

Segédanyag a Matematikai statisztika tantárgyhoz

2021. szeptember 8.

Leíró statisztika

Statisztikai **sokaság**: a megfigyelés tárgyát képező egyedek összessége, halmaza. Röviden sokaságnak hívjuk.

A sokaság egysége: a sokaság egy eleme.

Statisztikai **ismérv** (röv.: ismérv): a sokaság egyedeit jellemző tulajdonság. Az ismérvek típusai:

- minőségi ismérv: az egyedek számszerűen nem mérhető tulajdonsága
- mennyiségi ismérv: az egyedek számszerűen mérhető tulajdonsága. Két fajtájukat különböztetjük meg:
 - diszkrét: véges vagy megszámlálhatóan sok értéket vehet fel
 - folytonos: egy adott intervallumon belül kontinuum számosságú értéket felvehet
- időbeli ismérv: az egységek időbeli elhelyezésére szolgáló rendezőelvek
- területi ismérv: az egységek térbeli elhelyezésére szolgáló rendezőelvek

Mérési szintek:

- Névleges (nominális) mérési skála: a számok csak ún. kódszámok, amik a sokaság egyedeinek azonosítására szolgálnak. Ezek között matematikai relációkat és műveleteket nincs értelme végezni. Pl. a hallgatók neme.
- Sorrendi (ordinális) skála: a sokaság egyedeinek valamely tulajdonság alapján sorba való rendezése. Pl. a hallgatók jegyei egy tárgyból.
- Intervallumskála (különbségi skála): a skálaértékek különbségei is valós információt adnak a sokaság egyedeiről. A skálán a nullpont meghatározása önkényes. Ilyen skálákhoz mértékegység is tartozik. Pl. hőmérséklet.
- Arányskála: a skálának van valódi nullpontja is. Minden matematikai művelet elvégezhető ezekkel a számokkal. Pl. a hallgatók magassága.

Tipikusan a minőségi ismérvek mérési szintje nominális, esetleg sorrendi skála; a mennyiségi ismérvek mérési szintje különbségi vagy arányskála; a területi ismérvek mérési szintje nominális skála; az időbeli ismérvek pedig különbségi skála.

Néha az intervallum- vagy arányskálán mérhető tulajdonságokat *metrikus ismérveknek* nevezik.

Statisztikai tábla a statisztikai sorok összefüggő rendszere.

A statisztikai táblák fajtái:

- Egyszerű tábla: nincs benne összegző sor
- Csoportosító tábla: egyetlen összegző sort tartalmaz
- Kombinációs vagy *kontingenciatábla*: legalább két összegző sort tartalmaz

Definíció. z -kvantilis: $q(z) = q_z = \inf\{x : F(x) \geq z\}$, és amennyiben F invertálható, akkor $q_z = F^{-1}(z)$ -re egyszerűsödik ($0 < z < 1$)

Fontos speciális kvantilisok: **kvantilisok**:

- $Q_1 := q_{\frac{1}{4}} \rightsquigarrow$ alsó kvantilis
- $Q_2 = Me := q_{\frac{1}{2}} \rightsquigarrow$ **medián** (középső mintaelem)
- $Q_3 := q_{\frac{3}{4}} \rightsquigarrow$ felső kvantilis

Definíció. Módusz: abszolút folytonos eloszlás esetén a sűrűségfüggvény maximumhelye(i), diszkrét eloszlás esetén pedig az eloszlás maximumhelye(i). Tehát

- $Mo = \operatorname{argmax}_{x \in \mathbb{R}} f(x)$, ha X abszolút folytonos;
- $Mo = \operatorname{argmax}_{x_1, x_2, \dots} P(X = x_i)$, ha X diszkrét.

Nem biztos, hogy létezik, és ha létezik, akkor se biztos, hogy egyértelmű.

Definíció. Ferdeség (skewness): $\operatorname{skew}(X) = \frac{E(X-EX)^3}{(DX)^3}$

Értelmezése:

- $\operatorname{skew}(X) = 0 \Rightarrow$ az eloszlás szimmetrikus
- $\operatorname{skew}(X) > 0 \Rightarrow$ az eloszlás balra ferdült
- $\operatorname{skew}(X) < 0 \Rightarrow$ az eloszlás jobbra ferdült

Definíció. Csúcsosság (kurtosis): $\operatorname{kurt}(X) = \frac{E(X-EX)^4}{(DX)^4} - 3$

Értelmezés:

- $\operatorname{kurt}(X) = 0 \Rightarrow$ az eloszlás csúcsossága a standard normáliséval megegyező
- $\operatorname{kurt}(X) < 0 \Rightarrow$ az eloszlás laposabb a st. norm.-nál
- $\operatorname{kurt}(X) > 0 \Rightarrow$ az eloszlás csúcsosabb a st. norm.-nál

Minta: X_1, \dots, X_n valószínűségi változó sorozat, jel. $\mathbf{X} = (X_1, \dots, X_n)^T$

A továbbiakban feltesszük, hogy függetlenek és azonos eloszlásúak – ezt röviden *i.i.d. mintának* hívjuk (independent, identically distributed).

Az elméleti értékeket nagy, a konkrét, realizált mintából számolt értékeket mindig kis betű fogja jelölni, azaz minta esetén x_1, \dots, x_n .

Statisztika: a minta valamely függvénye: $T : \mathbf{X} \mapsto \dots$

Becslés: a minta eloszlásának ismeretlen paraméterét közelíti a minta segítségével.

Megi.: Minden becslés statisztika.

Néhány lényeges statisztika:

- **Rendezett minta:** $X_1^* \leq \dots \leq X_n^*$ nem csökkenő sorrendbe tesszük a mintaelemeket
- **Terjedelem:** $R = X_n^* - X_1^*$ (R=range)
- **Mintaátlag:** $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$
- **Tapasztalati szórás:** $S_n = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}$

Értelmezése: az átlagtól való átlagos eltérés abszolút mértékegységben

• *Korrigált tapasztalati szórás:* $S_n^* = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$

• *Szórási együttható:* $V = \frac{S_n}{\bar{X}}$

Értelmezése: az átlagtól való átlagos eltérés százalékban

Megj.: relatív szórásnak is hívják

• *Tapasztalati eloszlásfüggvény:* $F_n(x) = \frac{\sum_{i=1}^n I(X_i < x)}{n}$

ahol $I(X_i < x) = \begin{cases} 1 & \text{ha } X_i < x \\ 0 & \text{ha } X_i \geq x \end{cases} \rightsquigarrow$ karakterisztikus függvény

• *Tapasztalati z-quantilis:* Realizált mintából sokféleképpen számolható, interpolációs módszer:

1.) Sorszám megállapítása: $(n+1)z = e + t$ (e: egészrész, t: törtrész)

2.) $q_z = x_e^* + t(x_{e+1}^* - x_e^*)$

Értelmezése: a mintaelemek z-ed része legfeljebb a q_z értéket veszi fel, $(1-z)$ -ed része pedig legalább q_z .

• *Interkvartilis terjedelem:* $IQR = Q_3 - Q_1$

• *Tapasztalati módusz:* a legtöbbször előforduló érték.

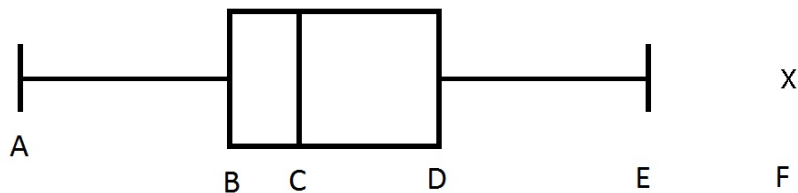
Értelmezése: a minta tipikus, leggyakrabban előforduló értéke.

• *Tapasztalati ferdeség:* $\frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^3}{S_n^3}$

• *Tapasztalati csúcsosság:* $\frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^4}{S_n^4} - 3$

Tétel. (Glivenko-Cantelli) A tapasztalati eloszlásfüggvény 1 valószínűséggel egyenletesen tart a valódi eloszlásfüggvényhez, formálisan $P\left(\limsup_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0\right) = 1$.

Boxplot ábra: (ez fekvő, de lehet álló is) ahol a betűk a következő értékeket jelentik:



- $A = \max\{x_1^*, Q_1 - 1, 5 \cdot IQR\}$;
- $B = Q_1$;

- $C = Me$;
- $D = Q_3$;
- $E = \min\{x_n^*, Q_3 + 1, 5 \cdot IQR\}$;
- F : kieső értékek, azokat tüntetjük fel pontokként, amik A-n vagy E-n kívülre esnek.

Az adatelemzés lépései:

- Adathibák keresése, irreális adatok, értékek törlése; esetleg korrigálása
- Alkalmos osztályközös gyakorisági sor készítése
- Középértékek kiszámítása
 - Átlag (számtani vagy mértani – amelyiknek értelme van)
 - Helyzeti középértékek: módusz és medián
- Szóródási mutatók kiszámítása
 - Terjedelem és interkvartilis terjedelem
 - Szórás és relatív szórás
- Alakmutatók kiszámítása
 - Ferdeség
 - Csúcsosság
- Ábrák készítése:
 - Sűrűséghisztogram
 - Boxplot ábra

Nevezetes diszkrét eloszlások:

Eloszlás neve	Jelölése	Eloszlása	EX	D ² X
Karakterisztikus (indikátorvált.)	Ind(p)	$P(X = 1) = p$ $P(X = 0) = 1 - p$	p	$p(1 - p)$
Geometriai (Pascal)	Geo(p)	$P(X = k) = p(1 - p)^{k-1}$ $k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hipergeometriai	Hipgeo(N, M, n)	$P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$ $k = 0, 1, \dots, n$	$n \frac{M}{N}$	$n \frac{M}{N} (1 - \frac{M}{N}) (1 - \frac{n-1}{N-1})$
Binomiális	Bin(n, p)	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ $k = 0, 1, \dots, n$	np	$np(1 - p)$
Negatív binomiális	NegBin(n, p)	$P(X = k) = \binom{k-1}{n-1} p^n (1-p)^{k-n}$ $k = n, n+1, \dots$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Poisson	Poi(λ)	$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ $k = 0, 1, \dots$	λ	λ

Nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D ² X
Egyenletes	$E(a, b)$	$\begin{cases} 0 & \text{ha } x \leq a \\ \frac{x-a}{b-a} & \text{ha } a < x \leq b \\ 1 & \text{ha } b < x \end{cases}$	$\begin{cases} \frac{1}{b-a} & \text{ha } a < x \leq b \\ 0 & \text{különben} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponenciális	$\text{Exp}(\lambda)$	$\begin{cases} 1 - e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\begin{cases} \lambda e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Standard norm.	$N(0, 1^2)$	$\Phi(x) = \dots$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad x \in \mathbb{R}$	0	1
Normális	$N(m, \sigma^2)$	\dots	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad x \in \mathbb{R}$	m	σ^2

További nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D ² X
Cauchy	$Cauchy(a, b)$ $a \in \mathbb{R}, b > 0$	$\frac{1}{\pi} \arctg\left(\frac{x-a}{b}\right) + \frac{1}{2}$	$\frac{1}{\pi b \left[1 + \left(\frac{x-a}{b}\right)^2\right]} \quad x \in \mathbb{R}$	\nexists	\nexists
Pareto*	$Pareto(\alpha, \beta)$ $\alpha, \beta > 0$	$\begin{cases} 1 - \left(\frac{\beta}{x}\right)^\alpha & \text{ha } x \geq \beta \\ 0 & \text{ha } x < \beta \end{cases}$	$\begin{cases} \frac{\alpha}{\beta} \left(\frac{\beta}{x}\right)^{\alpha+1} & \text{ha } x \geq \beta \\ 0 & \text{ha } x < \beta \end{cases}$	$\frac{\alpha\beta}{\alpha-1}$	$\frac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)}$

* A Pareto-eloszlásnak akkor van véges várható értéke a képletnek megfelelően, ha $\alpha > 1$, szórásnégyzete pedig akkor, ha $\alpha > 2$.

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D ² X
Khí-négyzet	$\chi_k^2 \quad k \in \mathbb{N}$	\dots	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \quad x \in \mathbb{R}$	k	$2k$
Gamma	$\Gamma(\alpha, \lambda) \quad \alpha, \lambda > 0$	\dots	$\begin{cases} \frac{1}{\Gamma(\alpha)} \lambda^\alpha e^{-\lambda x} x^{\alpha-1} & \text{ha } x \geq 0 \\ 0 & \text{ha } x < 0 \end{cases}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Béta	$Beta(\alpha, \beta)$ $\alpha, \beta > 0$	\dots	$\begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} & x \in [0; 1] \\ 0 & \text{különben} \end{cases}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Lognormális	$LN(m, \sigma^2)$ $m \in \mathbb{R}, \sigma > 0$	\dots	$\begin{cases} \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\log x - m)^2}{2\sigma^2}} & \text{ha } x \geq 0 \\ 0 & \text{ha } x < 0 \end{cases}$	$e^{m+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2m+\sigma^2}$

Matematikai statisztika – becslélmélet

Most belekezdünk a matematikai statisztikába, a korábbi minta fogalma egy fokkal absztraktabb formában fog visszaköszönni.

Definíció. Statisztikai mező. $(\Omega, \mathcal{A}, \mathcal{P})$ hármas, ahol \mathcal{P} pedig eloszlások egy családja és minden $P \in \mathcal{P}$ -re (Ω, \mathcal{A}, P) valószínűségi mező.

\mathcal{P} -t gyakran paraméteresen adjuk meg: $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$, ahol $\Theta \subseteq \mathbb{R}^p$ összefüggő és nyílt halmaz, amit **paraméterter**nek hívunk.

Definíció. Minta. $\mathbf{X} : (\Omega, \mathcal{A}) \rightarrow \mathcal{X}$ leképezés, ahol \mathcal{X} neve: **mintatér**.

Feladat: annak a meghatározása, hogy a \mathcal{P} eloszláscsalád melyik tagja írja le legjobban a valóságot, a vizsgált jelenséget. Ennek érdekében veszünk mintát. Erőfeszí-

téseink jelentős része arra fog irányulni, hogy a valamilyen szempontból "legjobb" P -t vagy paraméteres esetben ezzel ekvivalens módon, a "legjobb" ϑ paramétert megtaláljuk.

Jelölés. A továbbiakban a valószínűség, sűrűségfüggvény, várható érték és szórásmátrix) alsó indexben lévő ϑ arra fog utalni, hogy egy paraméteres statisztikai mező van a feladat háttérben és ϑ -val jelöljük az ismeretlen, de érdeklődésünk középpontjában lévő paramétert.

Definíció. Likelihood függvény: Legyen $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. minta

- $L(\vartheta; \mathbf{x}) = f_\vartheta(\mathbf{x}) = \prod_{i=1}^n f_\vartheta(x_i)$, ha az eloszlás folytonos
- $L(\vartheta; \mathbf{x}) = P_\vartheta(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n P_\vartheta(X_i = x_i)$, ha az eloszlás diszkrét

Definíció. Log-likelihood függvény: $l(\vartheta; \mathbf{x}) = \log L(\vartheta; \mathbf{x})$

Legyen $g : \Theta \rightarrow \mathbb{R}^k$ függvény. Célunk az \mathbf{X} minta alapján $g(\vartheta)$ becslése.

Definíció. Torzítatlan becslés. $T(\mathbf{X})$ statisztika torzítatlan becslése $g(\vartheta)$ -nak, ha $E_\vartheta T(\mathbf{X}) = g(\vartheta) \quad \forall \vartheta \in \Theta$ -ra.

Definíció. Torzítás (bias). $b_T(\vartheta) = E_\vartheta T(\mathbf{X}) - g(\vartheta)$

Definíció. Legyenek $T_1(\mathbf{X})$ és $T_2(\mathbf{X})$ torzítatlan becslései $g(\vartheta)$ -nak. Ekkor azt mondjuk, hogy $T_1(\mathbf{X})$ **hatásosabb** $T_2(\mathbf{X})$ -nél, ha $D_\vartheta^2(T_1(\mathbf{X})) \leq D_\vartheta^2(T_2(\mathbf{X}))$ minden $\vartheta \in \Theta$ esetén.

Definíció. Hatásos becslés. A $T(\mathbf{X})$ torzítatlan becslést hatásosnak nevezük, ha minden torzítatlan becslésnél hatásosabb.

Tétel. A hatásos becslés egyértelmősége.

Ha $T_1(\mathbf{X})$ és $T_2(\mathbf{X})$ hatásos becslései $g(\vartheta)$ -nak, akkor minden paraméterértékre 1 valószínűséggel megegyeznek, azaz $P_\vartheta(T_1(\mathbf{X}) = T_2(\mathbf{X})) = 1 \quad \forall \vartheta \in \Theta$ esetén.

Megjegyzés. Egy becslésről nem egyszerű belátni, hogy hatásos. Hatásos becslés keresésének alapja a Blackwell-Rao tétel.

Definíció. Aszimptotikus torzítatlanság. A $T_n(\mathbf{X})$ becsléssorozat ($n = 1, 2, \dots$) aszimptotikusan torzítatlan becslése a $g(\vartheta)$ -nak, ha $E_\vartheta T_n(\mathbf{X}) \xrightarrow{n \rightarrow \infty} g(\vartheta) \quad \forall \vartheta \in \Theta$ esetén.

Definíció. Gyenge konzisztencia. A $T_n(\mathbf{X})$ becsléssorozat ($n = 1, 2, \dots$) gyengén konzisztens becslése a $g(\vartheta)$ -nak, ha $T_n(\mathbf{X}) \xrightarrow[n \rightarrow \infty]{P} g(\vartheta) \quad \forall \vartheta \in \Theta$ esetén.

Tétel. Elégséges feltétel gyenge konzisztenciára.

Ha $E_\vartheta T_n(\mathbf{X}) \xrightarrow{n \rightarrow \infty} g(\vartheta)$ és $D_\vartheta^2 T_n(\mathbf{X}) \xrightarrow{n \rightarrow \infty} 0$, akkor T_n becsléssorozat gyengén konzisztens becslése $g(\vartheta)$ -nak.

Definíció. Erős konzisztencia. A $T_n(\mathbf{X})$ becsléssorozat ($n = 1, 2, \dots$) erősen konzisztens becslése a $g(\vartheta)$ -nak, ha $T_n(\mathbf{X}) \xrightarrow[n \rightarrow \infty]{1 \text{ vsz.}} g(\vartheta) \quad \forall \vartheta \in \Theta$ esetén.

Állítás.

- A tapasztalati eloszlásfüggvény torzítatlan és erősen konzisztens becslése az eloszlásfüggvénynek.
- A mintaátlag torzítatlan és erősen konzisztens becslése a várható értéknek.
- A tapasztalati szórásnégyzet *torzított*, de aszimptotikusan torzítatlan és erősen konzisztens becslése a szórásnégyzetnek.
- A tapasztalati szórásnégyzet torzítatlan és erősen konzisztens becslése a korrigált szórásnégyzetnek.

A sűrűségfüggvény becslése nem triviális probléma, két módszer erre:

- Hisztogram (sűrűséghisztogram)
- Parzen-Rosenblatt becslés

A sűrűségfüggvény becslése magfüggvény segítségével n elemű mintából:

Parzen-Rosenblatt becslés: $f_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)$, ahol

- h_n alkalmas 0-hoz tartó sorozat, ami a mintapont körüli intervallum hossza felének felel meg
- $K(x)$ az úgynevezett magfüggvény; a leggyakrabban a Gauss-féle magfüggvényt használjuk, ami $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

A Parzen-Rosenblatt becslés

Tétel. A Parzen-Rosenblatt becslés konzisztenciája. Alkalmas feltételek esetén h_n -re és a K magfüggvényre, az $f_n(x)$ Parzen-Rosenblatt becslés aszimptotikusan torzítatlan és erősen konzisztens becslése a valódi sűrűségfüggvénynek.

Paraméterbecslési módszerek

- **Maximum likelihood módszer (ML-módszer):** Azt a paraméterértéket keressük, ahol a likelihood fv. a legnagyobb értéket veszi fel: $\max_{\vartheta} L(\vartheta; \mathbf{x})$
Amennyiben a függvény deriválható ϑ szerint, akkor a maximumot kereshetjük a szokásos módon, az első és második deriváltak segítségével, azonban a feladatunkat jelentősen megnehezíti, hogy olyan n -szeres szorzatot kellene deriválni, amelyiknek minden tagjában ott van az a változó, ami szerint deriválnunk kellene. Ezért likelihood függvény helyett a log-likelihood függvény maximumhelyét keressük.
- **Momentum módszer:** A mintából számítható tapasztalati momentumokat ($m_i := \frac{\sum_j x_j^i}{n}$) egyenlővé tesszük az elméleti momentumokkal ($M_i(\vartheta) := E_{\vartheta} X^i$), az elsőől kezdve, mégpedig annyit, amennyi paraméter van. Tehát p darab ismeretlen paraméter esetén a következő p ismeretlenes egyenletrendszert oldjuk meg:

$$\begin{aligned} M_1(\vartheta) &= m_1 \\ &\vdots \\ M_p(\vartheta) &= m_p \\ \text{Megjegyzés: } m_1 &= \bar{x} \end{aligned}$$

Fisher-tétel: Ha ϑ ML-becslése $\hat{\vartheta}$, akkor tetszőleges g függvény esetén $g(\vartheta)$ ML-becslése $g(\hat{\vartheta})$.

Definíció. Elégséges statisztika.

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező, \mathbf{X} minta, $B \in \mathcal{A}$. A T statisztikát **elégséges statisztikának** nevezzük, ha a $P_{\vartheta}(\mathbf{X} \in B | T(\mathbf{X}))$ feltételes eloszlásnak létezik ϑ -tól nem függő változata.

Megjegyzés. Ez egy elég absztrakt fogalom. Elégséges statisztikát a Neyman-féle faktorizációs tétel segítségével (kicsit lejjebb) tudunk keresni és arra lesz jó, hogy segítségével bizonyos szempontból optimális becslést találjunk.

Megjegyzés. az elégséges statisztika minden lényeges információt tartalmaz az ismeretlen ϑ paraméterre vonatkozóan.

Tétel. Neyman-féle faktorizációs tétel.

"Szép" statisztikai mezőn a T statisztika akkor és csak akkor elégséges, ha léteznek olyan g_{ϑ} nemnegatív és h függvények, hogy $L(\vartheta; \mathbf{x}) = g_{\vartheta}(T(\mathbf{x})) \cdot h(\mathbf{x}) \quad \forall \vartheta \in \Theta$ és λ -m.m. $\mathbf{x} \in \mathcal{X}$ esetén.

Állítás. A $T(\mathbf{X}) = \mathbf{X}^*$ rendezett minta elégséges statisztika.

Ebben a részben tegyük fel, hogy a paraméterter 1 dimenziós.

Definíció. Fisher-információ.

Tegyük fel, hogy a log-likelihood függvény ϑ szerint deriválható. Ekkor az \mathbf{X} n elemű mintában lévő Fisher-információ: $I_{\mathbf{X}}(\vartheta) \equiv I_n(\vartheta) = E_{\vartheta} ([\partial_{\vartheta} l(\vartheta; \mathbf{X})]^2)$.

Megi.: $I_{\mathbf{X}}(\vartheta)$ azt az (absztrakt) információmennyiséget méri, amelyet az \mathbf{X} minta a paraméterre vonatkozóan magában hordoz.

A Fisher-információ kiszámítása bizonyos, úgynevezett regularitási feltételek esetén egyszerűbbé válik.

Definíció. 1. regularitási feltétel. $E_{\vartheta}(\partial_{\vartheta} l(\vartheta, \mathbf{X})) = 0$

Állítás. $E_{\vartheta}(\partial_{\vartheta} l(\vartheta, \mathbf{X})) = 0 \iff \partial_{\vartheta} \int_{\mathbf{x} \in \mathcal{X}} f_{\vartheta}(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{x} \in \mathcal{X}} \partial_{\vartheta} f_{\vartheta}(\mathbf{x}) d\mathbf{x}$, azaz "be lehet deriválni" az integráljel mögé.

Állítás. Ha teljesül az 1. regularitási feltétel, akkor a Fisher-információt kiszámolhatjuk az alábbi módon: $I_n(\vartheta) = n \cdot I_1(\vartheta) = n \cdot D_{\vartheta}^2 \partial_{\vartheta} l(\vartheta, X_1)$.

Tétel. Cramér-Rao egyenlőtlenség.

Tegyük fel, hogy $T(\mathbf{X})$ statisztika torzítatlan becslése $g(\vartheta)$ -nak és teljesül az 1.

regularitási feltétel. Ekkor minden $\vartheta \in \Theta$ -ra $D_{\vartheta}^2(T(\mathbf{X})) \geq \underbrace{\frac{(g'(\vartheta))^2}{I_n(\vartheta)}}_{\text{információs határ}}$.

Megjegyzés. Ha minden ϑ -ra egyenlőség teljesül a Cramér-Rao egyenlőtlenységben, akkor T hatásos becslés. Ennek az egyenlőtlenységnek a vizsgálata tehát lehetőséget ad arra, hogy blackwellizálás nélkül hatásos becslést találjunk.

Megjegyzés. Előfordulhat, hogy a statisztika szórásnégyzete nagyobb az információs határnál, viszont a statisztika hatásos. Példa erre i.i.d. exponenciális mintánál az $\frac{n-1}{\sum_{i=1}^n X_i}$ statisztika.

Definíció. χ^2 -eloszlás: Az X val. változó n szabadságfokú χ^2 -eloszlású (jel.: $X \sim \chi_n^2$), ha $X = U_1^2 + \dots + U_n^2$, ahol $U_i \sim N(0, 1) \forall i$ -re és függetlenek egymástól.

Definíció. t-eloszlás: Az X valószínűségi változó n szabadságfokú Student-féle t-eloszlást követ (jel.: $X \sim t_n$), ha $X = \frac{Z}{\sqrt{\frac{Y_n}{n}}}$, ahol $Z \sim N(0, 1)$ és $Y_n \sim \chi_n^2$ függetlenek egymástól.

Definíció. F-eloszlás: Az X valószínűségi változó m és n szabadságfokú F-eloszlást követ (jel.: $X \sim F_{m,n}$), ha $X = \frac{Y_m}{\frac{Z_n}{n}}$, ahol $Y_m \sim \chi_m^2$ és $Z_n \sim \chi_n^2$ függetlenek egymástól.

Mostantól α egy 0-hoz közeli pozitív szám lesz (például $0,05 = 5\%$), és vezessük be a következő jelöléseket az eloszlások kvantiliseire:

- u_{α} : $N(0, 1)$ eloszlás $(1 - \alpha)$ -kvantilise, azaz $u_{\alpha} = \Phi^{-1}(1 - \alpha)$
- $z_{\alpha} := u_{1-\alpha}$ (sok könyvben ezt használják)
- $t_{n,\alpha}$: n szabadságfokú t-eloszlás $(1 - \alpha)$ -kvantilise
- $\chi_{n,\alpha}^2$: n szabadságfokú χ^2 -eloszlás α -kvantilise
- $F_{m,n}^{\alpha}$: m, n szabadságfokú F-eloszlás α -kvantilise

Definíció. Konfidencia intervallum: Adott α -hoz legalább $(1 - \alpha)$ valószínűséggel tartalmazza az adott paramétert (vagy annak egy függvényét): $P_{\vartheta} \left(T_1(\mathbf{X}) < \hat{\vartheta} < T_2(\mathbf{X}) \right) \geq 1 - \alpha$.

Gyakran keresünk szimmetrikus konfidencia intervallumot, ilyenkor $T_1 = T_2 =: \Delta$, és az intervallum $\hat{\vartheta} \pm \Delta$ alakba írható.

Állítás. Legyen $X_1, \dots, X_n \sim N(m, \sigma^2)$ i.i.d. minta. Ekkor

- m -re konfidencia intervallum
 - ha σ ismert, akkor $\bar{x} \pm u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
 - ha σ ismeretlen, akkor $\bar{x} \pm t_{n-1, \frac{\alpha}{2}} \frac{s_n^*}{\sqrt{n}}$
- σ^2 -re konfidencia intervallum: $\left[\frac{(n-1) \cdot (s_n^*)^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1) \cdot (s_n^*)^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right]$

Intervallumbecslés az ismeretlen ϑ paraméterre a paraméter ML-becslésének aszimptotikája alapján n elemű mintából: $\hat{\vartheta} \pm u_{\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{I_n(\hat{\vartheta})}}$

Hipotézisvizsgálat

Hipotézis \sim valami állítás, aminek igazságát vizsgálni szeretnénk

Paramétertér: $\Theta = \Theta_0 \cup^* \Theta_1 \rightarrow$ "valóság"

Mintatér: $\mathcal{X} = \mathcal{X}_e \cup^* \mathcal{X}_k \rightarrow$ "látzat" - MINTÁBÓL

\mathcal{X}_k : kritikus tartomány - azon \mathbf{X} megfigyelések halmaza, amikre *elutasítjuk* a nullhipotézist

\mathcal{X}_e : elfogadási tartomány - azon \mathbf{X} megfigyelések halmaza, amikre *elfogadjuk* a nullhipotézist

Hipotézisvizsgálati feladat:

$H_0 : \vartheta \in \Theta_0 \rightsquigarrow$ nullhipotézis

$H_1 : \vartheta \in \Theta_1 \rightsquigarrow$ ellenhipotézis vagy alternatív hipotézis

Tehát ha $\mathbf{X} \in \mathcal{X}_e$, akkor elfogadjuk H_0 -t; ha $\mathbf{X} \in \mathcal{X}_k$, akkor pedig elutasítjuk H_0 -t. Amennyiben a Θ_0 halmaz egyelemű, akkor azt mondjuk, hogy H_0 egyszerű. H_1 -re ugyanígy.

Az \mathcal{X} mintatér felosztását általában egy statisztika (neve: próbastatisztika) segítségével végezzük el:

legyen $T: \mathcal{X} \rightarrow \mathbb{R}$, $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) > c\}$ c neve: kritikus érték
 $\mathcal{X}_e = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \leq c\}$

"Valóság"	Döntés	
	H_0 -t	
H_0 teljesül (Θ_0)	helyes döntés	elsőfajú hiba
H_0 nem teljesül (Θ_1)	másodfajú hiba	helyes döntés

$P(\text{elsőfajú hiba}) = \alpha(\vartheta) = P_{\vartheta}(\mathcal{X}_k)$, ahol $\vartheta \in \Theta_0$

$P(\text{másodfajú hiba}) = \beta(\vartheta) = P_{\vartheta}(\mathcal{X}_e)$, ahol $\vartheta \in \Theta_1$

Erőfüggvény: $\psi: \Theta_1 \rightarrow \mathbb{R}$, $\psi(\vartheta) = P_{\vartheta}(\mathcal{X}_k)$

Terjedelem: $\alpha = \sup \{\alpha(\vartheta) : \vartheta \in \Theta_0\}$

p-érték: az az α terjedelem, ami esetén a próbastatisztika értéke egyenlő a kritikus értékkel: $T(\mathbf{x}) = c_{\alpha}$.

A p-érték a legkisebb terjedelem, amire még elutasítjuk a H_0 -t. Ha egy próbát számítógép segítségével végzünk el, rendszerint a p-érték révén tudunk dönteni: ha $(p\text{-érték}) < \alpha$, akkor elvetjük H_0 -t.

Ha mind H_0 , mind H_1 egyszerű, akkor adott α terjedelemhez lehet legerősebb próbát találni, ezt pedig úgy hívják, hogy *valószínűség-hányados próba*. A hipoté-

ziseket folytonos esetre írunk fel. Diszkrétre a sűrűségfüggvény helyett a konkrét eloszlást kell írni.

$$H_0 : f = f_0$$

$$H_1 : f = f_1$$

A valószínűség-hányados próba kritikus tartománya: $\mathcal{X}_k = \left\{ \mathbf{x} : \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > c_\alpha \right\}$

Tehát azokat az \mathbf{x} -eket, amire az $\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$ nagy, bepakoljuk a kritikus tartományba egészen addig, míg az adott α terjedelmet el nem érjük. Diszkrét esetben ehhez általában véletlenítésre van szükség, azaz bizonyos \mathbf{x} -ek esetén nem 1 vagy 0, hanem egy, e két szám közé eső (jelöljük p_α -val) valószínűséggel vetjük el a nullhipotézist.

A hipotézisvizsgálat menete próbastatisztika és kritikus érték segítségével:

1. A terjedelem (α) lefixálása, ami jellemzően 1% és 10% közötti, tipikusan 5%
Megbízhatóság = $1 - \alpha$, általában %-osan írjuk
2. Nullhipotézis (H_0) felírása – sokévi, megszokott, elvárt értékeknek megfelelő paramétertartomány
3. Alternatív hipotézis (H_1) felírása – a minta alapján bennünket érdeklő kérdésnek megfelelő paramétertartomány
4. A probléma megoldására alkalmas próba vagy próbák kiválasztása – feltételek ellenőrzése
5. Próbastatisztika kiszámítása
6. Kritikus érték kiszámítása, kritikus tartomány (\mathcal{X}_k) megállapítása
7. Döntés:
 - (a) $\mathbf{x} \in \mathcal{X}_k \rightsquigarrow$ **erős döntés**, H_1 -et elfogadjuk, H_0 -t elvetjük/elutasítjuk
 - (b) $\mathbf{x} \in \mathcal{X}_e \rightsquigarrow$ **gyenge döntés**, H_1 -et elutasítjuk, H_0 -t nem tudjuk elutasítani
8. Szöveges értelmezés: α terjedelem/elsőfajú hiba valószínűsége mellett azt állíthatjuk/nem állíthatjuk, hogy ...

A hipotézisvizsgálat menete p -érték segítségével:

1. A terjedelem (α) lefixálása
2. Nullhipotézis (H_0) felírása
3. Alternatív hipotézis (H_1) felírása
4. A probléma megoldására alkalmas próba vagy próbák kiválasztása
5. p -érték megállapítása, rendszerint számítógép segítségével
6. Döntés:
 - (a) $p\text{-érték} < \alpha \Leftrightarrow \mathbf{x} \in \mathcal{X}_k \Leftrightarrow H_1$ -et elfogadjuk
 - (b) $p\text{-érték} > \alpha \Leftrightarrow \mathbf{x} \in \mathcal{X}_e \Leftrightarrow H_0$ -t nem tudjuk elutasítani
7. Szöveges értelmezés

Most néhány nevezetes próbát mutatunk be a normális eloszlás várható értékére, illetve szórására. Az α végig a próba terjedelmét jelöli, ami előre adott.

I. Próbák normális eloszlás várható értékére

1.) Egymintás próbák

a.) Egymintás u-próba

$X_1, \dots, X_n \sim N(m, \sigma^2)$, ahol σ ismert, m paraméter

$$\begin{array}{lll} \text{a.) } H_0 : m = m_0 & \text{b.) } H_0 : m = m_0 & \text{c.) } H_0 : m = m_0 \\ & H_1 : m \neq m_0 & H_1 : m > m_0 \\ & & H_1 : m < m_0 \end{array}$$

A próbastatisztika: $T(\mathbf{X}) = u = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$ H_0 esetén $\sim N(0, 1)$

A kritikus tartományok:

$$\begin{array}{l} \text{a.) } \mathcal{X}_k = \{ \mathbf{x} : |u| > u_{\alpha/2} \} \\ \text{b.) } \mathcal{X}_k = \{ \mathbf{x} : u > u_\alpha \} \\ \text{c.) } \mathcal{X}_k = \{ \mathbf{x} : u < -u_\alpha \} \end{array}$$

b.) Egymintás t-próba

$X_1, \dots, X_n \sim N(m, \sigma^2)$, ahol σ , m paraméter

$$\begin{array}{lll} \text{a.) } H_0 : m = m_0 & \text{b.) } H_0 : m = m_0 & \text{c.) } H_0 : m = m_0 \\ & H_1 : m \neq m_0 & H_1 : m > m_0 \\ & & H_1 : m < m_0 \end{array}$$

A próbastatisztika: $T(\mathbf{X}) = t = \frac{\bar{X} - m_0}{\frac{s_n^*}{\sqrt{n}}}$ H_0 esetén $\sim t_{n-1}$

A kritikus tartományok:

$$\begin{array}{l} \text{a.) } \mathcal{X}_k = \{ \mathbf{x} : |t| > t_{n-1, \alpha/2} \} \\ \text{b.) } \mathcal{X}_k = \{ \mathbf{x} : t > t_{n-1, \alpha} \} \\ \text{c.) } \mathcal{X}_k = \{ \mathbf{x} : t < -t_{n-1, \alpha} \} \end{array}$$

2.) Kétmintás próbák

$X_1, \dots, X_n \sim N(m_1, \sigma_1^2)$

$Y_1, \dots, Y_m \sim N(m_2, \sigma_2^2)$

Az elvégzendő próbák $H_0 : m_1 = m_2$ nullhipotézis esetén:

	a két minta független		a két minta nem független
σ_1 és σ_2 ismert	a.) kétmintás u-próba		egymintás u-próba a különbségekre
σ_1 és σ_2 ismeretlen	előzetes F-próba		egymintás t-próba a különbségekre
	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	
	b.) kétmintás t-próba	c.) Welch-próba	

a.) kétmintás u-próba

m_1, m_2 paraméterek, σ_1, σ_2 ismert

$H_0 : m_1 = m_2$ és H_1 : ami a szöveggörnyezetben értelmes

A próbastatisztika: $u = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$ H_0 esetén $\sim N(0, 1)$

b.) kétmintás t-próba

$m_1, m_2, \sigma_1 = \sigma_2$ paraméterek

$H_0 : m_1 = m_2$ és H_1 : ami a szöveggörnyezetben értelmes

A próbatatiztika: $t = \sqrt{\frac{nm}{n+m}} \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(n-1)(s_1^*)^2+(m-1)(s_2^*)^2}{n+m-2}}} \stackrel{H_0 \text{ esetén}}{\sim} t_{n+m-2}$

c.) Welch-próba

$m_1, m_2, \sigma_1 \neq \sigma_2$ paraméterek

$H_0 : m_1 = m_2$ és H_1 : ami a szöveggörnyezetben értelmes

A próbatatiztika: $t' = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(s_1^*)^2}{n} + \frac{(s_2^*)^2}{m}}} \stackrel{H_0 \text{ esetén}}{\sim} t_f$, ahol

$$\frac{1}{f} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$$

$$c = \frac{\frac{(s_1^*)^2}{n}}{\frac{(s_1^*)^2}{n} + \frac{(s_2^*)^2}{m}}, \text{ ha } s_1^* > s_2^*$$

II. Próbák normális eloszlás szórására

1.) Egymintás próba: χ^2 -próba

$X_1, \dots, X_n \sim N(m, \sigma^2)$, ahol m és σ ismeretlen paraméterek

$H_0 : \sigma = \sigma_0$ és $H_1 : \sigma \neq \sigma_0$

A próbatatiztika: $h = \frac{(n-1)(S_n^*)^2}{\sigma_0^2} \stackrel{H_0 \text{ esetén}}{\sim} \chi_{n-1}^2$

Kritikus tartomány: $\mathcal{X}_k = \{ \mathbf{x} : h < \chi_{n-1, \alpha/2}^2 \text{ vagy } h > \chi_{n-1, 1-\alpha/2}^2 \}$

Az ellenhipotézis lehet egyoldali is, ilyenkor a kritikus tartomány értelemszerűen módosul.

2.) Kétmintás próba: F -próba

$X_1, \dots, X_n \sim N(m_1, \sigma_1^2)$ $Y_1, \dots, Y_m \sim N(m_2, \sigma_2^2)$ $m_1, m_2, \sigma_1, \sigma_2$ paraméterek

$H_0 : \sigma_1 = \sigma_2$ és H_1 : ami a szöveggörnyezetben értelmes

A próbatatiztika: $F = \frac{(S_1^*)^2}{(S_2^*)^2} \stackrel{H_0 \text{ esetén}}{\sim} F_{n-1, m-1}$

χ^2 -próbák

a.) Diszkrét illeszkedésvizsgálat

Feladat: adott egy $\mathbf{X} = (X_1, \dots, X_n)$ n elemű minta, és azt akarjuk eldönteni, hogy a minta egy általunk "remélt" eloszlásból származik-e. *Diszkrét* illeszkedésvizsgálatnál feltesszük, hogy a mintaelemek r különböző értéket vehetnek fel: $P(X_i = x_j) = p_j$ $j = 1, \dots, r$. Jelöljük N_j -vel a gyakoriságokat, azaz azt, hogy az n elemű mintában hány darab x_j szerepel.

Osztályok	1	2	...	r	Összesen
Valószínűségek	p_1	p_2	...	p_r	1
Gyakoriságok	N_1	N_2	...	N_r	n

H_0 : a valószínűségek: $\mathbf{p}=(p_1, \dots, p_r)$

H_1 : nem ezek a valószínűségek

A próbatatiztika: $T_n = \sum_{i=1}^r \frac{(N_i - np_i)^2}{np_i} \stackrel{H_0 \text{ esetén}}{\rightarrow} \chi_{r-1}^2$ eloszlásban, ha $n \rightarrow \infty$

A kritikus tartomány: $\mathcal{X}_k = \{ \mathbf{x} : T_n(\mathbf{x}) > \chi_{r-1, 1-\alpha}^2 \}$

Becsléses illeszkedésvizsgálat: csak annyit "sejtünk", hogy a minta valamilyen eloszlású, viszont a paramétereiről nincs sejtésünk. Ilyenkor amennyiben ML-módszerrel becsüljük meg az s darab ismeretlen paramétert, akkor a próbatatiztika: $T_n \stackrel{H_0 \text{ esetén}}{\rightarrow} \chi_{r-1-s}^2$ eloszlásban, ha $n \rightarrow \infty$.

Nagyon fontos: a próba csak akkor hajtható végre, amennyiben az egyes osztályokban elegendő számú gyakoriság szerepel. Nem egyértelmű, milyen határvonalat húzzunk meg. Hüvelykujjszabályként azt lehet mondani, hogy legalább 4-6 gyakoriság szerepeljen a cellákban és np_i legalább 4 legyen minden osztályra. Amennyiben kevés gyakoriság van a cellákban, akkor az érintett osztályokat össze kell vonni.

b.) Diszkrét homogenitásvizsgálat

Feladat: van két **független** minta, mindkettő egy közös szempont szerint r osztály egyikébe sorolva. Azt kell eldönteni, hogy a két minta azonos eloszlásúnak tekinthető-e.

Osztályok	1	2	...	r	Összesen
1. minta					
Valószínűségek	p_1	p_2	...	p_r	1
Gyakoriságok	N_1	N_2	...	N_r	n
2. minta					
Valószínűségek	q_1	q_2	...	q_r	1
Gyakoriságok	M_1	M_2	...	M_r	m

H_0 : a valószínűségek: $(p_1, \dots, p_r) = (q_1, \dots, q_r)$

H_1 : nem ezek a valószínűségek

A próbatat.: $T_{n,m} = nm \sum_{i=1}^r \frac{(\frac{N_i}{n} - \frac{M_i}{m})^2}{\frac{N_i + M_i}{n+m}} \stackrel{H_0 \text{ esetén}}{\rightarrow} \chi_{r-1}^2$ eloszlásban, ha $n \rightarrow \infty$

A kritikus tartomány: $\mathcal{X}_k = \{ \mathbf{x} : T_{n,m}(\mathbf{x}) > \chi_{r-1, 1-\alpha}^2 \}$

c.) Függetlenségvizsgálat

Feladat: van egy minta, két szempont szerint csoportosítva. Azt kell eldönteni, hogy a két szempont független-e egymástól.

$p_{i,j}$ =P(egy megfigyelés az (i,j) osztályba kerül)

$N_{i,j}$ =ennyi megfigyelés kerül az (i,j) osztályba

A mintavétel eredménye:

	2. szempont					Összesen	
	1	...	j	...	s		
1. szempont	1	N_{11}	...	N_{1j}	...	N_{1s}	$N_{1\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	i	N_{i1}	...	N_{ij}	...	N_{is}	$N_{i\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	r	N_{r1}	...	N_{rj}	...	N_{rs}	$N_{r\bullet}$
Összesen	$N_{\bullet 1}$...	$N_{\bullet j}$...	$N_{\bullet s}$	n	

ahol $N_{i\bullet} = \sum_{j=1}^s N_{ij}$ és $N_{\bullet j} = \sum_{i=1}^r N_{ij}$

H_0 : a szempontok függetlenek, azaz $p_{i,j} = p_{i\bullet} \cdot p_{\bullet j} \forall i, j$ -re

H_1 : nem azok

A próbat statisztika: $T_n = n \left(\sum_{i=1}^r \sum_{j=1}^s \frac{N_{ij}^2}{N_{i\bullet} N_{\bullet j}} - 1 \right) \xrightarrow{H_0 \text{ esetén}} \chi_{(r-1)(s-1)}^2$ eloszlásban,

ha $n \rightarrow \infty$

A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi_{(r-1)(s-1), 1-\alpha}^2\}$

Ha $r = s = 2$, akkor a próbat statisztika $T_n = n \cdot \frac{(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\bullet}N_{2\bullet}N_{\bullet 1}N_{\bullet 2}}$ -re egyszerűsödik, az aszimptotikus eloszlás pedig 1 szabadságfokú χ^2 .

Folytonos illeszkedésvizsgálat

Azt akarjuk ellenőrizni, hogy egy X_1, \dots, X_n független, azonos eloszlású minta egy adott (fix paraméterű) folytonos eloszlásból származik-e. Tehát formálisan

$H_0 : F_{X_1}(x) = F(x) \forall x \in \mathbb{R}$, ahol F egy adott eloszlás eloszlásfüggvénye

$H_1 : \exists x \in \mathbb{R} : F_{X_1}(x) \neq F(x)$

Kolmogorov-Szmirnov próba

Próbat statisztika: $\sqrt{n}D_n(\mathbf{x}) = \sqrt{n} \cdot \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$

A próbat statisztika eloszlása H_0 esetén az ún. Kolmogorov-eloszláshoz tart (n $\rightarrow \infty$), melynek eloszlásfüggvénye $1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2}$. Jelöljük K_α -val

a Kolmogorov-eloszlás α -kvantilisét.

A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{X} : \sqrt{n}D_n(\mathbf{X}) > K_{1-\alpha}\}$

Megj.: D_n kiszámításához elég csak a mintapontokban, azaz F_n ugrópontjaiban tekinteni az eltérést: $D_n(\mathbf{x}) = \max_{1 \leq i \leq n} \max \left\{ \left| \frac{i-1}{n} - F(x_i^*) \right|, \left| \frac{i}{n} - F(x_i^*) \right| \right\}$, ahol x_i^*

az i -edik rendezett mintaelem ($i = 1, \dots, n$).

Megj.: azt is megtehetjük, hogy a mintából osztályközös gyakorisági sort hozunk létre – azaz mesterséges osztályozást készítünk –, majd χ^2 -próbát hajtunk végre.

Ezt az eljárást *diszkrétizálás*nak hívjuk.

A Kolmogorov-Szmirnov próba messze nem a legjobb próba folytonos illeszkedésvizsgálatra, sokkal inkább a legegyszerűbb és legrégebbi. További próbák folytonos illeszkedésvizsgálatra:

- **Anderson-Darling próba:** nagyon hasonló a Kolmogorov-Szmirnov próbához, de jóval nagyobb nála az ereje. Fő erőssége: megfelelő paraméterezéssel beállítható, az eloszlás alsó vagy felső részére koncentrálnunk-e.
- **Shapiro-Wilk próba:** a legerősebb próba annak ellenőrzésére, hogy a minta normális eloszlásból származik-e. A standard normális eloszlás rendezett mintáján alapul.

Nemparaméteres próbák az u -próbák/ t -próbák helyett

Ezek a próbák az egymintás és kétmintás u -próbák/ t -próbák helyett hajtandók végre, amennyiben nem teljesül, hogy a minta normális eloszlású (például vannak benne kiugró értékek). Az egymintás próba erre a célra az előjel próba és a Wilcoxon-próba, a kétmintás esetben pedig független minták esetén a Mann-Whitney próba. Kétmintás esetben amennyiben összefüggők a minták, akkor a különbségekre Wilcoxon-próbát vagy előjel próbát hajtunk végre.

Ezeket a próbákat mi csak folytonos eloszlásból származó mintákra fogjuk használni, ami a lejjebb szereplő formulákban is tükröződni fog, de kisebb módosításokkal a próbák diszkrét eloszlásokra is alkalmazhatók.

1.) Egymintás próbák

Adott egy X_1, \dots, X_n folytonos eloszlásból származó minta. A nullhipotézis azt állítja, hogy a medián egy előre megadott m_0 számmal egyenlő-e. Az ellenhipotézis egy- és kétoldali is lehet.

A nullhipotézis: $H_0 : m = m_0$, ahol m a mediánt jelöli. Másképp: $H_0 : P(X - m_0 > 0) = 1/2$.

a.) Előjel próba

A próbat statisztika: $T = \sum_{i=1}^n I(X_i > m_0) \rightsquigarrow$ megszámoljuk, hány mintaelem nagyobb m_0 -nál. A próbat statisztika H_0 teljesülése esetén $Bin(n, \frac{1}{2})$ eloszlású, ezáltal a p -érték

- kétoldali ellenhipotézis esetén: $\frac{2}{2^n} \sum_{i=0}^{\min(T, n-T)} \binom{n}{i}$
- baloldali ellenhipotézis esetén: $\frac{1}{2^n} \sum_{i=0}^T \binom{n}{i}$
- jobboldali ellenhipotézis esetén: $\frac{1}{2^n} \sum_{i=0}^{n-T} \binom{n}{i}$

b.) Wilcoxon-próba

A próba végrehajtásához legyen $Y_i := |X_i - m_0|$, $i = 1, \dots, n$. Jelölje R_i az i -edik mintaelem rangját: $R_i = \sum_{j=1}^n I(Y_i \geq Y_j)$.

A próbastatisztika azon rangok összege, amikre $X_i - m_0$ pozitív: $T^+ = \sum_{i: X_i > m_0} R_i$

A kritikus érték(ek) meghatározása viszonylag macerás. Kis mintákra speciális táblázatokból kell kinézni az értékeket, nagy mintákra pedig a nullhipotézis esetén a lenormált próbastatisztika közel standard normális eloszlású, így ennek a kvantiliseit lehet használni.

2.) Kétmintás próbák

Adott egy X_1, \dots, X_n minta és egy tőle független Y_1, \dots, Y_m minta, mindkettő folytonos eloszlásból. A nullhipotézis azt állítja, hogy a két eloszlás mediánja megegyezik-e. Az ellenhipotézis egy- és kétoldali is lehet.

A nullhipotézis formálisan: $H_0 : P(X > Y) = P(X < Y)$.

a.) Mann-Whitney próba

A próbastatisztika: $W = \sum_{i=1}^n \sum_{j=1}^m I(X_i > Y_j)$.

A kritikus érték(ek) meghatározása hasonlóan megy, mint a Wilcoxon-próbánál. Kis mintákra speciális táblázatokból kell kinézni az értékeket, nagy mintákra pedig a próbastatisztikát le kell normálni és használhatjuk a standard normális eloszlás kvantiliseit.

Regressziószámítás

Lineáris regresszió

Feladat: adottak $(x_1, y_1), \dots, (x_n, y_n)$ pontok, ezekre szeretnénk egyenest illeszteni (neve: *regressziós egyenes*) legkisebb négyzetek módszerével, azaz keressük azon a és b konstansokat, melyekre $\sum_{i=1}^n (y_i - (ax_i + b))^2$ minimális.

A modell: $y_i = ax_i + b + \varepsilon_i$, ahol $E\varepsilon_i = 0$ és $D^2\varepsilon_i = \sigma^2 < \infty$ ($i = 1, \dots, n$)

Paraméterbecslés: $\hat{a} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$, $\hat{b} = \bar{y} - \hat{a}\bar{x}$

Reziduuumok (becsült hibák): $\hat{\varepsilon}_i = y_i - \hat{a}x_i - \hat{b}$ $i = 1, \dots, n$

Reziduális négyzetösszeg: $\text{RNÖ} = \sum \hat{\varepsilon}_i^2 = \sum (y_i - \bar{y})^2 - \frac{\sum(x_i - \bar{x})(y_i - \bar{y})^2}{\sum(x_i - \bar{x})^2}$

A hibák szórásnégyzetének becslése: $\hat{\sigma}^2 = \frac{\text{RNÖ}}{n-2}$

A becsült paraméterek szórásnégyzete: $D^2(\hat{a}) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$, $D^2(\hat{b}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)$

Tapasztalati korrelációs együttható: $R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$. Ennek négyzetét, R^2 -et, *determinációs együtthatónak* hívjuk, és ezzel mérjük a modell jóságát. Az R^2 mutatja meg, hogy százalékban a modell az Y változékonyságából mennyit

magyaráz meg. Értéke 0 és 1 között lehet, ha 0-hoz közeli, akkor a modell gyengén teljesít, ha 1-hez, akkor jól.

A lineáris regresszió egy modellezési eszköz. A statisztikai modellek is olyanok, mint a hétköznapi értelemben használt modellek: csak leképezni próbálják a valóságot, de visszaadni nem tudják. Ebből kifolyólag körültekintéssel kell megalkotni minden modellt.

A lineáris regresszió feltevései:

- **Linearitás:** a magyarázó változó és a kimeneti változó magyarázó változó értékeire feltételes átlagai között lineáris az összefüggés
- **Homoszkedaszticitás:** a reziduálisok (hibatag) σ szórása megegyezik a magyarázó változó minden értékére (azaz nem függ az i indextől)
- **Függetlenség:** adataink egyszerű véletlen mintából származnak, tehát függetlenek egymástól
- **Normalitás:** a magyarázóváltozó minden rögzített értékére, a kimeneti változó normális eloszlású. Ez gyakorlatilag megegyezik a reziduálisok normalitásával.

Hipotézisvizsgálat a lineáris modellben – a modell jóságára vonatkozó tesztek:

$H_0 : a = 0$

$H_1 : a \neq 0$

Próbastatisztika: $t = \frac{\hat{a}}{D(\hat{a})}$

H_0 esetén $t \sim t_{n-1}$, azaz t -próbát kell végezni a t próbastatisztika segítségével.

Amennyiben el tudjuk vetni, hogy a regressziós egyenes meredeksége 0, akkor azt mondjuk, hogy szignifikáns az összefüggés a két változó között.

A legtöbb statisztikai programcsomag azt a próbát is végrehajtja, ami az összes regressziós együttható együttes nullaságát vizsgálja. Sima lineáris regressziónál:

$H_0 : a = b = 0$

$H_1 : a \neq 0$ vagy $b \neq 0$

Ezeket a hipotéziseket egy F-próbával lehet ellenőrizni. A nullhipotézis elvetése esetén azt mondhatjuk, hogy a modell nem teljesen értelmetlen. Ha ezzel az F-próbával nem tudjuk elvetni a nullhipotézist, akkor abszolút nincs értelme az adott modellel dolgozni.

Dummy (bináris) változó: két értéket vesz fel (például férfi és nő). Ilyenkor az egyik értékhez célszerű 0-t, a másikhoz 1-et rendelni. A regressziós elemzés eredményei továbbra is érvényben maradnak, amennyiben a magyarázóváltozó dummy változó.

Többváltozós regresszióelemzés

Egy n elemű mintából a következő adatok állnak rendelkezésre: $(y_i, x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$, azaz van egy kimeneti és p magyarázóváltozónk. Ebben az esetben is alkalmazható lineáris regresszió, a modell a

következőféleképpen írható fel:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n \quad \text{általában } n \gg p$$

ε továbbra is a fentebbi módon definiált hibatag, β_0 a tengelymetszet, amit az eddigiekben b -vel jelöltünk. A modell ebben az esetben nem egy egyenest, hanem egy hipersíkot generál.

y elnevezései: eredményváltozó, függő változó, endogén változó

x_i -k elnevezései: magyarázóváltozók, független változók, exogén változók

Az együtthatók kiszámolásához írjuk fel a modellt mátrix alakban: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, ahol

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

A legkisebb négyzetek becslését ebben az esetben a következőféleképpen kapjuk:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

Ehhez persze fel kell tennünk az inverz létezését. Azt az esetet, mikor nem létezik az inverz itt nem tárgyaljuk. $X^T X$ inverze létezik abban az esetben, ha a magyarázóváltozók lineárisan függetlenek, azaz egyik magyarázóváltozót sem tudjuk kifejezni a többi lineáris kombinációjaként. Amennyiben fennáll valamilyen szintű (lineáris) összefüggés a magyarázó változók között, abban az esetben *multikollinearitás*ról beszélünk. Ezt a gyakorlatban érdemes megvizsgálni, mert hatással lehet a modell validitására (azaz, hogy mennyire írja le jól a valóságot, valóban azt adják-e vissza, amit várunk).

Reziduálisok: $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$

Reziduális négyzetösszeg: $\text{RNÖ} := \|\hat{\boldsymbol{\varepsilon}}\|^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Teljes négyzetösszeg: $\text{NÖ} = \sum_{i=1}^n (y_i - \bar{y})^2$

Determinációs együttható: $R^2 = 1 - \frac{\text{RNÖ}}{\text{NÖ}} = \frac{\text{NÖ} - \text{RNÖ}}{\text{NÖ}} \rightsquigarrow$ az eredményváltozó változékonyságának hány %-át magyarázza a regressziós modell. Értéke 0 és 1 között lehet. Minél nagyobb, annál jobb.

Gyakori modellválasztási kritériumok:

- Korrigált determinációs együttható: $R_{\text{adj}}^2 = 1 - \frac{n-1}{n-r-1} \frac{\text{RNÖ}}{\text{NÖ}} \rightsquigarrow$ minél nagyobb, annál jobb
- **Akaike-féle információs kritérium:** $AIC = 2k - 2 \log \hat{L}$, ahol
 - k : a becsülendő paraméterek száma, a regressziós modellben $k = p + 1$

– \hat{L} a likelihood-függvény értéke akkor, ha az ML-becslést használjuk (normális eloszlású hibáknál ez megegyezik a legkisebb négyzetes becsléssel) Minél kisebb, annál jobb.

- **Bayes-féle információs kritérium:** $BIC = \log n \cdot k - 2 \log \hat{L} \rightsquigarrow$ minél kisebb, annál jobb

A regresszióelemzés lépései:

- az eredményváltozó(k) és a lehetséges magyarázóváltozók kiválasztása
- adatgyűjtés
- adattisztítás, adathibák korrekciója
- pontdiagrammal a potenciális modellek kiválasztása (lineáris, négyzetes, logisztikus stb.)
- paraméterbecslés
- modelldiagnosztika – az együtthatók szignifikanciája, a modell együttes jó-sága
- legjobb modell kiválasztása, "modellépítés" – több módszer/mutató közül választhatunk: korrigált R^2 , cross-validation, AIC/BIC információs kritériumok stb.
- előrejelzés

A továbbiakban néhány fontos speciális témát ismertetünk röviden a regresszió témakörön belül.

Logisztikus regresszió (logit modell)

Olyan regresszió, amikor az eredményváltozó egy valószínűség, arra szeretnénk regressziós modellt felépíteni. Legyen Y egy 0 vagy 1 értékű valószínűségi változó, az $Y = 1$ esemény jelöli egy bizonyos esemény bekövetkeztét, míg $Y = 0$ azt, hogy nem következik be. Jelölje $q = P(Y = 1)$ -et, ezt szeretnénk magyarázni. Odds-hányadosnak nevezzük a $\frac{q}{1-q}$ hányadost, ami azt mutatja meg, hány-szoros eséllyel következik be az a bizonyos esemény ahhoz képest, hogy nem következik be. Az odds-hányados logaritmusát logodds-hányadosnak hívják és az a fő feltevésünk (reményünk), hogy ezt lineáris modellel megfelelően tudjuk már magyarázni, azaz ha q az eredményváltozó, akkor a következő modellt építjük fel: $\log\left(\frac{q}{1-q}\right) = b_0 + b_1 x_1 + \dots + b_p x_p$, ahol b_0, \dots, b_p az ismeretlen, becsülendő paraméterek.

A $g(x) = \log\left(\frac{x}{1-x}\right)$ függvényt logit függvénynek szokták hívni, innen a modellt logit modellnek keresztelték el. Itt ezt a g függvényt *linkfüggvénynek* hívják, mert nem a q -t modellezzük lineáris modellel, hanem $g(q)$ -t, tehát g függvény teremt meg a kapcsolatot az eredményváltozó és a magyarázóváltozók között.

A paraméterek becslése csak numerikus módszerekkel lehetséges, nincs rájuk "szép" képlet, szemben az előző fejezetben tárgyalt lineáris modellnél.

Szóráselmzés (ANOVA) / vegyes kapcsolat elemzése

A szóráselmzés a lineáris modell egyik legfontosabb alkalmazása. Az eljárásnak több elnevezése van: szóráselmzés = variancia-analízis = ANOVA (analysis of variance).

A modell: $y_{i,j} = b_i + \varepsilon_{i,j}$, ahol

- $i = 1, \dots, p \rightsquigarrow$ csoportok vagy osztályok száma
- $j = 1, \dots, n_i \rightsquigarrow$ mintaelemszám egy osztályon belül
- $n = \sum_{i=1}^p n_i \rightsquigarrow$ teljes mintaelemszám
- $\varepsilon_{i,j} \sim N(0, \tau^2)$ függetlenek, ahol $\tau > 0$

Feladat: annak eldöntése, hogy $b_1 = \dots = b_k$, azaz a csoporthoz tartozás nem befolyásolja az ismerv értékét.

Vezessünk be jelöléseket:

- $\bar{y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i,j} \rightsquigarrow$ részátlagok vagy csoportátlagok
- $\bar{y}_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{i,j} = \frac{1}{n} \sum_{i=1}^p n_i \bar{y}_{i\bullet} \rightsquigarrow$ teljes átlag
- $\sigma_i = \sqrt{\frac{1}{n_i} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{i\bullet})^2} \rightsquigarrow$ részszerések
- $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{\bullet\bullet})^2} \rightsquigarrow$ teljes szórás
- $SSB = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{i\bullet})^2 = \sum_{i=1}^p n_i \sigma_i^2 \rightsquigarrow$ csoportokon belüli eltérés-négyzetösszeg
- $SSK = \sum_{i=1}^p n_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \rightsquigarrow$ csoportok közötti teljes eltérés-négyzetösszeg
- $SST = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{\bullet\bullet})^2 \rightsquigarrow$ teljes eltérés-négyzetösszeg

Állítás. $SST = SSB + SSK$

A kapcsolat erősségét mérő mutató a **szórásnégyzet-hányados**: $H^2 = 1 - \frac{SSB}{SST} = \frac{SSK}{SST}$

Tulajdonságai:

- $H^2 = 0$ esetén a két ismerv között nincs kapcsolat, DE (!) ekkor nem feltétlen függetlenek egymástól (analógia: korrelálatlanságból nem következik a függetlenség)
- $H^2 = 1$ esetén a két ismerv között függvényszerű kapcsolat van
- $0 < H^2 < 1$ esetén a két ismerv között sztochasztikus kapcsolat van

- erős a kapcsolat, ha $H = \sqrt{H^2}$ közel van 1-hez és gyenge a kapcsolat, ha 0-hoz

Megjegyzés: ez nem más, mint a regresszió az R^2 .

Tekintsük az alábbi hipotézisvizsgálati feladatot: $H_0 : b_1 = \dots = b_k$
 $H_1 : \text{nem igaz } H_0$

A hipotézisekről egy F -próbával döntünk, a F próbastatisztika kiszámításához az ún. **ANOVA táblázat**ot szokás elkészíteni:

Szóródás forrása	Szabadságfok	Négyzetösszegek	Tapasztalati szórásnégyzetek	
Külső (between group)	$p - 1$	SSK	$MSK = \frac{SSK}{p-1}$	$F = \frac{MSK}{MSB} = \frac{\frac{SSK}{p-1}}{\frac{SSB}{n-p}}$
Belső (within group)	$n - p$	SSB	$MSB = \frac{SSB}{n-p}$	
Teljes	$n - 1$	SST	-	

Tétel. Ha teljesülnek a modell feltételei és a H_0 nullhipotézis, akkor $F \sim F_{p-1, n-p}$, azaz a próbastatisztika F -eloszlást követ.

Állítás: $\frac{\bar{y}_{i\bullet} - b_i}{\sqrt{MSB}} \sqrt{n_i} \sim t_{n-p}$ és $\frac{\bar{y}_{i\bullet} - \bar{y}_{j\bullet} - (b_i - b_j)}{\sqrt{MSB}} \sqrt{\frac{n_i n_j}{n_i + n_j}} \sim t_{n-p}$

Ezek alapján konfidenciaintervallumokat lehet készíteni b_i együtthatókra és a $b_i - b_j$ különbségekre.

Konfidenciaintervallumok:

- b_i -re: $\bar{y}_{i\bullet} \pm t_{n-p; \alpha/2} \sqrt{\frac{MSB}{n_i}}$
- $b_i - b_j$ -re: $\bar{y}_{i\bullet} - \bar{y}_{j\bullet} \pm t_{n-p; \alpha/2} \sqrt{MSB} \sqrt{\frac{n_i + n_j}{n_i n_j}}$