek.

### Bizonyítás vázlat

A nevezetes statisztikák eloszlása, a Steiner és a Fisher-Cochran tételek alkalmazásával.

## $H_0$ vizsgálata

- ▶ a próbastatisztika  $F = \frac{Q_r}{Q_e}(n-2)$ , ami  $H_0$  mellett  $\mathcal{F}(1, n-2)$  eloszlású.
- ▶  $F_\alpha = F_{1, n-2}^{-1}(1-\alpha)$  és  $\mathcal{X}_k = \{\mathbf{x} : F \geq F_\alpha\}$ .

## Állítás

$$Q_r = Q_t \cdot r_n^2(x, Y).$$

## Bizonyítás

$$\sum_{i=1}^n (\hat{a}x_i - \hat{b} - \bar{Y})^2 = \hat{a}^2 \sum_{i=1}^n (x_i - \bar{x})^2 = r_n^2(x, Y) \frac{S_Y^2}{S_x^2} n S_x^2.$$

## Következmény

$$Q_e = Q_t(1 - r_n^2(x, Y)) \quad \text{és} \quad \frac{Q_r}{Q_e} = \frac{r_n^2(x, Y)}{1 - r_n^2(x, Y)}.$$

## Pearson

A háttérváltozó  $(X, Y)^\top \sim \mathcal{N}_2(\mathbf{m}, \mathbf{C})$ .  $H_0 : X, Y$  fgn.

$t = \sqrt{n-2} \frac{r_n(X, Y)}{\sqrt{1-r_n^2(X, Y)}} \xrightarrow{\mathcal{D}} t(n-2)$   $H_0$  mellett, ha  $n \rightarrow \infty$  tehát

$$\mathcal{X}_k = \{\mathbf{x} : |t| \geq \Phi_{n-2}^{-1}(1 - \alpha/2)\}.$$

## Spearman

$(X, Y)^\top$  abszolút folytonos háttérváltozó.  $H_0 : X, Y$  fgn.

Tekintsük az  $(X_1, Y_1), \dots, (X_n, Y_n)$  mintát, legyen  $R_i$  az  $X_i$  és  $S_i$  az  $Y_i$  rangszáma. A **Spearman-féle rangkorrelációs együttható**

$$\rho_n = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^n (S_i - \bar{S})^2}}.$$

$\sqrt{n-1} \rho_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$   $H_0$  mellett, ha  $n \rightarrow \infty$ , így

$$\mathcal{X}_k = \{\mathbf{x} : |\sqrt{n-1} \rho_n| \geq \Phi^{-1}(1 - \alpha/2)\}.$$

# Többváltozós regresszió

$\mathbf{X}$  egy  $d$  dimenziós **magyarázó** változó, a regressziós függvény  $E((Y - g(\mathbf{X}))^2)$ -t minimalizálja. Ha  $Y$  és  $\mathbf{X}$  együtt többdimenziós normális eloszlású, a legjobb regressziós függvény a lineáris.

## Definíció

$r_Y(\mathbf{x}) = r(\ell(\mathbf{X}), Y)$  **a többszörös korrelációs együttható**, ahol  $\ell(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$  **a lineáris regressziós függvény**.

## Lineáris regresszió (determinisztikus változók esete)

A modell  $Y_i = \mathbf{a}^\top \mathbf{x}_i + b + \varepsilon$ ,  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ .

$$H_0 : \mathbf{a} = \mathbf{0}.$$

- ▶ Vesszük  $\mathbf{a}$ ,  $b$  legkisebb négyzetes becsléseit,  $\hat{Y}_i = \hat{\mathbf{a}}^\top \mathbf{x}_i + \hat{b}$  és képezzük  $Q_t$ ,  $Q_r$  és  $Q_e$  négyzetösszegeket.
- ▶ A próbastatisztika most  $F = \frac{Q_r}{Q_e} \cdot \frac{n - d - 1}{d}$ , ami  $H_0$  mellett  $\mathcal{F}(d, n - d - 1)$  eloszlású.

## Megjegyzések

- ▶ A többszörös korrelációs együttható nem általánosítása a korrelációs együtthatónak (de négyzetükre igaz).
- ▶  $\mathbf{X}$  komponenseit **független**,  $Y$ -t **függő változó**nak nevezik.
- ▶ A többszörös korrelációs együttható értelmezése:  $\mathbf{X}$  komponensei  $Y$  varianciájának  $r_{Y(\mathbf{X})}^2$  részét magyarázzák.  $r_{Y(\mathbf{X})}^2$ -t **meghatározottsági együttható**nak is nevezik.
- ▶ Többdimenziós normális eloszlású változók esetén is alkalmazható a varianciaanalízis lineáris modellben.
- ▶  $\frac{Q_r}{Q_e} = \frac{R^2}{1-R^2}$ , ahol  $R$  a becsült többszörös korrelációs együttható.  $H_0: r_{Y(\mathbf{X})} = 0$ -val ekvivalens.
- ▶ A nemlineáris regressziót gyakran a lineárisra vezetik vissza.

## Ajánlott irodalom

- [1] Bolla Marianna, Krámlí András: *Statisztikai következtetések elmélete*, Typotex
- [2] Bolla Marianna, Krámlí András, Nagy-György Judit: *Többváltozós statisztikai módszerek (digitális jegyzet)*, Tankönyvtár
- [3] Fazekas István (szerk.): *Bevezetés a matematikai statisztikába*, Kossuth Egyetemi Kiadó
- [4] Móri F. Tamás, Szeidl László, Zempléni András: *Matematikai statisztika példatár*, ELTE Eötvös Kiadó
- [5] Nagy-György Judit, Osztényiné Krauczi Éva, Székely László: *Valószínűesszámitás és statisztika példatár*, Polygon Jegyzettár
- [6] Viharos László: *A sztochasztika alapjai*, Polygon Jegyzettár