

Bevezetés a matematikai statisztikába

Nagy-György Judit

Szegedi Tudományegyetem, Bolyai Intézet

Tekintsünk egy ξ valószínűségi változót.

Statisztikai minta (n elemű minta) $\xi_1, \dots, \xi_n$ fae vv, eloszlásuk megegyezik a ξ **háttérváltozó** eloszlásával.

Ha ξ_i változó az x_i értéket veszi fel ($i = 1, \dots, n$), akkor azt mondjuk, hogy $x_1, \dots, x_n$ a minta **realizációja**.

A matematikai statisztika alapvető feladatai:

- ▶ ξ háttérváltozó eloszlásának, egyéb mutatóinak **becslése** (pontbecslések, intervallumbecslések),
- ▶ ξ háttérváltozó eloszlására vonatkozó **hipotézisek vizsgálata** (statisztikai próbák).

Definíció

Legyen f egy n -változós függvény. A mintaelemek $f(\xi_1, \dots, \xi_n)$ függvényét **statisztikának** nevezzük.

Alapstatisztikák

- ▶ $\bar{\xi} = \frac{\xi_1 + \dots + \xi_n}{n}$ mintaátlag,
- ▶ $\min_i \{\xi_i\}$ legkisebb mintaelem,
- ▶ $\max_i \{\xi_i\}$ legnagyobb mintaelem,
- ▶ $\max_i \{\xi_i\} - \min_i \{\xi_i\}$ mintaterjedelem,
- ▶ empirikus (tapasztalati) medián: a sorbarendezett mintaelemek közül a középső (vagy ha n páros, a középső kettő átlaga).
- ▶ ...

Legyen θ a ξ eloszlásának egy paramétere, $\hat{\theta}_n := \hat{\theta}(\xi_1, \dots, \xi_n)$.

Definíció

Az $\hat{\theta}_n$ statisztika **torzítatlan** becslése θ -nak, ha

$$E(\hat{\theta}_n) = \theta.$$

Definíció

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

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

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

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

Példa

Legyen $\xi_1, \dots, \xi_n$ minta, ahol a ξ háttérváltozónak létezik a σ szórása. Legyen $\theta = E(\xi)$ ismeretlen paraméter.

Tekintsük a ξ_1 és a $\bar{\xi}$ statisztikákat.

- ▶ Nyilvánvalóan ξ_1 még gyengén sem konzisztens becslése θ -nak, míg a nagy számok erős törvénye miatt $\bar{\xi}$ erősen konzisztens becslés.

- ▶ Mindkét statisztika torzítatlan becslése θ -nak:

$$E(\xi_1) = E(\xi) = \theta,$$

$$E(\bar{\xi}) = \frac{1}{n}E(\xi_1 + \dots + \xi_n) = \frac{1}{n} \cdot n \cdot E(\xi) = \theta.$$

- ▶ $D^2(\bar{\xi}) = \frac{1}{n^2}D^2(\xi_1 + \dots + \xi_n) = \frac{n}{n^2} \cdot D^2(\xi) = \frac{\sigma^2}{n} < \sigma^2 = D^2(\xi_1).$

Vagyis a két torzítatlan becslés közül $\bar{\xi}$ a **hatásosabb**.

Definíció

Legyen $k_{n,x} = \sum_{i=1}^n I(\xi_i < x)$, ahol I indikátorfüggvény és

$$F_n(x) = \frac{k_{n,x}}{n}.$$

Az F_n függvényt *empirikus eloszlásfüggvény*nek nevezzük.

F_n tulajdonságai

A minta bármely realizációját tekintve F_n

- ▶ monoton nemcsökkenő,
- ▶ balról folytonos,
- ▶ $\lim_{n \rightarrow \infty} F_n(x) = 1$ és $\lim_{n \rightarrow -\infty} F_n(x) = 0$.

Állítás

$x \in \mathbb{R}$ rögzített. $F_n(x)$ torzítatlan és erősen konzisztens becslése $F(x)$ -nek.

Bizonyítás.

$k_{x,n}$ binomiális eloszlású n és $F(x)$ paraméterekkel. Tehát

$$E(F_n(x)) = E\left(\frac{k_{x,n}}{n}\right) = \frac{1}{n}E(k_{x,n}) = \frac{1}{n} \cdot nF(x) = F(x).$$

Az erős konzisztencia nagy számok erős törvényéből közvetlenül következik.

Tétel (A matematikai statisztika alaptétele; Glivenco–Cantelli)

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

A sűrűségfüggvény becslése

Definíció

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

$$f_n(x) = \frac{\nu_k}{n|I_k|}, \quad \text{ha } x \in I_k.$$

Az f_n függvényt **sűrűséghisztogramnak** nevezzük.

f_n tulajdonságai

A minta bármely realizációját tekintve

- ▶ $f_n \geq 0$,
- ▶ $\int_{-\infty}^{\infty} f_n(x) dx = 1$.

Megjegyzés

Általában egyenlő hosszú az intervallumokra osztják $\mathbb{R}$ -t.

A várható érték becslése

Definíció

Az $E_n(\xi) = \bar{\xi}$ statisztikát **empirikus (tapasztalati) várható érték**nek nevezzük.

Állítás

$E_n(\xi)$ torzítatlan, és ha $E(\xi^2) < \infty$, akkor erősen konzisztens becslése is $E(\xi)$ -nek.

Bizonyítás

A várható érték additivitása miatt

$$E(E_n(\xi)) = E\left(\frac{\xi_1 + \dots + \xi_n}{n}\right) = \frac{E(\xi_1) + \dots + E(\xi_n)}{n} = \frac{1}{n}nE(\xi).$$

Az állítás második fele éppen a nagy számok erős törvénye.

Definíció

Empirikus (tapasztalati) variancia:

$$V_n(\xi) = \frac{1}{n} \sum_{i=1}^n (\xi_i - E_n(\xi))^2$$

Empirikus (tapasztalati) szórás: $D_n(\xi) = \sqrt{V_n(\xi)}$

Állítás

$$V_n(\xi) = \frac{1}{n} \sum_{i=1}^n (\xi_i^2) - E_n^2(\xi) = E_n(\xi_i^2) - E_n^2(\xi).$$

Bizonyítás

$$\begin{aligned} V_n(\xi) &= \frac{1}{n} \sum_{i=1}^n \xi_i^2 - 2E_n(\xi) \frac{1}{n} \sum_{i=1}^n \xi_i + \frac{1}{n} \cdot nE_n^2(\xi) \\ &= E_n(\xi_i^2) - E_n^2(\xi). \end{aligned}$$

Tétel

Ha $D(\xi)$ létezik, akkor $E(V_n(\xi)) = \frac{n-1}{n}D^2(\xi)$.

Bizonyítás

$$\begin{aligned}E(V_n(\xi)) &= E\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^n \xi_i\right)^2\right) \\&= \frac{1}{n}\sum_{i=1}^n E(\xi_i^2) - \frac{1}{n^2}\left(\sum_{i=1}^n E(\xi_i^2) + 2\sum_{i<j} E(\xi_i\xi_j)\right) \\&= \frac{n-1}{n^2}\sum_{i=1}^n E(\xi_i^2) - \frac{2}{n^2}\sum_{i<j} E(\xi_i)E(\xi_j) \\&= \frac{n-1}{n^2} \cdot n \cdot E(\xi^2) - \frac{2}{n^2} \cdot \frac{n(n-1)}{2} E^2(\xi) \\&= \frac{n-1}{n}E(\xi^2) - \frac{n-1}{n}E(\xi)^2.\end{aligned}$$

Definíció

Korrigált empirikus (tapasztalati) variancia:

$$V_n^*(\xi) = \frac{n}{n-1} V_n(\xi) = \frac{1}{n-1} \sum_{i=1}^n \xi_i^2 - E_n^2(\xi)$$

Korrigált empirikus (tapasztalati) szórás: $D_n^*(\xi) = \sqrt{V_n^*(\xi)}$

Következmény

$V_n^*(\xi)$ torzítatlan becslése $D^2(\xi)$ -nek.

Megjegyzés

Az viszont nem következik ebből, hogy $D_n^*(\xi)$ torzítatlan becslése lenne $D(\xi)$ -nek.

Tétel

Ha $D(\xi)$ létezik, akkor $V_n(\xi)$ és $V_n^*(\xi)$ is erősen konzisztens becslése $D^2(\xi)$ -nek.

Bizonyítás

A nagy számok erős törvény alapján

$$E_n(\xi) = \frac{\xi_1 + \dots + \xi_n}{n} \xrightarrow{mb} E(\xi), \quad n \rightarrow \infty$$

valamint

$$E_n(\xi^2) = \frac{\xi_1^2 + \dots + \xi_n^2}{n} \xrightarrow{mb} E(\xi^2), \quad n \rightarrow \infty.$$

Így

$$V_n(\xi) = E_n(\xi^2) - E_n^2(\xi) \xrightarrow{mb} E(\xi^2) - E^2(\xi) = D^2(\xi), \quad n \rightarrow \infty.$$

Tekintsük (ξ, η) háttérváltozót és $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ mintát.

Definíció

ξ és η empirikus kovarianciája

$$C_n(\xi, \eta) = \frac{1}{n} \sum_{i=1}^n (\xi_i - E_n(\xi))(\eta_i - E_n(\eta)).$$

Állítás

$$C_n(\xi, \eta) = \frac{1}{n} \sum_{i=1}^n \xi_i \eta_i - E_n(\xi)E_n(\eta)$$

Bizonyítás

$$\begin{aligned} C_n(\xi, \eta) &= \frac{1}{n} \left(\sum_{i=1}^n \xi_i \eta_i - E_n(\xi) \sum_{i=1}^n \eta_i - E_n(\eta) \sum_{i=1}^n \xi_i + n E_n(\xi) E_n(\eta) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \xi_i \eta_i - 2 E_n(\xi) E_n(\eta) + E_n(\xi) E_n(\eta). \end{aligned}$$

Definíció

ξ és η (Pearson-féle) empirikus korrelációs együtthatója

$$r_n(\xi, \eta) = \frac{C_n(\xi, \eta)}{D_n(\xi)D_n(\eta)}, \text{ ha } D_n(\xi)D_n(\eta) \neq 0, \text{ és } 0 \text{ különben.}$$

Állítás

$|r_n(\xi, \eta)| \leq 1$. Továbbá $|r_n(\xi, \eta)| = 1$ pontosan akkor teljesül, ha ξ_i, η_i pontpárok között lineáris összefüggés van.

Bizonyítás

A Cauchy-Bunyakovszkij-Schwarz egyenlőtlenség alapján

$$\begin{aligned} |nC_n(\xi, \eta)| &\leq \sum_{i=1}^n |\xi_i - E_n(\xi)| \cdot |\eta_i - E_n(\eta)| \\ &\leq \sqrt{\sum_{i=1}^n (\xi_i - E_n(\xi))^2 \sum_{i=1}^n (\eta_i - E_n(\eta))^2} \\ &= \sqrt{nV_n(\xi) \cdot nV_n(\eta)} = nD_n(\xi)D_n(\eta). \end{aligned}$$

Megjegyzések

- ▶ $C_n(\xi, \xi) = V_n(\xi)$.
- ▶ $C_n(\xi, \eta)$ valamint $r_n(\xi, \eta)$ előjele a kapcsolat irányára utal.
- ▶ $|r_n(\xi, \eta)|$ a kapcsolat szorosságát jelzi. Ha $r_n(\xi, \eta) = 1$, akkor pozitív, ha $r_n(\xi, \eta) = -1$, akkor negatív lineáris kapcsolat.

Állítás

$C_n(\xi, \eta)$ erősen konzisztens becslése $C(\xi, \eta)$ -nak, valamint $r_n(\xi, \eta)$ erősen konzisztens becslése $r(\xi, \eta)$ -nak.

Bizonyítás

A nagy számok erős törvénye alapján egyszerűen adódik:

$$E_n(\xi\eta) - E_n(\xi)E_n(\eta) \xrightarrow{mb} E(\xi\eta) - E(\xi)E(\eta) = C(\xi, \eta), \quad n \rightarrow \infty,$$

$$\frac{C_n(\xi, \eta)}{\sqrt{V_n(\xi)V_n(\eta)}} \xrightarrow{mb} \frac{C(\xi, \eta)}{D(\xi)D(\eta)}, \quad n \rightarrow \infty.$$

Legyen θ ismeretlen paraméter, erre szeretnénk becslést kapni.

Definíció

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

$$L(\theta) = P_{\theta}(\xi_1 = x_1, \dots, \xi_n = x_n) = P_{\theta}(\xi = x_1) \cdot \dots \cdot P_{\theta}(\xi = x_n).$$

Definíció

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

$$L(\theta) = f_{\theta, \xi_1, \dots, \xi_n}(x_1, \dots, x_n) = f_{\theta}(x_1) \cdot \dots \cdot f_{\theta}(x_n).$$

Definíció

A θ paraméter **maximum-likelihood becslése** (röviden ML becslése) $\hat{\theta}$, ha

$$L(\hat{\theta}) = \max_{\theta} L(\theta).$$

A szorzat maximumhelyének meghatározása helyett sokszor könnyebb egy összeg maximumhelyét megadni, ezért vezetjük be a következő fogalmat.

Definíció

*Tekintsük az $x_1, \dots, x_n$ realizációt, θ paramétert, és a hozzájuk tartozó $L(\theta)$ likelihood-függvényt. A **log-likelihood függvény** $\ell(\theta) = \ln L(\theta)$.*

A (természetes alapú) logaritmus függvény monoton növekvő, ezért $\ell(\theta)$ éppen ott veszi fel szélsőértékeit, ahol $L(\theta)$.

Megjegyzés

ML becslést konkrét realizáció esetén kapunk, ha minden realizációra tekintjük, akkor kapunk becslést θ -ra.

Legyen A egy esemény $p = P(A)$ valószínűséggel, ahol $0 < p < 1$. Tekintsük $\xi = I(A)$ háttérváltozót (A indikátorváltozója), ez Bernoulli-eloszlású p paraméterrel.

Tegyük fel, hogy $x_1, \dots, x_n$, az ebből vett minta realizációja nem csupa 0 vagy csupa 1, és legyen $k_n(A) = \sum_{i=1}^n x_i$.

$$L(p) = p^{k_n(A)}(1-p)^{n-k_n(A)}, \text{ ha } p \in (0, 1)$$

tehát

$$\ell(p) = k_n(A) \ln p + (n - k_n(A)) \ln(1 - p)$$

maximumhelyét keressük $(0, 1)$ -en.

$\ell(p)$ a $(0, 1)$ -en kétszer deriválható, így

$$\begin{aligned}\ell'(\hat{p}) &= \frac{k_n(A)}{\hat{p}} - \frac{n - k_n(A)}{1 - \hat{p}} = 0, \\ \ell''(\hat{p}) &= \frac{-k_n(A)}{\hat{p}^2} - \frac{n - k_n(A)}{(1 - \hat{p})^2} < 0\end{aligned}$$

megoldása $\hat{p} = k_n(A)/n$ lesz p ML becslése.

Megjegyzések

- ▶ Ha $p = 0$ vagy 1 , akkor $\ell(p)$ nem értelmezett.
- ▶ Ha a realizáció $1, \dots, 1$, akkor $L(p) = p^n$ ha $p \in [0, 1]$, maximumhelye 1 , de nem deriválható 1 -ben. $(0, 1)$ -en nincs szélsőértéke $L(p)$ -nek.
- ▶ Hasonló a helyzet $0, \dots, 0$ realizáció esetén, $L(p) = (1 - p)^n$ ha $p \in [0, 1]$.
- ▶ Ha $K_n(A)$ a bekövetkezések száma, $E(K_n(A)/n) = np/n = p$.

Másik lehetőség p ML becslésére, ha tekintünk egy $0 < p < 1$ paraméterű geometriai eloszlású ξ háttérváltozót. Ekkor

$$\begin{aligned}L(p) &= p(1-p)^{x_1-1} \cdot \dots \cdot p(1-p)^{x_n-1} \\&= p^n (1-p)^{\sum_{i=1}^n x_i - n} \\ \ell(p) &= n \ln p + \left(\sum_{i=1}^n x_i - n \right) \ln(1-p).\end{aligned}$$

$\ell(p)$ $(0, 1)$ -en kétszer deriválható, így

$$\begin{aligned}\ell'(\hat{p}) &= \frac{n}{\hat{p}} - \frac{\sum_{i=1}^n x_i - n}{1 - \hat{p}} = 0, \\ \ell''(\hat{p}) &= \frac{-n}{\hat{p}^2} - \frac{\sum_{i=1}^n x_i - n}{(1 - \hat{p})^2} < 0\end{aligned}$$

megoldása $\hat{p} = n / \sum_{i=1}^n x_i$. Belátható, hogy ez nem torzítatlan becslése p -nek.

Poisson eloszlás paraméterének ML becslése

$$L(\lambda) = \prod_{i=1}^n \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right) = \lambda^{\sum_{i=1}^n x_i} \cdot \left(\prod_{i=1}^n x_i! \right)^{-1} \cdot e^{-n\lambda}, \quad \lambda > 0.$$

tehát

$$\ell(\lambda) = \sum_{i=1}^n x_i \ln \lambda + c - n\lambda, \quad \lambda > 0$$

kétszer deriválható $(0, \infty)$ -en,

$$\ell'(\hat{\lambda}) = \frac{\sum_{i=1}^n x_i}{\hat{\lambda}} - n = 0, \quad \ell''(\hat{\lambda}) = -\frac{\sum_{i=1}^n x_i}{\hat{\lambda}^2} < 0$$

megoldása $\hat{\lambda} = \sum_{i=1}^n x_i / n$, ha $\sum_{i=1}^n x_i \neq 0$.

Ha $\sum_{i=1}^n x_i = 0$, akkor $L(\lambda) = e^{-n\lambda}$, ennek nincs szélsőértéke a $(0, \infty)$ intervallumon, így ekkor λ -nak nincs ML becslése.

Exponenciális eloszlás paraméterének ML becslése

$$L(\lambda) = \prod_{i=1}^n \left(\lambda e^{-\lambda x_i} \right) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}, \quad \lambda > 0.$$

tehát

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^n x_i, \quad \lambda > 0$$

kétszer deriválható $(0, \infty)$ -en,

$$\ell'(\hat{\lambda}) = \frac{n}{\hat{\lambda}} - \sum_{i=1}^n x_i = 0, \quad \ell''(\hat{\lambda}) = \frac{-n}{\hat{\lambda}^2} < 0$$

megoldása $\hat{\lambda} = n / \sum_{i=1}^n x_i$.

Belátható, hogy $1/E_n(\xi)$ nem torzítatlan becslése λ -nak.

Intervallum jobb végpontjának ML becslése

Legyen ξ háttérváltozó egyenletes eloszlású a $[0, \theta]$ intervallumon ($f(x) = \theta^{-1}$, ha $0 \leq x \leq \theta$ és 0 különben).

$$L(\theta) = \begin{cases} \theta^{-n}, & \text{ha } x_1, \dots, x_n \leq \theta, \\ 0 & \text{különben} \end{cases} = \begin{cases} \theta^{-n}, & \text{ha } \max_{i=1}^n x_i \leq \theta, \\ 0 & \text{különben.} \end{cases}$$

$L(\theta)$ monoton nő, ezért maximumhelye $\hat{\theta} = \max_{i=1}^n x_i$.

Megjegyzések

- ▶ Mivel $L(\theta)$ -nak a maximumhelye szakadási pont, deriválással nem lehet meghatározni $\hat{\theta}$ -t.
- ▶ Belátható, hogy $E(\max_{i=1}^n \xi_i) = \frac{n}{n+1}\theta$.
- ▶ Ha a $(0, \theta)$ intervallumból vennénk a mintát, akkor nem létezne ML becslés.

Intervallum helyének ML becslése

Legyen ξ háttérváltozó egyenletes eloszlású a $[\theta, \theta + 1]$ intervallumon ($f(x) = 1$, ha $\theta \leq x \leq \theta + 1$ és 0 különben).

$$\begin{aligned} L(\theta) &= \begin{cases} 1, & \text{ha } \theta \leq x_1, \dots, x_n \leq \theta + 1, \\ 0 & \text{különben} \end{cases} \\ &= \begin{cases} 1, & \text{ha } \max_{i=1}^n x_i - 1 \leq \theta \leq \min_{i=1}^n x_i, \\ 0 & \text{különben.} \end{cases} \end{aligned}$$

$L(\theta)$ maximumhelyei a $(\max_{i=1}^n x_i - 1, \min_{i=1}^n x_i)$ intervallum pontjai.

Megjegyzés

Belátható, hogy $E(\max_{i=1}^n \xi_i - 1) = \theta - 1/(n+1)$, valamint $E(\min_{i=1}^n \xi_i) = \theta + 1/(n+1)$, tehát a kettő átlaga torzítatlan becslés θ -ra.

Normális eloszlás paramétereinek ML becslése

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

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

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

ami kétszer deriválható. Képezzük a parciális deriváltakat:

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

$$\frac{\partial \ell}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0, \quad (2)$$

L lehetséges szélsőértékhelyei ezek közös zérushelyei.

(1)-ből kapjuk, hogy

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n},$$

ezt (2)-be helyettesítve pedig

$$\begin{aligned}\frac{n}{\hat{\sigma}} &= \frac{1}{\hat{\sigma}^3} \sum_{i=1}^n (x_i - \hat{\mu})^2 \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \\ \hat{\sigma} &= \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2}.\end{aligned}$$

Belátható, hogy $(\hat{\mu}, \hat{\sigma})$ valóban maximumhelye $L(\mu, \sigma)$ -nak.

Normális eloszlásból származtatott eloszlások

Legyenek $\xi_0, \dots, \xi_{m+n}$ független, standard normális eloszlású vv-k.

Definíció

- ▶ $\chi^2 = \xi_1^2 + \dots + \xi_n^2$ vv n szabadságfokú χ^2 eloszlású $F_{\chi^2, n}$ eloszlásfüggvénnyel.
- ▶ $\frac{\xi_0 \sqrt{n}}{\sqrt{\xi_1^2 + \dots + \xi_n^2}}$ vv n szabadságfokú Student (t) eloszlású Φ_n eloszlásfüggvénnyel.
- ▶ $\frac{n}{m} \cdot \frac{\xi_1^2 + \dots + \xi_m^2}{\xi_{m+1}^2 + \dots + \xi_{m+n}^2}$ vv (m, n) szabadságfokú F eloszlású $F_{m, n}$ eloszlásfüggvénnyel.

Megjegyzések

- ▶ Bizonyos paraméterértékek esetén a χ^2 , Student és F eloszlások inverz eloszlásfüggvényeinek néhány értékét statisztikai táblázatok tartalmazzák.
- ▶ A $t(n)$ eloszlás szimmetrikus: ha $\xi \sim t(n)$, akkor $-\xi \sim t(n)$.
- ▶ Belátható, hogy ha $\xi_n \sim t(n)$ és $\xi \sim N(0, 1)$, akkor $\xi_n \rightarrow \xi$ eloszlásban, ha $n \rightarrow \infty$ (vagyis $\Phi_n(x) \rightarrow \Phi(x)$, ha $n \rightarrow \infty$).
- ▶ Ha $\xi \sim F(n, m)$, akkor $1/\xi \sim F(m, n)$. Ebből

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

- ▶ Ha $\xi \sim t(n)$, akkor $\xi^2 \sim F(1, n)$.
- ▶ Ha $\xi \sim \chi^2(m)$, $\eta \sim \chi^2(n)$ függetlenek, akkor

$$\frac{\xi/m}{\eta/n} \sim F(m, n).$$

Tétel

Legyen $\xi_1, \dots, \xi_n$ egy $N(\mu, \sigma^2)$ eloszlásból vett minta. A következők teljesülnek:

- ▶ $E_n(\xi) \sim N\left(\mu, \frac{\sigma^2}{n}\right),$
- ▶ $\frac{n}{\sigma^2} V_n(\xi) \sim \chi^2(n-1),$
- ▶ $E_n(\xi)$ és $V_n(\xi)$ függetlenek,
- ▶ $\frac{E_n(\xi) - \mu}{\sqrt{V_n^*/n}} \sim t(n-1).$

Tétel

Legyen $\xi_1, \dots, \xi_{n_1}$ egy $N(\mu_1, \sigma_1^2)$, $\eta_1, \dots, \eta_{n_2}$ egy $N(\mu_2, \sigma_2^2)$ eloszlásból vett független minta. A következők teljesülnek:



$$\frac{E_{n_1}(\xi) - E_{n_2}(\eta) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

► ha $\sigma_1 = \sigma_2$, akkor

$$\frac{E_{n_1}(\xi) - E_{n_2}(\eta) - (\mu_1 - \mu_2)}{\sqrt{n_1 V_{n_1}(\xi) + n_2 V_{n_2}(\eta)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

Student eloszlású $n_1 + n_2 - 2$ szabadságfokkal,



$$\frac{V_{n_1}^*(\xi) \sigma_2^2}{V_{n_2}^*(\eta) \sigma_1^2} \sim F(n_1 - 1, n_2 - 1).$$

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

Definíció

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

A konfidencia-intervallum szerkesztése általában

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

Konfidencia intervallum normális eloszlás várható értékére

Legyen $\xi_1, \dots, \xi_n$ $N(\mu, \sigma^2)$ -ből vett minta. μ -re keresünk $1 - \alpha$ megbízhatósági szintű konfidencia intervallumot. Tudjuk, hogy $E_n(\xi) \sim N(\mu, \sigma^2/n)$. Válasszuk a $Z_n(\mu) = \frac{E_n(\xi) - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ statisztikát, és szerkesszünk köré intervallumot:

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

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

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

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

Most nézzük meg azt az esetet, ha σ ismeretlen. Legyen

$$Z_n(\mu) = \frac{E_n(\xi) - \mu}{\sqrt{V_n^*(\xi)/n}} \sim t(n-1).$$

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

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

Konfidencia intervallum normális eloszlás szórására

Legyen $Z_n(\sigma) = nV_n(\xi)/\sigma^2 \sim \chi^2(n-1)$.

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

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

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

Konf. intervallum normális eloszlások vé. különbségére

Legyenek $\xi_1, \dots, \xi_{n_1} \sim N(\mu_1, \sigma^2)$ és $\eta_1, \dots, \eta_{n_2} \sim N(\mu_2, \sigma^2)$ fgn. minták. $\mu_1 - \mu_2$ -re keresünk konfidencia intervallumot.

$$Z_{n_1, n_2} = \frac{E_{n_1}(\xi) - E_{n_2}(\eta) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim N(0, 1).$$

Mivel $P(-x < Z_{n_1, n_2} < x) = 2\Phi(x) - 1$, az $x_\alpha = \Phi^{-1}(1 - \alpha/2)$ választással kapjuk:

$$\begin{aligned} 1 - \alpha &= P \left(-x_\alpha < \frac{E_{n_1}(\xi) - E_{n_2}(\eta) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} < x_\alpha \right) \\ &= P \left(\bar{\xi} - \bar{\eta} - x_\alpha \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} < \mu_1 - \mu_2 < \bar{\xi} - \bar{\eta} + x_\alpha \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \right). \end{aligned}$$

Ha σ ismeretlen, akkor legyen

$$Z_{n_1, n_2} = \frac{E_{n_1}(\xi) - E_{n_2}(\eta) - (\mu_1 - \mu_2)}{D_*} \sim t(n_1 + n_2 - 2),$$

ahol

$$D_*^2 = (n_1 V_{n_1}(\xi) + n_2 V_{n_2}(\eta)) \frac{n_1 + n_2}{n_1 n_2 (n_1 + n_2 - 2)}.$$

$P(-x < Z_{n_1, n_2} < x) = 2\Phi_{n-1}(x) - 1$, így $x_\alpha = \Phi_{n_1+n_2-2}^{-1}(1 - \alpha/2)$ választással kapjuk:

$$1 - \alpha = P\left(-x_\alpha < \frac{E_{n_1}(\xi) - E_{n_2}(\eta) - (\mu_1 - \mu_2)}{D_*} < x_\alpha\right)$$

$$= P(E_{n_1}(\xi) - E_{n_2}(\eta) - x_\alpha D_* < \mu_1 - \mu_2 < E_{n_1}(\xi) - E_{n_2}(\eta) + x_\alpha D_*).$$

Konf. intervallum normális eloszlások szóráshányadosára

Legyenek $\xi_1, \dots, \xi_{n_1} \sim N(\mu_1, \sigma_1^2)$ és $\eta_1, \dots, \eta_{n_2} \sim N(\mu_2, \sigma_2^2)$ fgn. minták. σ_1/σ_2 -re keresünk konfidencia intervallumot. legyen

$$Z_{n_1, n_2} = \frac{V_{n_1}^*(\xi) \cdot \sigma_2^2}{V_{n_2}^*(\eta) \cdot \sigma_1^2} \sim F(n_1 - 1, n_2 - 1).$$

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

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

Legyen θ a háttéreloszlás egy ismeretlen paramétere a C **paramétertér** egy eleme, továbbá $C_0 \cup C_1 = C$, $C_0 \cap C_1 = \emptyset$, $C_0, C_1 \neq \emptyset$. A

$$H_0 : \theta \in C_0, \quad H_1 : \theta \in C_1$$

nullhipotézist és **alternatív hipotézist** vizsgáljuk.

A hipotézisek közötti döntés egy S_n **tesztstatisztika** (**próbastatisztika**) és egy $K(\alpha)$ **kritikus tartomány** segítségével történik: pontosan akkor vetjük el H_0 -t, ha $S_n \in K(\alpha)$.

Ha $P_{H_0}(S_n \in K(\alpha)) \leq \alpha$, akkor α **terjedelmű** próbáról beszélünk. α -t nevezik **szignifikancia-szint**nek is.

- ▶ H_0 általában $\theta = \theta_0$ alakú, H_1 pedig $\theta \neq \theta_0$ (**kétoldali próba**), $\theta < \theta_0$ vagy $\theta > \theta_0$ (**egyoldali próba**).
- ▶ α -t általában 0,05-nak vagy 0,01-nak, néha 0,1-nek választják.

Tévedési lehetőségek:

- ▶ **Elsőfajú hiba:** H_0 teljesül, de elvetjük. Ennek valószínűsége $p_1 = P_{H_0}(S_n \in K(\alpha)) \leq \alpha$.
- ▶ **Másodfajú hiba:** H_0 nem teljesül, de elfogadjuk. Ennek valószínűsége $p_2 = P_{H_1}(S_n \notin K(\alpha))$.

Definíció

$e_n(\alpha, \theta) = 1 - p_2 = P_{H_1}(S_n \in K(\alpha))$ a próba **ereje**, az e_n függvény az **erőfüggvény**.

Definíció

Egy próba **konzisztens**, ha $\lim_{n \rightarrow \infty} e_n(\alpha, \theta) = 1$.

Megjegyzések

- ▶ A kritikus tartomány csökkentésével az elsőfajú hiba csökken, a másodfajú hiba nő.
- ▶ A konzisztencia azt jelenti, hogy a mintaelemszám növelésével a másodfajú hiba teszőlegesen kicsivé tehető.

Általános módszer próbák szerkesztésére

- ▶ Keresünk egy S_n statisztikát, aminek ismert az eloszlása, ha H_0 teljesül. Jelölje az eloszlásfüggvényét G .
- ▶ Egyoldali próba esetén meghatározunk egy s_α **kritikus értéket**, amelyre $P_{H_0}(S_n < s_\alpha) = G(s_\alpha) = 1 - \alpha$: $s_\alpha = G^{-1}(1 - \alpha)$.
- ▶ Kétoldali próba esetén $s_\alpha^{(1)}$ és $s_\alpha^{(2)}$ kritikus értékeket keresünk, amelyekre $P_{H_0}(s_\alpha^{(1)} \leq S_n < s_\alpha^{(2)}) = G(s_\alpha^{(2)}) - G(s_\alpha^{(1)}) = 1 - \alpha$: $s_\alpha^{(1)} = G^{-1}(\alpha/2)$ és $s_\alpha^{(2)} = G^{-1}(1 - \alpha/2)$.
- ▶ Döntés: H_0 elfogadjuk, ha egyoldali esetben $S_n \leq s_\alpha$, kétoldali esetben $s_\alpha^{(1)} \leq S_n \leq s_\alpha^{(2)}$.

Megjegyzés

Ha több módon választhatjuk meg S_n statisztikát (illetve a kritikus értékeket), akkor válasszuk azt, ahol az erőfüggvény nagyobb.

u-próba (μ -próba)

Tekintsünk egy $N(\mu, \sigma^2)$ eloszlásból vett mintát. Tegyük fel, hogy σ ismert. Rögzített μ_0 esetén vizsgáljuk a következő hipotéziseket:

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

- ▶ Tekintsük a következő próbastatisztikát:

$$u = \frac{E_n(\xi) - \mu_0}{\sigma/\sqrt{n}},$$

ami H_0 teljesülése esetén standard normális eloszlású.

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

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

legyen amiből $u_\alpha = \Phi^{-1}(1 - \alpha/2)$ adódik.

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

A hibák valószínűségei:

- ▶ Elsőfajú hiba valószínűsége:

$$p_1 = P_{H_0}(|u| > u_\alpha) = 1 - P_{H_0}(-u_\alpha \leq u \leq u_\alpha) = \alpha.$$

- ▶ Felhasználva, hogy $\frac{E_n(\xi) - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, a másodfajú hiba valószínűsége:

$$\begin{aligned} p_2 &= P_{H_1}(|u| \leq u_\alpha) \\ &= P_{H_1}\left(-u_\alpha \leq \frac{E_n(\xi) - \mu_0}{\sigma/\sqrt{n}} \leq u_\alpha\right) \\ &= P_{H_1}\left(-u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \leq \frac{E_n(\xi) - \mu}{\sigma/\sqrt{n}} \leq u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) - \Phi\left(-u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right). \end{aligned}$$

Állítás

Az u -próba konzisztens.

Bizonyítás

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

- ▶ Ha $\mu > \mu_0$, akkor $\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \rightarrow -\infty$, ha $n \rightarrow \infty$. Tehát
 $\lim_{n \rightarrow \infty} e_n(\alpha, \mu) = 1 - \lim_{x \rightarrow -\infty} \Phi(x) + \lim_{x \rightarrow -\infty} \Phi(x) = 1.$
- ▶ Ha $\mu_0 < \mu$, akkor $\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \rightarrow \infty$, ha $n \rightarrow \infty$. Tehát
 $\lim_{n \rightarrow \infty} e_n(\alpha, \mu) = 1 - \lim_{x \rightarrow \infty} \Phi(x) + \lim_{x \rightarrow \infty} \Phi(x) = 1.$

Megjegyzés

Könnyen látható, hogy az α terjedelmű u -próba pontosan akkor fogadja el H_0 -t, ha μ_0 benne van az ismert σ esetén μ -re konstruált $1 - \alpha$ megbízhatósági szintű konfidencia intervallumban.

Most a következő hipotéziseket vizsgáljuk:

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

- ▶ A próbastatisztika most is

$$u = \frac{E_n(\xi) - \mu_0}{\sigma / \sqrt{n}},$$

ami H_0 teljesülése esetén standard normális eloszlású.

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

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

legyen amiből $u_\alpha = \Phi^{-1}(1 - \alpha)$ adódik.

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

Állítás

Az egyoldali u -próba konzisztens.

Bizonyítás

Felhasználva, hogy $\frac{E_n(\xi) - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, a másodfajú hiba valószínűsége:

$$\begin{aligned} p_2 &= P_{H_1} \left(\frac{E_n(\xi) - \mu_0}{\sigma/\sqrt{n}} \leq u_\alpha \right) \\ &= P_{H_1} \left(\frac{E_n(\xi) - \mu}{\sigma/\sqrt{n}} \leq u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right) \\ &= \Phi \left(u_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right). \end{aligned}$$

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

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

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

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

- Tekintsük a következő próbastatisztikát:

$$t = \frac{E_n(\xi) - \mu_0}{\sqrt{V_n^*/n}},$$

ami H_0 teljesülése esetén $n-1$ szabadságfokú Student eloszlású.

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

$$P_{H_0}(|t| < t_\alpha) = \Phi_{n-1}(t_\alpha) - \Phi_{n-1}(-t_\alpha) = 2\Phi_{n-1}(t_\alpha) - 1 = 1 - \alpha$$

legyen amiből $t_\alpha = \Phi_{n-1}^{-1}(1 - \alpha/2)$ adódik.

- Döntés: ha $|t| \leq t_\alpha$, akkor H_0 -t elfogadjuk, különben elvetjük.

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

- ▶ A próbastatisztika most is

$$t = \frac{E_n(\xi) - \mu_0}{\sqrt{V_n^*/n}},$$

ami H_0 teljesülése esetén $n-1$ szabadságfokú Student eloszlású.

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

$$P_{H_0}(t < t_\alpha) = \Phi_{n-1}(t_\alpha) = 1 - \alpha$$

legyen amiből $t_\alpha = \Phi_{n-1}^{-1}(1 - \alpha)$ adódik.

- ▶ Döntés: ha $t \leq t_\alpha$, akkor H_0 -t **elfogadjuk**, különben elvetjük.

Megjegyzések

- ▶ Belátható, hogy a két- és egyoldali t-próba is konzisztens.
- ▶ Könnyen látható, hogy az α terjedelmű kétoldali t-próba pontosan akkor fogadja el H_0 -t, ha μ_0 benne van az ismeretlen σ esetén μ -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.

Tekintsük a $\xi_1, \dots, \xi_{n_1} \sim N(\mu_1, \sigma^2)$ és $\eta_1, \dots, \eta_{n_2} \sim N(\mu_2, \sigma^2)$ független mintákat.

$$H_0 : \mu_1 - \mu_2 = \Delta_0, \quad H_1 : \mu_1 - \mu_2 \neq \Delta_0.$$

- ▶ A próbastatisztika:

$$t = \frac{E_{n_1}(\xi) - E_{n_2}(\eta) - \Delta_0}{\sqrt{n_1 V_{n_1}(\xi) + n_2 V_{n_2}(\eta)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}},$$

amiről tudjuk, hogy $n_1 + n_2 - 2$ szabadságfokú Student eloszlású H_0 teljesülése esetén.

- ▶ A kritikus érték α szinthez $t_\alpha = \Phi_{n_1+n_2-2}^{-1}(1 - \alpha/2)$.
- ▶ Döntés: ha $|t| \leq t_\alpha$, akkor H_0 -t elfogadjuk, különben elvetjük.

Megjegyzések

- ▶ Belátható, hogy a kétmintás t-próba konzisztens.
- ▶ Könnyen látható, hogy az α terjedelmű kétmintás t-próba pontosan akkor fogadja el a nullhipotézist, ha μ_0 benne van az ismeretlen szórás esetén $\mu_1 - \mu_2$ -re konstruált $1 - \alpha$ megbízhatósági szintű konfidencia intervallumban.
- ▶ A kétmintás t-próbának van egyoldali változata is, ahol $H_0 : \mu_1 - \mu_2 = \Delta_0$, $H_1 : \mu_1 - \mu_2 > \Delta_0$, a t próbastatisztika ugyanaz, a kritikus érték $t_\alpha = \Phi_{n_1+n_2-2}^{-1}(1 - \alpha)$, a döntés pedig ha $t \leq t_\alpha$, akkor H_0 -t elfogadjuk, különben elvetjük.
- ▶ A kétmintás t-próba feltétele a szórások egyezése, amit az alábbi F-próbával tesztelünk. Ha a szórások nem egyenlőek, akkor $\mu_1 - \mu_2$ -re vonatkozó hipotéziseket az ún. **Welch**-próbával vizsgáljuk.

Tekintsük a $\xi_1, \dots, \xi_{n_1} \sim N(\mu_1, \sigma_1^2)$ és $\eta_1, \dots, \eta_{n_2} \sim N(\mu_2, \sigma_2^2)$ független mintákat.

$$H_0 : \frac{\sigma_1}{\sigma_2} = \tau_0, \quad H_1 : \frac{\sigma_1}{\sigma_2} \neq \tau_0.$$

- ▶ A próbastatisztika:

$$F = \frac{V_{n_1}^*}{V_{n_2}^* \cdot \tau_0},$$

ami $n_1 - 1, n_2 - 1$ szabadságfokú F eloszlású H_0 esetén.

- ▶ A kritikus értékek α szinthez:

$$f_{1,\alpha} = F_{n_1-1, n_2-1}^{-1}(\alpha/2) \text{ és } f_{2,\alpha} = F_{n_1-1, n_2-1}^{-1}(1 - \alpha/2).$$

- ▶ Döntés: ha $f_{1,\alpha} \leq F \leq f_{2,\alpha}$, akkor H_0 -t elfogadjuk, különben elvetjük.

A gyakorlatban az F-próba módosított változatát alkalmazzák, ami azon alapszik, hogy H_0 teljesülése esetén $1/F \sim F(n_2 - 1, n_1 - 1)$, és ennek megfelelően $f_{1,\alpha} < 1$, valamint $f_{2,\alpha} > 1$.

- Tehát a módosított próbastatisztika

$$F = \max \left\{ \frac{V_{n_1}^*}{V_{n_2}^* \cdot \tau_0}, \frac{V_{n_2}^* \cdot \tau_0}{V_{n_1}^*} \right\} \geq 1.$$

- A kritikus érték α szinthez

$$F_\alpha = \begin{cases} F_{n_1-1, n_2-1}^{-1}(1 - \alpha/2), & \text{ha } V_{n_1}^*/V_{n_2}^* \geq \tau_0, \\ F_{n_2-1, n_1-1}^{-1}(1 - \alpha/2), & \text{ha } V_{n_1}^*/V_{n_2}^* < \tau_0. \end{cases}$$

- Akkor fogadjuk el H_0 -t, ha $F < F_\alpha$.

Szemléletesen: a mintát úgy választjuk, hogy az eredeti változatban $F \geq 1$ teljesül, és csak a felső kritikus értékkel hasonlítjuk össze. Ez magyarázza, hogy az F táblázatokban csak $F_{m,n}^{-1}(p)$ értékek 1-hez közeli p -kre vannak feltüntetve.

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

Legyen (ξ, η) olyan vektorváltozó, amely komponenseinek létezik a várható értéke: $E(\xi) = \mu_1$ és $E(\eta) = \mu_2$, továbbá $\nu = \xi - \eta$ normális eloszlású. Tekintsük a $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ mintát. A

$$H_0 : \mu_1 = \mu_2$$

hipotézist vizsgáljuk. Az eljárás a következő:

- ▶ Képezzük a $\nu_i = \xi_i - \eta_i$ változókat, amik függetlenek és normális eloszlásúak $\mu = \mu_1 - \mu_2$ várható értékkel.
- ▶ A $\nu_1, \dots, \nu_n$ minta segítségével **egymintás t-próbával** teszteljük a $H_0 : \mu = 0$ nullhipotézist. Ha $H_1 : \mu_1 \neq \mu_2$, akkor a kétoldali, ha $H_1 : \mu_1 > \mu_2$, akkor az egyoldali változatot alkalmazzuk.

Nemparaméteres próbák esetén nem feltételezzük, hogy olyan eloszláscsaládból vesszük a háttérváltozó eloszlását, amely véges sok paraméterrel leírható.

Leggyakoribb nemparaméteres tesztek

- ▶ események valószínűségére vonatkozó próbák,
- ▶ eloszlásfüggvények tesztelése,
- ▶ kvantilisekre (általában mediánra) vonatkozó próbák,
- ▶ homogenitásvizsgálatok,
- ▶ függetlenségvizsgálatok,
- ▶ véletlenszerűség tesztelése.

Itt csak a legáltalánosabban használt nemparaméteres próbákat (kétoldali változataikat) tekintjük át.

Valószínűségek tesztelése

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

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

- ▶ Végezzünk n független megfigyelést, és jelölje φ_i az A_i bekövetkezéseinek számát.

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

H_0 esetén eloszlásban $\chi^2(r-1)$ -hez tart ($n \rightarrow \infty$), így

- ▶ a kritikus érték $\chi_\alpha^2 = F_{\chi^2, r-1}^{-1}(1 - \alpha)$.
- ▶ Döntés: (nagy mintaelemszám esetén) ha $\chi^2 < \chi_\alpha^2$, akkor elfogadjuk H_0 -t.

Tiszta illeszkedésvizsgálat χ^2 próbával

Legyen $\xi_1, \dots, \xi_n$ minta ismereten F eloszlásfüggvénnyel.

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

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

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

- ▶ Legyen $p_i = F_0(x_i) - F_0(x_{i-1})$.
- ▶ 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 χ^2 próbával.

Becsléses illeszkedésvizsgálat χ^2 próbával

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

ahol $\theta_1, \dots, \theta_k$ ismeretlen paraméterek.

Vegyük a paraméterek $\hat{\theta}_1, \dots, \hat{\theta}_k$ becsléseit és legyen $F_0 = F_{\hat{\theta}_1, \dots, \hat{\theta}_k}$. Eljárásunk ugyanaz lesz, mint a tiszta illeszkedésvizsgálatnál, csak most a próbastatisztika közelítőleg $\chi^2(r - k - 1)$ eloszlású H_0 mellett.

Megjegyzés

Ha ξ értékkészlete véges: $x_1, \dots, x_r$, akkor nincs szükség osztópontokra, hanem az $A_i = \{\xi = x_i\}$ választással a valószínűségek tesztelését közvetlenül alkalmazhatjuk.

Természetesen becsléses illeszkedésvizsgálat esetén a szabadságfok ebben az esetben is csökken a becsült paraméterek számával.

Homogenitásvizsgálat χ^2 próbával

Tekintsünk két mintát: $\xi_1, \dots, \xi_{n_1}$ és $\eta_1, \dots, \eta_{n_2}$.

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

- Legyenek most is $-\infty = x_0 < x_1 < \dots < x_{r-1} < x_r = \infty$ osztópontok és μ_i az $[x_{i-1}, x_i)$ intervallumba eső első, ν_i a második mintabeli elemek száma.

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

ami közelítőleg $\chi^2(r-1)$ eloszlású H_0 esetén.

- Tehát most is $\chi_\alpha^2 = F_{\chi^2, r-1}^{-1}(1 - \alpha)$.
- Döntés: ha $\chi^2 < \chi_\alpha^2$, akkor elfogadjuk H_0 -t.

Megjegyzés

Diszkrét véges értékű változók esetén itt nincs szükség az osztópontokra: intervallumokba esés helyett x_i értéket keresünk.

Függetlenségvizsgálat χ^2 próbával

Legyen ξ értékkészlete $x_1, \dots, x_r$ és η -é $y_1, \dots, y_s$.

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

azaz ξ és η függetlenek. Tekintsük a $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ mintát.

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

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

ami H_0 mellett közelítőleg $\chi^2((r-1)(s-1))$ eloszlást követ.

- ▶ Tehát $\chi_\alpha^2 = F_{\chi^2, (r-1)(s-1)}^{-1}(1 - \alpha)$.
- ▶ Döntés: ha $\chi^2 < \chi_\alpha^2$, akkor **elfogadjuk H_0 -t**.

Tiszta illeszkedésvizsgálat

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

- ▶ A próbat statisztika

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

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

- ▶ A kritikus érték $k_\alpha = K^{-1}(1 - \alpha)/\sqrt{n}$.
- ▶ Döntés: elfogadjuk H_0 -t, ha $\Delta_n < k_\alpha$.

Megjegyzések

- ▶ Δ_n meghatározásához elég $2n$ különbséget meghatározni, mivel F_0 monoton növény, F_n pedig lépcsős függvény.
- ▶ Egyoldali változatban a próbat statisztika abszolútérték nélküli és közelítőleg Szmirnov-eloszlású.

Homogenitásvizsgálat Kolmogorov-Szmirnov próbával

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

Tekintsük a $\xi_1, \dots, \xi_{n_1}$ és $\eta_1, \dots, \eta_{n_2}$ mintákat és a hozzájuk tartozó F_{n_1} és G_{n_2} empirikus eloszlásfüggvényeket.

- ▶ A próbastatisztika

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

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

- ▶ Tehát $k_\alpha = \sqrt{\frac{n_1 + n_2}{n_1 n_2}} K^{-1}(1 - \alpha)$.
- ▶ Döntés: elfogadjuk H_0 -t, ha $\Delta_n < k_\alpha$.

Az illeszkedésvizsgálat megjegyzései itt is érvényesek.

Egymintás előjelpróba (binomiális teszt)

$$H_0 : P(\xi < x_0) = p_0,$$

azaz a p_0 -kvantilis éppen x_0 .

- ▶ $S_n = nF_n(x_0)$ a H_0 mellett $\text{binom}(n, p_0)$ eloszlású.
- ▶ A kritikus értékek c_1 és c_2 , amelyekre éppen

$$P_{H_0}(S_n \leq c_1) = \sum_{k=0}^{c_1} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \frac{\alpha}{2},$$

$$P_{H_0}(S_n \geq c_2) = \sum_{k=c_2}^n \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \frac{\alpha}{2}.$$

- ▶ Ha n nagy, CHT miatt $\frac{S_n - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0, 1)$ közelítőleg,

$$c_{1,2} = np_0 \mp \sqrt{np_0(1-p_0)} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

- ▶ Döntés: ha $c_1 < S_n < c_2$, akkor H_0 -t elfogadjuk.

Az előjelpróba kétmintás változata. $\xi_1, \dots, \xi_{n_1}$ és $\eta_1, \dots, \eta_{n_2}$ független, folytonos eloszlású minták ($n_1 + n_2$ páros).

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

- ▶ Vegyük az egyesített minta mediánját és jelölje M_{n_1, n_2} az első mintabeli mediánnál kisebb elemek számát. H_0 mellett

$$P(M_{n_1, n_2} = k) = \frac{\binom{n_1}{k} \binom{\frac{n_1+n_2}{2} - k}{\frac{n_1+n_2}{2}}}{\binom{\frac{n_1+n_2}{2}}{\frac{n_1+n_2}{2}}}, \quad k = 0, \dots, n_1.$$

- ▶ $E(M_{n_1, n_2}) = \frac{n_1}{2}$ és $D^2(M_{n_1, n_2}) = \frac{n_1 n_2}{4(n_1 + n_2 + 1)}$.
- ▶ Ha n_1 és n_2 nagy, akkor $\frac{M_{n_1, n_2} - E(M_{n_1, n_2})}{D(M_{n_1, n_2})} \sim N(0, 1)$ közelítőleg.
- ▶ Döntés: H_0 -t elfogadjuk, ha $\left| \frac{M_{n_1, n_2} - E(M_{n_1, n_2})}{D(M_{n_1, n_2})} \right| < \Phi^{-1} \left(1 - \frac{\alpha}{2} \right)$.

Legyen ξ folytonos, a mediánra szimmetrikus eloszlású vv.

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

- ▶ Rendezzük sorba $|\xi_1 - m_0|, \dots, |\xi_n - m_0|$ mintát és legyen R_i **rangsám** $|\xi_i - m_0|$ sorszáma ebben a rendezett mintában.
- ▶ Legyen $T^+ = \sum_{i: \xi_i > m_0} R_i$. H_0 mellett
 - ▶ kis n esetén T^+ eloszlásának $t(\alpha, n)$ kvantiliseit táblázat tartalmazza.
 - ▶ Nagy n -re $\frac{T^+ - E(T^+)}{D(T^+)} \sim N(0, 1)$ közelítőleg,
 $E(T^+) = \frac{1}{2} \sum_{i=1}^n i = \frac{n(n+1)}{4}$ és $D^2(T^+) = \frac{n(n+1)(2n+1)}{24}$.
- ▶ Döntés: **H_0 -t elfogadjuk**, ha
 - ▶ kis n esetén $t(\alpha/2, n) < T^+ < \frac{n(n+1)}{2} - t(\alpha/2, n)$,
 - ▶ nagy n -re $\left| \frac{T^+ - E(T^+)}{D(T^+)} \right| < \Phi^{-1}(1 - \alpha/2)$.

Páros Wilcoxon-próba

ξ, η folytonos eloszlásúak.

$$H_0 : P(\xi < \eta) = \frac{1}{2}$$

Tekintsük a $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ mintát. Legyen $\nu_i = \xi_i - \eta_i$, így

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

amit Wilcoxon-próbával tesztelhetünk, ha ν eloszlása folytonos és a mediánjára szimmetrikus.

Megjegyzések

- ▶ A páros t-próba nemparaméteres alternatívájaként alkalmazzák.
- ▶ Az előjelpróbának is hasonlóan konstruálhatjuk meg a páros változatát, de ott nem kell feltenni ν szimmetriáját.

Kétmintás Wilcoxon (Mann-Whitney) próba

$\xi_1, \dots, \xi_{n_1}$ és $\eta_1, \dots, \eta_{n_2}$ független, folytonos eloszlású minták.

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

- ▶ Rendezzük sorba az egyesített mintát és legyen R_i a ξ_i rangja.
- ▶ $R = \sum_{i=1}^{n_1} R_i$. H_0 mellett
 - ▶ kis elemszám esetén R eloszlásának $r(\alpha, n, m)$ kvantiliseit táblázat tartalmazza.
 - ▶ Nagy elemszámnál $\frac{R - E(R)}{D(R)} \sim N(0, 1)$ közelítőleg, ahol $E(R) = \frac{n_1(n_1 + n_2 + 1)}{2}$ és $D^2(R) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$.
- ▶ Döntés: H_0 -t elfogadjuk, ha
 - ▶ kis elemszám esetén $r(\alpha/2, n_1, n_2) < R < n_1(n_1 + n_2 + 1) - r(\alpha/2, n_1, n_2)$,
 - ▶ nagy elemszámmra $\left| \frac{R - E(R)}{D(R)} \right| < \Phi^{-1}(1 - \alpha/2)$.

Spearman-féle rangkorreláció

Definíció

Tekintsük a $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ mintát. Legyen R_i a ξ_i és S_i az η_i rangszáma. A **Spearman-féle rangkorrelációs együttható**

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

Tulajdonságok

- ▶ $\rho_n = \frac{\sum_{i=1}^n (R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})}{n(n^2-1)/12} = 1 - \frac{6 \sum_{i=1}^n (R_i - S_i)^2}{n(n^2-1)}.$
- ▶ $|\rho_n| \leq 1$. Ha $\rho_i = 1$, akkor $\xi_i \leq \xi_j$ implikálja $\eta_i \leq \eta_j$ -t, ha $\rho_i = -1$, akkor $\xi_i \leq \xi_j$ implikálja $\eta_i \geq \eta_j$ -t minden $i \neq j$ -re.

Függetlenségteszt

$H_0 : \xi$ és η függetlenek.

- ▶ H_0 mellett $\sqrt{n-1}\rho_n \sim N(0, 1)$ aszimptotikusan.
- ▶ Döntés: ha $|\sqrt{n-1}\rho_n| < \Phi^{-1}(1 - \alpha/2)$, elfogadjuk H_0 -t.

Véletlenszerűség tesztelése

I. Adott egy két jelből álló sorozat. H_0 : sorrendjük véletlenszerű.

xxxyyxyyyy

Az azonos elemekből álló maximális sorozatok a **futamok**. Az x -ek száma n , az y -oké m , a futamok számát jelölje W . Legyen

$$I_k = \begin{cases} 1, & \text{ha a } k\text{-adik elem futam eleje,} \\ 0 & \text{különben.} \end{cases}$$

- ▶ H_0 mellett $I_1 = 1$ és $k > 1$ -re I_k Bernoulli-eloszlású $\frac{nm}{\binom{n+m}{2}}$ paraméterrel, valamint $W = \sum_{i=1}^{n+m} I_i$.
- ▶ H_0 mellett $E(W) = 1 + \frac{2nm}{n+m}$ és $D^2(W) = \frac{2nm(2nm-n-m)}{(n+m)^2(n+m-1)}$.
- ▶ Ha n és m nagy, akkor $\frac{W-E(W)}{D(W)}$ közelítőleg $N(0, 1)$ eloszlású.

II. Vegyünk egy n hosszú 2 jelből állt sorozatot. H_0 : véletlen forrásból származik (fae), x valószínűsége $1/2$. Ebben az esetben

- ▶ $I_2, \dots, I_n \sim B(1/2)$ fae., amiből $W - 1 \sim \text{binom}(n - 1, 1/2)$,
- ▶ így $E(W) = \frac{n+1}{2}$ és $D^2(W) = \frac{n-1}{4}$.
- ▶ CHT alapján nagy n -re $\frac{W-E(W)}{D(W)}$ közelítőleg $N(0, 1)$ eloszlású.

Döntés (I-II): H_0 -t elfogadjuk, ha $\left| \frac{W-E(W)}{D(W)} \right| < \Phi^{-1} \left(1 - \frac{\alpha}{2} \right)$.

III. H_0 : $x_1, \dots, x_n$ véletlen minta realizációja. Legyen m a medián és alkalmazzuk az II-t a $\text{sgn}(x_1 - m), \dots, \text{sgn}(x_n - m)$ előjelsorozatra.

Futamteszt homogenitásvizsgálatra (Wald-Wolfowitz)

$\xi_1, \dots, \xi_{n_1}$ és $\eta_1, \dots, \eta_{n_2}$ független, folytonos eloszlású minták.

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

Rendezzük az egyesített mintát és ebben a ξ - η jelsorozatra alkalmazzuk I-t.

Ezek a módszerek több változó együttes vizsgálatára vonatkoznak.

Alapvető típusaik:

- ▶ többdimenziós eloszlásokra vonatkozó hipotézisvizsgálatok (általában többdimenziós normális eloszlás paramétereire vonatkoznak, itt nem térünk ki rájuk),
- ▶ varianciaanalízisek,
- ▶ klasszifikációs módszerek (diszkrimináló illetve klaszterező eljárások),
- ▶ dimenziócsökkentés,
- ▶ ...

Egyszempontos varianciaanalízis (ANOVA)

Tekintsük a $\xi_{i,j} \sim N(\mu_i, \sigma^2)$, $i = 1, \dots, r$, $j = 1, \dots, n_i$ független mintákat.

A várható értékek felírhatók a következő formában: $\mu_i = \mu + a_i$, ahol μ a várható értékek n_i értékekkel súlyozott átlaga, a_i pedig az i -edik csoportthatás.

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

azaz minden csoportthatás 0.

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

$$\bar{\xi}_i = \frac{\sum_{j=1}^{n_i} \xi_{i,j}}{n_i}, \quad \bar{\xi} = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} \xi_{i,j}}{n}.$$

$$SST = \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{i,j} - \bar{\xi})^2,$$

$$SSB = \sum_{i=1}^r n_i (\bar{\xi}_i - \bar{\xi})^2, \quad SSW = \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{i,j} - \bar{\xi}_i)^2.$$

Állítás

$$SST = SSB + SSW.$$

Bizonyítás

$$\begin{aligned} SST &= \sum_{i=1}^r \sum_{j=1}^{n_i} ((\xi_{i,j} - \bar{\xi}_i) + (\bar{\xi}_i - \bar{\xi}))^2 \\ &= SSW + SSB + 2 \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{i,j} - \bar{\xi}_i)(\bar{\xi}_i - \bar{\xi}) \\ &= SSW + SSB + 2 \sum_{i=1}^r (\bar{\xi}_i - \bar{\xi}) \underbrace{\sum_{j=1}^{n_i} (\xi_{i,j} - \bar{\xi}_i)}_0. \end{aligned}$$

Visszatérve H_0 vizsgálatára:

- ▶ a próbastatisztika

$$F = \frac{SSB}{SSW} \cdot \frac{n-r}{r-1},$$

ami H_0 mellett $F(r-1, n-r)$ eloszlású.

- ▶ Tehát a kritikus érték α -hoz $F_\alpha = F_{r-1, n-r}^{-1}(1-\alpha)$.
- ▶ Döntés: ha $F < F_\alpha$, akkor H_0 -t elfogadjuk, különben elvetjük.

Megjegyzések

- ▶ ANOVA a kétmintás t-próba általánosítása: könnyen látható, hogy ha $r = 2$ és $\Delta_0 = 0$, akkor $F = t^2$, és $F_\alpha = t_\alpha^2$.
- ▶ Itt is feltétel a szórások egyezése, ami az általánosított F -próbával tesztelhető. Ha a szórások nem egyenlőek, akkor ANOVA helyett a többmintás Welch-próbát alkalmazzák.
- ▶ Az ANOVA-t diszkrét és normális eloszlású változó függetlenségének tesztelésére is használják.
- ▶ Van többszemponyos varianciaanalízis is.

Lineáris regresszió

Tekintsük a (ξ, η) véletlen vektort. Keressük η -t legjobban közelítő lineáris függvényt, azaz a -t és b -t, amelyre $E((\eta - (a\xi + b))^2)$ négyzetes eltérés minimális.

- ▶ $E((\eta - a\xi - b)^2) = E((\eta - a\xi)^2) - 2b \cdot E(\eta - a\xi) + b^2$
minimális, ha $b = E(\eta - a\xi) = E(\eta) - aE(\xi)$.
- ▶ $E((\eta - a\xi - E(\eta) + aE(\xi))^2) = D^2(\eta) + a^2 D^2(\xi) - 2a \cdot \text{Cov}(\eta, \xi)$,
aminek maximumhelye $a = \frac{\text{Cov}(\xi, \eta)}{D^2(\xi)} = r(\xi, \eta) \frac{D(\eta)}{D(\xi)}$,
- ▶ amiből $b = E(\eta) - \frac{\text{Cov}(\xi, \eta)}{D^2(\xi)} E(\xi)$.

Definíció

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

Megjegyzés

Ha ξ, η függetlenek, akkor $a = r(\xi, \eta) \frac{D(\eta)}{D(\xi)} = 0$.

Regressziós egyenes paramétereinek ML becslése

Tekintsük a $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ mintát és az $(x_1, y_1), \dots, (x_n, y_n)$ realizációt. Tegyük fel, hogy $\eta = a\xi + b + \varepsilon$, ahol $\varepsilon \sim N(0, \sigma^2)$.

Célunk a és b paraméterek becslése.

Az egyszerűség kedvéért tegyük fel, hogy x_i értékek nem véletlen minta realizációi, hanem rögzített értékek, így az $\eta_i = ax_i + b + \varepsilon_i$ modellt kapjuk, ahol $\varepsilon_1, \dots, \varepsilon_n \sim N(0, \sigma^2)$ függetlenek. Tehát $\eta_1, \dots, \eta_n$ függetlenek, $\eta_i \sim N(ax_i + b, \sigma^2)$. Jelölje η_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 négyzetek módszere).

Megjegyzés

$Q(a, b)$ az $(x_1, y_1), \dots, (x_n, y_n)$ pontok együttes négyzetes távolsága az $y = ax + b$ egyenestől.

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

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

A lehetséges szélsőérték helyek 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

$$\hat{a} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$
$$= \frac{n^2 C_n(\xi, \eta)}{n^2 V_n(\xi)} = r_n(\xi, \eta) \frac{D_n(\eta)}{D_n(\xi)}.$$

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

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

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

Definíció

Az $\hat{a}, \hat{b}$ értékeket *becsült lineáris regressziós együtthatóknak*, az $y = \hat{a}x + \hat{b}$ egyenest *becsült lineáris regressziós egyenesnek* nevezzük.

Megjegyzés

A becsült lineáris regressziós egyenest η_{n+1} érték becslésére használjuk, ha x_{n+1} adott csak.

Varianciaanalízis lineáris modellben

Tekintsük a $\eta = ax + b + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ regressziós modellt és az $\eta_1, \dots, \eta_n$ mintát.

$$H_0 : a = 0.$$

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

$$SST = \sum_{i=1}^n (\eta_i - \bar{\eta})^2, \quad SSR = \sum_{i=1}^n (\hat{\eta}_i - \bar{\eta})^2, \quad SSE = \sum_{i=1}^n (\eta_i - \hat{\eta}_i)^2.$$

► a próbastatisztika $F = \frac{SSR}{SSE}(n-2),$

ami H_0 mellett $F(1, n-2)$ eloszlású.

► Tehát a kritikus érték α -hoz $F_\alpha = F_{1, n-2}^{-1}(1 - \alpha).$

► Döntés: ha $F < F_\alpha$, akkor H_0 -t elfogadjuk, különben elvetjük.

Állítás

$$SST = SSR + SSE.$$

Bizonyítás

$$SST = \sum_{i=1}^n ((\eta_i - \hat{\eta}_i) + (\hat{\eta}_i - \bar{\eta}))^2 = SSR + SSE + \underbrace{2 \sum_{i=1}^n (\eta_i - \hat{\eta}_i)(\hat{\eta}_i - \bar{\eta})}_{2 \sum_{i=1}^n (\eta_i - \hat{a}x_i - \hat{b})(\hat{a}x_i + \hat{b} - \bar{\eta})} = -\hat{a} \frac{\partial Q}{\partial a} - (\hat{b} - \bar{\eta}) \frac{\partial Q}{\partial b} = 0.$$

Állítás

$$SSR = SST \cdot r_n^2(x, \eta).$$

Bizonyítás

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

Következmény

$$SSE = SST(1 - r_n^2(x, \eta)) \quad \text{és} \quad \frac{SSR}{SSE} = \frac{r_n^2(x, \eta)}{1 - r_n^2(x, \eta)}.$$

Többsváltozós regresszió

A modell $\eta = a_1x_1 + \dots + a_rx_r + b + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$.

$$H_0 : a_1 = \dots = a_r = 0.$$

- ▶ Vesszük $a_1, \dots, a_r, b$ legkisebb négyzetes becsléseit, $\hat{\eta}_i = \hat{a}_1x_{i,1} + \dots + \hat{a}_rx_{i,r} + \hat{b}_i$ és képezzük SST , SSR és SSE négyzetösszegeket.
- ▶ A próbastatisztika most $F = \frac{SSR}{SSE} \cdot \frac{n - r - 1}{r}$, ami H_0 mellett $F(r, n - r - 1)$ eloszlású.

Megjegyzések

- ▶ Az x_i -k a teljes variancia $r_n^2(x, \eta)$ -részét magyarázzák.
- ▶ Az általános regressziós modell $\eta = f(\xi) + \varepsilon$, ahol f néhány paraméterrel jellemzett függvény. Általában a nemlineáris regressziót a lineárisra vezetik vissza.

Legyen $\mathbf{X}$ egy d -dimenziós véletlen vektor $\mathbf{m}$ várható érték vektorral és pozitív definit $\mathbf{C}$ kovarianciamátrixszal. Cél a változók számának csökkentése minél kevesebb információ vesztesével.

$\mathbf{C}$ szimmetrikus, spektrálfelbontása

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T,$$

- ▶ ahol $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_d)$ a $\mathbf{C}$ mátrix $\lambda_1 \geq \dots \geq \lambda_d > 0$ sajátértékeinek diagonális mátrixa,
- ▶ $\mathbf{V}$ oszlopvektorai pedig a hozzá tartozó ortonormált sajátvektorrendszer a megfelelő sorrendben.

Definíció

A $\mathbf{Y} = \mathbf{V}^{-1}(\mathbf{X} - \mathbf{m}) = \mathbf{V}^T(\mathbf{X} - \mathbf{m})$ az $\mathbf{X}$ főkomponensvektora, $\mathbf{Y}$ k -adik komponense (Y_k) a k -adik főkomponens.

Állítás

$\mathbf{Y}$ kovarianciamátrixa Λ .

Bizonyítás

$$\mathbf{E}(\mathbf{Y}) = \mathbf{E}(\mathbf{V}^T(\mathbf{X} - \mathbf{m})) = \mathbf{V}^T(\mathbf{E}(\mathbf{X}) - \mathbf{m}) = \mathbf{0},$$

$$\begin{aligned}\mathbf{E}(\mathbf{Y}\mathbf{Y}^T) &= \mathbf{E}(\mathbf{V}^T(\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T\mathbf{V}) \\ &= \mathbf{V}^T\mathbf{E}((\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T)\mathbf{V} = \mathbf{V}^T\mathbf{C}\mathbf{V} = \Lambda.\end{aligned}$$

Tétel

$\mathbf{x}_1, \dots, \mathbf{x}_d$ ortonormált vektorrendszer. Ekkor

$$\sum_{i=1}^k \mathbf{x}_i^T \mathbf{C} \mathbf{x}_i \leq \sum_{i=1}^k \lambda_i, \quad k = 1, \dots, d.$$

Ha $\mathbf{x}_i = \mathbf{v}_i, i = 1, \dots, d$, akkor egyenlőség teljesül.

Következmény

Y_k szórása *maximális* az $\mathbf{z}^T(\mathbf{X} - \mathbf{m})$ alakú vv-k között, amelyre $\|\mathbf{x}\| = 1$ és $\mathbf{z}^T(\mathbf{X} - \mathbf{m})$ *korrelálatlan* Y_i -vel minden $i < k$ -ra.

Tétel

A $k(< d)$ -dimenziós $\mathbf{P}$ projekciók közül $E(\|\mathbf{X} - \mathbf{P}\mathbf{X}\|)$ kifejezést a $\mathbf{v}_1, \dots, \mathbf{v}_k$ által feszített altérre vetítő projekció minimalizálja.

(Tehát $\mathbf{X}$ négyzetes középben vett legjobb k -dimenziós becslésének első k koordinátája a sajátvektorok bázisában az első k főkomponens, a többi pedig 0.)

k megválasztása

$$\frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_d}$$

hányados azt mutatja, hogy az első k főkomponens a teljes variancia hányadrészét magyarázza.

Megjegyzés

A gyakorlatban egy minta empirikus kovarianciamátrixából indulunk, ennek sajátértékei és sajátvektorai lesznek λ_i és $\mathbf{v}_i$ ($i = 1, \dots, d$) becslései.

Definíció

Legyen $\mathbf{X}$ egy d -dimenziós véletlen vektor $\mathbf{m}$ várható érték vektorral és $\mathbf{C}$ kovarianciamátrixszal. A *k -faktor modell* a következő:

$$\mathbf{X} = \mathbf{A}\mathbf{Y} + \mathbf{W} + \mathbf{m},$$

- ▶ ahol $\mathbf{A}$ *átviteli mátrix vagy faktorsúlymátrix* $d \times k$ -as,
- ▶ az $\mathbf{Y}$ *közös faktor* k -dimenziós véletlen vektor, amelyre $E(\mathbf{Y}) = \mathbf{0}$ és $E(\mathbf{Y}\mathbf{Y}^T) = \mathbf{I}_k$, valamint
- ▶ $\mathbf{W}$ *egyedi faktor* egy d -dimenziós véletlen vektor, amelyre $E(\mathbf{W}) = \mathbf{0}$, $E(\mathbf{W}\mathbf{W}^T) = \mathbf{D}$ diagonális mátrix, valamint $E(\mathbf{Y}\mathbf{W}^T) = \mathbf{0}$.

A modell paraméterei $\mathbf{A}$ és $\mathbf{D}$, feltevése, hogy a faktorkomponensek korrelálatlanok, továbbá a közös faktorok szórásnégyszete 1.

Állítás

$\mathbf{C} = \mathbf{A}\mathbf{A}^T + \mathbf{D}$. (Ezt nevezzük a modell mátrixalakjának.)

Bizonyítás

$X_i = \sum_{j=1}^k a_{i,j} Y_j + W_i + m_i \quad (1 \leq i \leq d)$ szórásnégyzete
 $c_{ii} = \sum_{j=1}^k a_{i,j}^2 + d_{i,i}$, mivel a faktorkomponensek korrelálatlanok.

($\sum_{j=1}^k a_{i,j}^2$ az X_i változó **kommunalitása**, $d_{i,i}$ az **egyedi variancia**.)

Tétel

Adott $k < d$ esetén a k -faktor modellnek pontosan akkor van megoldása, ha van olyan $d \times d$ -s $\mathbf{D}$ nemnegatív elemű diagonális mátrix, hogy $\mathbf{C} - \mathbf{D}$ egy legfeljebb k -adrangú, pozitív szemidefinit.

Bizonyítás

Ha van megoldás, akkor $\mathbf{C} - \mathbf{D} = \mathbf{A}\mathbf{A}^T$ rangja legfeljebb k , mivel $\mathbf{A}$ egy $d \times k$ -as mátrix. Fordítva, ha $\mathbf{C} - \mathbf{D}$ szimmetrikus, rangja $r(\leq k)$, akkor felírható $\mathbf{A}\mathbf{A}^T$ alakban, ahol $\mathbf{A}$ $d \times r$ -es.

Megjegyzések

- ▶ Ha $\mathbf{C} = \mathbf{D}$ -nek van $\mathbf{A}\mathbf{A}^T$ alakú felbontása, akkor mivel ez nem egyértelmű: ha $\mathbf{U}$ egy ortogonális mátrix, akkor $\mathbf{B} = \mathbf{A}\mathbf{U}^T$ -ra $\mathbf{A}\mathbf{A}^T = \mathbf{B}\mathbf{B}^T$, sokszor további feltételeket is tesznek $\mathbf{A}$ -ra.
- ▶ Közelítő megoldásokat szoktak adni (ML; legkisebb négyzetek módszere; főkomponensanalízis módosított változata;...).

A faktorsúlyok értelmezése és a faktorok rotációja

$\text{Cov}(X_i, Y_j) = E((X_i - m_i)Y_j) = E((\sum_{j=1}^k a_{i,j} Y_j + W_i)Y_j) = a_{i,j}$
a faktorsúlyok korrelálatlansága miatt.

- ▶ X_i az Y_j faktorról akkor áll szoros kapcsolatban, ha $|a_{i,j}|$ nagy.
- ▶ A faktorok értelmezése szempontjából az a jó, ha minden i -re kevés (esetleg egy) $|a_{i,j}|$ nagy, a többi kicsi.
- ▶ A **rotáció** célja olyan $\mathbf{U}$ $p \times p$ -s ortogonális mátrix keresése, amelyre az

$$\mathbf{X} = \mathbf{X} = \mathbf{A}^* \mathbf{Y}^* + \mathbf{W} + \mathbf{m}$$

modell, ahol $\mathbf{A}^* = \mathbf{A}\mathbf{U}$ és $\mathbf{Y}^* = \mathbf{U}^T \mathbf{Y}$, teljesíti a fentieket.

Legyen $\mathbf{X}$ változó $N(\mathbf{m}_1, \mathbf{C})$ vagy $N(\mathbf{m}_2, \mathbf{C})$ eloszlású. Ezt az **osztálybatarozást** szeretnénk eldönteni a tér egy $\mathcal{X}_1 \cup \mathcal{X}_2$ partícionálásával. Legyen p_i az i -edik osztályba esés valószínűsége, f_i az i -edik osztályhoz tartozó sűrűségfüggvény. (Általában $\mathbf{m}_i$ -t és $\mathbf{C}$ -t a mintából becsüljük.) Két megközelítést tekintünk.

I. A

$$\int_{\mathcal{X}_1} p_2 f_2(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{X}_2} p_1 f_1(\mathbf{x}) d\mathbf{x}$$

veszteségfüggvényt minimalizáló partícionálást keresünk.

Belátható, hogy a veszteségfüggvény akkor minimális, ha $\mathbf{m}_j^T \mathbf{C}^{-1} \mathbf{x}$ maximális ($j = 1, 2$). Ez alapján

$$L(\mathbf{x}) = (\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{C}^{-1} \mathbf{x} + c$$

alakú **Fisher-féle diszkriminancia-függvényt** használjuk: $\mathbf{x}$ pontosan akkor tartozik az első osztályba, ha $L(\mathbf{x}) > 0$.

II. Legyen $\mathbf{A} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$. Tekintsük a Rayleigh-hányadost:

$$R(\mathbf{v}) = \frac{\mathbf{v}^T \mathbf{A} \mathbf{v}}{\mathbf{v}^T \mathbf{C} \mathbf{v}}$$

Heurisztika: R maximumhelyének irányában jól tudunk szeparálni. Legyenek $\mathbf{C}^{-1}\mathbf{A}$ sajátértékei $\lambda_1 \geq \dots \geq \lambda_d$ és $\mathbf{v}_1, \dots, \mathbf{v}_d$ hozzájuk tartozó sajátvektorok a **kanonikus főirányok**.

► Belátható, hogy R maximumhelye $\mathbf{v}_1$.

► Egy $L(\mathbf{x}) = \mathbf{v}_1^T \mathbf{x} + b$

alakú **diszkriminanciafüggvény** segítségével döntünk: $\mathbf{x}$ akkor esik az első osztályba, ha $L(\mathbf{x}) > 0$.

Megjegyzések

- A két diszkriminanciafüggvény pozitív konstans faktorban tér el, ugyanazt az osztályozást eredményezik.
- A szeparálás jóságát a **Wilks-féle** $\Lambda = (1 + \hat{\lambda}_1)^{-1}$ statisztikával jellemezzük. Ha az osztályok jól szeparálhatók, akkor Λ kicsi.

Logisztikus regresszió

Legyen η ismeretlen p paraméterű Bernoulli-változó és $\mathbf{x}$ egy d -dimenziós vektor. Tegyük fel, hogy $p = p(\mathbf{x})$, ezt a függést a következő **logisztikus regressziós modell (logitmodell)** írja le:

$$\ln \frac{p}{1-p} = \mathbf{a}^T \mathbf{x} + b.$$

- ▶ Tekintsük egy n elemű mintát. Meghatározzuk $\hat{\mathbf{a}}, \hat{b}$ ML becsléseket (kihasználva, hogy az $\mathbf{x}_i$ vektorhoz tartozó n_i mintaelem összege binomiális eloszlású n_i és $p(\mathbf{x}_i)$ paraméterekkel). Ebből kapjuk p becslését:

$$\hat{p} = \frac{1}{1 + e^{-\hat{\mathbf{a}}^T \mathbf{x} - \hat{b}}}.$$

- ▶ η értékei szerinti osztálybatartozás előrejelzése ennek segítségével úgy történik, hogy rögzítünk egy $0 < p_0 < 1$ értéket, és $\hat{\eta} = 1$, ha $\hat{p} > p_0$.

Különböző eloszlásból vett minták esetén nem tudjuk, melyik mintaelem melyik **osztályba (klaszterbe)** tartozik, esetleg az osztályok száma is ismeretlen. Célunk az osztályok meghatározása.

Tekintsük az $\mathbf{x}_1, \dots, \mathbf{x}_n$ realizációt. Tegyük fel, hogy van k klaszter: $\mathcal{C}_1, \dots, \mathcal{C}_k$. Legyen $\bar{\mathbf{x}}_{\mathcal{C}} = \frac{1}{c} \sum_{\mathbf{x}_j \in \mathcal{C}} \mathbf{x}_j$ a $\mathcal{C}$ klaszterközéppontja.

Legyen $SST = \sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2$, $SSW = \sum_{i=1}^k \sum_{\mathbf{x}_j \in \mathcal{C}_i} \|\mathbf{x}_j - \bar{\mathbf{x}}_{\mathcal{C}_i}\|^2$, $SSB = \sum_{i=1}^k |\mathcal{C}_i| \cdot \|\bar{\mathbf{x}}_{\mathcal{C}_i} - \bar{\mathbf{x}}\|^2$. Most is $SST = SSW + SSB$.

A klasztereket úgy szeretnénk meghatározni, hogy a

$$\frac{SSW}{SSB} = \frac{SSW}{SST - SSW}$$

kifejezést **minimalizálják**, ami egyenértékű SSW minimalizálásával.

SSW minimumának meghatározása már két klaszter esetén is $2^{n-1} - 1$ eset vizsgálatát jelenti az n elemű mintán, ezért heurisztikákat alkalmaznak.

A klaszterező eljárásoknak alapvető fajtája van aszerint, hogy a klaszterek száma ismert-e.

k -közép módszer

Veszünk k kezdeti klaszterközéppontot, a mintaelemeket egyesével ahhoz a klaszterhez soroljuk, amihez hozzávéve SSW legkevesébe nő meg (amelyik középpontjához közelebb van), majd a középpontokat újraszámítjuk. Függ a kezdeti középpontoktól és a sorrendtől.

Hierarchikus módszerek

Vannak a klaszterszámot csökkentő (agglomeratív vagy összevonó) és növelő (divizív vagy felosztó) módszerek. Pl. előbbi esetében kezdetben minden pont külön klaszterbe tartozik, minden lépésben a két legközelebbi klasztert vonjuk össze, míg egy klaszter nem lesz. Függ a klaszterek távolságának definíciójától.

Klaszterek távolsága ($d(C_i, C_j)$)

- ▶ Legközelebbi szomszéd (nearest neighbor vagy single linkage):

$$\min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \|\mathbf{x} - \mathbf{y}\|.$$

- ▶ Legtávolabbi szomszéd (furthest neighbor vagy complete linkage):

$$\max_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \|\mathbf{x} - \mathbf{y}\|.$$

- ▶ Átlagos távolság (between-groups linkage vagy average linkage):

$$\sum_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \|\mathbf{x} - \mathbf{y}\| / (|C_i| \cdot |C_j|).$$

- ▶ Klaszterközpontok távolsága (centroid): $\|\bar{\mathbf{x}}_{C_i} - \bar{\mathbf{x}}_{C_j}\|$.

- ▶ Unió pontjainak átlagos távolsága (within-groups linkage):

$$\frac{1}{\binom{|C_i| + |C_j|}{2}} \sum_{\mathbf{x}, \mathbf{y} \in C_i \cup C_j} \|\mathbf{x} - \mathbf{y}\|.$$

- ▶ Ward:

$$\sum_{\mathbf{x} \in C_i \cup C_j} \|\mathbf{x} - \bar{\mathbf{x}}_{C_i \cup C_j}\|^2 - \sum_{\ell=1}^2 \sum_{\mathbf{x} \in C_\ell} \|\mathbf{x} - \bar{\mathbf{x}}_{C_\ell}\|^2 = \frac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} \|\bar{\mathbf{x}}_{C_i} - \bar{\mathbf{x}}_{C_j}\|^2.$$