

## Segédanyag a Matematikai statisztika tantárgyhoz

2019. május 16.

### Leíró statisztika

A statisztikai elemzések egyik legfontosabb eszközei a viszonyszámok. A **viszony-szám** két statisztikai adat hányadosa. Jelölések:  $V = \frac{A}{B}$ , ahol  $V$ : viszonyszám;  $A$ : a viszonyítás tárgya;  $B$ : a viszonyítás alapja.

A viszonyszámok fajtái:

- Megoszlási: a sokaság egy részét a sokaság egészéhez viszonyítjuk.
- Koordinációs: a sokaság egy részének a sokaság egy másik részéhez való viszonyítása.
- Dinamikus: két időpont vagy időszak adatának hányadosa.
- Intenzitási: különböző fajta adatok viszonyítása egymáshoz; gyakran a mértékegységük is eltérő.

Ha egy teljes sokaságra és annak  $m$  részére rendelkezésre áll a viszonyszám alapja és részei, akkor a viszonyszámokat ki tudjuk számolni a teljes sokaságra (jel.  $\bar{V}$ , ezt *összetett viszony-szám*nak hívják) és annak részeire is (jel.  $V_1, \dots, V_m$ ). Ekkor a teljes sokaságra számolt viszonyszám kiszámítási lehetőségei:

$$\bar{V} = \frac{\sum_{i=1}^m A_i}{\sum_{i=1}^m B_i} = \frac{\sum_{i=1}^m B_i V_i}{\underbrace{\sum_{i=1}^m B_i}_{\text{súlyozott számtani átlag}}} = \frac{\sum_{i=1}^m A_i}{\underbrace{\sum_{i=1}^m \frac{A_i}{V_i}}_{\text{súlyozott harmonikus átlag}}}$$

A leíró statisztikai szakirodalomban az  $i$  indexeket – pongyola módon – le szokták hagyni:

$$\bar{V} = \frac{\sum A}{\sum B} = \frac{\sum BV}{\sum B} = \frac{\sum A}{\sum \frac{A}{V}}$$

**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:  $T : \mathbf{X} \mapsto \dots$

**Beclés:** a minta eloszlásának ismeretlen paraméterét közelíti a minta segítségével.

Megj.: Minden beclés statisztika.

Néhány lényeges statisztika:

- **Rendezett minta:**  $X_1^* \leq \dots \leq X_n^*$  nem csökkenő sorrendbe tesszük a minta-elemeket
- **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.

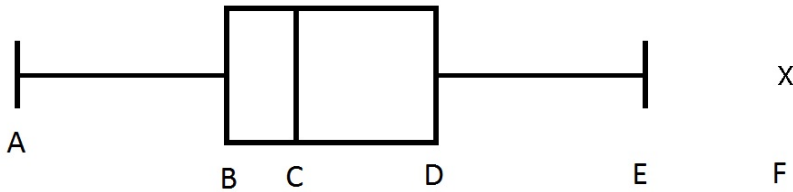
Értelmezése: a minta tipikus, leggyakrabban előforduló értéke.

• **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 pontoként, 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 – amelyeknek értelme van)
  - Helyzeti középértékek: módusz és medián
- Szóródási mutatók kiszámítása
  - Terjedelem és interkvartilis terjedelem
  - Szórás és relatív szórás
- Alakmutatók kiszámítása
  - Ferdeség
  - Csúcsosság
- Ábrák készítése:
  - Sűrűség-histogram
  - Boxplot ábra

Nevezetes diszkrét eloszlások:

Eloszlás neve	Jelölése	Eloszlása	EX	D <sup>2</sup> 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)	$P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$ $k = 0, 1, \dots, n$	$n \frac{M}{N}$	$n \frac{M}{N} \left(1 - \frac{M}{N}\right) \left(1 - \frac{n-1}{N-1}\right)$
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 <sup>2</sup> 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	Exp(λ)	$\begin{cases} 1 - e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\begin{cases} \lambda e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Standard norm.	N(0, 1 <sup>2</sup> )	$\Phi(x) = \dots$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ $x \in \mathbb{R}$	0	1
Normális	N(m, σ <sup>2</sup> )	$\dots$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$ $x \in \mathbb{R}$	m	σ <sup>2</sup>

További nevezetes abszolút folytonos eloszlások:

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

## Matematikai statisztika – becsléelmélet

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

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

$\mathcal{P}$ -t gyakran paraméresen 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. Erőfeszítéseink jelentős része arra fog irányulni, hogy a valamilyen szempontból "legjobb"  $P$ -t vagy paraméteres esetben ezzel ekvivalens módon, a "legjobb"  $\vartheta$  paramétert megtaláljuk.

**Jelölés.** A továbbiakban a valószínűség, sűrűségfüggvény, várható érték és szórásmátrix) alsó indexben lévő  $\vartheta$  arra fog utalni, hogy egy paraméteres statisztikai mező van a feladat hátterében és  $\vartheta$ -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. 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  $\vartheta$ -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  $\vartheta$  paraméterre vonatkozóan.

**Tétel. Neyman-féle faktorizációs tétel.**

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

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

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 nevezzük, ha minden torzítatlan becslésnél hatásosabb.

**Tétel. A hatásos becslés egyértelműsége.**

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 később tárgyalandó 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.

**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. Ez felel meg a mintapont körüli intervallum hossza felének.

**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.

Most egy kis emlékeztető következik arról, mi az a feltételes várható érték és hogyan kell kiszámítani (blackwellizáláshoz szükséges).

Az  $E(X|Y)$ -ra úgy gondolunk, mint egy valószínűségi változóra, konkrétan az  $Y$  valószínűségi változó egy mérhető  $h(Y)$  függvényére; és ha  $Y$  egy adott értéket vesz fel, azaz ha  $E(X|Y = y)$ , akkor mint konkrét számra. Számítása

- diszkrét esetben:

$$E(g(X)|Y) = \sum_i g(x_i)P(X = x_i|Y = j) \Big|_{j=Y}, \text{ ami azt jelenti, hogy először}$$

kiszámoljuk a végtelen összeget a feltételben lévő  $Y$  valószínűségi változó egy konkrét  $j$  értéke esetén, majd a  $j$  helyére visszaírjuk az  $Y$ -t.

- abszolút folytonos eloszlások esetén:

$$E(g(X)|Y) = \int_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)dx \Big|_{y=Y}$$

ahol  $f_{X|Y}(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{ha } f_Y(y) > 0 \\ 0 & \text{különben} \end{cases}$  a feltételes sűrűségfüggvény

**Állítás. Teljes valószínűség tétele folytonos esetben.**

Legyen  $A$  tetszőleges esemény,  $Y$  abszolút folytonos valószínűségi változó. Ekkor  $P(A) = \int_{-\infty}^{\infty} P(A|Y = y)f_Y(y) dy$ .

**Definíció. Teljes statisztika.** A  $T$  statisztika teljes, ha tetszőleges  $h$  mérhető függvényére  $E_{\vartheta}(h(T)) = 0 \quad \forall \vartheta$ -ra  $\iff h(T) = 0 \quad \mathcal{P}$ -majdnem mindenütt

**Tétel. Blackwell-Rao tétel.**

Tegyük fel, hogy  $T$  statisztika torzítatlan becslése  $g(\vartheta)$ -nak és  $S$  elégséges statisztika. Ekkor  $E(T|S)$  feltételes várható érték torzítatlan becslése  $g(\vartheta)$ -nak és  $D_{\vartheta}^2(E(T|S)) \leq D_{\vartheta}^2(T)$  minden  $\vartheta$ -ra.

Ha  $S$  még teljes is, akkor  $E(T|S)$  hatásos becslés.

Tehát a Blackwell-Rao tétel értelmében az alábbi lépéseket követve tudunk hatásos becslést kapni (az eljárás neve *blackwellizálás*):

1. az adott statisztikai mezőn keresünk egy  $S$  teljes elégséges statisztikát;
2. keresünk egy, a bennünket érdeklő  $g(\vartheta)$ -t torzítatlanul becslő  $T$  statisztikát;
3. kiszámoljuk az  $E(T|S)$  feltételes várható értéket – ami az  $S$  függvénye –, és ez lesz a hatásos becslés.

## 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  $\vartheta$  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 ( $M_i := E_{\vartheta} X^i$ ), az elsőtől kezdve, mégpedig annyit, amennyi paraméter van. Tehát  $p$  darab ismeretlen paraméter esetén a következő  $p$  ismeretlenes egyenlet-rendszert oldjuk meg:

$$\begin{aligned} M_1 &= m_1 \\ &\vdots \\ M_p &= m_p \end{aligned}$$

Megjegyzés:  $m_1 = \bar{x}$

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

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

**Definíció. Fisher-információ.**

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

Megj.:  $I_{\mathbf{X}}(\vartheta)$  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.**  $E_{\vartheta}(\partial_{\vartheta}l(\vartheta, \mathbf{X})) = \mathbf{0}$

**Állítás.**  $E_{\vartheta}(\partial_{\vartheta}l(\vartheta, \mathbf{X})) = \mathbf{0} \iff \partial_{\vartheta} \int_{\mathbf{x} \in \mathcal{X}} f_{\vartheta}(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{x} \in \mathcal{X}} \partial_{\vartheta} f_{\vartheta}(\mathbf{x}) d\mathbf{x}$ , 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  $\vartheta$ -ra egyenlőség teljesül a Cramér-Rao egyenlőtlenségben, akkor  $T$  hatásos becslés. Ennek az egyenlőtlenségnek a vizsgálata tehát lehetőséget ad arra, hogy blackwellizálás nélkül hatásos becslést találjunk.

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

**Definíció.  $\chi^2$ -eloszlás:** Az  $X$  val. változó  $n$  szabadságfokú  $\chi^2$ -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  $X = \frac{\frac{Y_m}{m}}{\frac{Z_n}{n}}$ , ahol  $Y_m \sim \chi_m^2$  és  $Z_n \sim \chi_n^2$  függetlenek egymástól.

Mostantól  $\alpha$  egy 0-hoz közeli pozitív szám lesz (például  $0,05 = 5\%$ ), és vezessük

be a következő jelöléseket az eloszlások kvantiliseire:

- $u_{\alpha}$ :  $N(0, 1)$  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  $\chi^2$ -eloszlás  $\alpha$ -kvantilise
- $F_{m,n}^{\alpha}$ :  $m, n$  szabadságfokú F-eloszlás  $\alpha$ -kvantilise

**Definíció. Konfidencia intervallum:** Adott  $\alpha$ -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.

**Definíció.** Legyen  $X_1, \dots, X_n$  minta egy  $\vartheta$  ismeretlen paraméterű eloszláscsaládból. A  $T(\mathbf{X})$  statisztikát **pivotal statisztikának** hívjuk, ha eloszlása nem függ a  $\vartheta$  paramétertől.

Például ha van egy  $N(m, \sigma^2)$  eloszlású mintánk, akkor a  $T(\mathbf{X}) = \sqrt{n} \frac{\bar{X} - m}{\sigma}$  statisztika pivotal, ugyanis standard normális eloszlású, tehát eloszlása nem függ a paraméterektől.

A pivotal statisztikák egyik fő haszna, hogy segítségükkel sok esetben konfidenciaintervallumot lehet készíteni. Erre példa a következő néhány nagyon fontos konfidenciaintervallum.

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

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

**Hipotézisvizsgálat**

Hipotézis  $\sim$  valami állítás, aminek igazságát vizsgálni szeretnénk

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

Mintatér:  $\mathcal{X} = \mathcal{X}_e \cup^* \mathcal{X}_k \longrightarrow$  "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

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  $\Theta_0$  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:

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

$$\mathcal{X}_e = \{\underline{x} \in \mathcal{X} : T(\underline{x}) \leq c\}$$

Döntés	$H_0$ -t	
	elfogadjuk ( $\mathcal{X}_e$ )	elutasítjuk ( $\mathcal{X}_k$ )
"Valóság"		
$H_0$ teljesül ( $\Theta_0$ )	helyes döntés	elsőfajú hiba
$H_0$ nem teljesül ( $\Theta_1$ )	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:**  $\alpha = \sup \{\alpha(\vartheta) : \vartheta \in \Theta_0\}$

Azt mondjuk, hogy az 1-es próba *erősebb* a 2-es próbánál, ha  $\alpha_1 = \alpha_2$  és  $\psi_1(\vartheta) \geq \psi_2(\vartheta) \forall \vartheta \in \Theta_1$ .

**Próbafüggvény:**  $\varphi: \mathcal{X} \rightarrow [0,1] \rightsquigarrow$  ennyi valószínűséggel vetem el a  $H_0$ -t a minta alapján

$$\mathbf{x} \in \mathcal{X}_k \Rightarrow \varphi(\underline{x}) = 1$$

$$\mathbf{x} \in \mathcal{X}_e \Rightarrow \varphi(\underline{x}) = 0$$

**p-érték:** az az  $\alpha$  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  $\alpha$  terjedelemez lehet legerősebb próbát találni, ezt pedig úgy hívják, hogy *valószínűség-hányados próba*. A hipotéziseket folytonos esetre íróm 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:  $\mathcal{X}_k = \left\{ \mathbf{x} : \frac{f_1(\underline{x})}{f_0(\underline{x})} > c_{\alpha} \right\}$

Tehát azokat az  $\mathbf{x}$ -eket, amire az  $\frac{f_1(\underline{x})}{f_0(\underline{x})}$  nagy, bepakoljuk a kritikus tartományba egészen addig, míg az adott  $\alpha$  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  $p_{\alpha}$ -val) valószínűséggel vetjük el a nullhipotézist.

Most néhány nevezetes próbát mutatunk be a normális eloszlás várható értékére, illetve szórására. Az  $\alpha$  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  $\sigma$  ismert,  $m$  paraméter

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

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

A kritikus tartományok:

$$\text{a.) } \mathcal{X}_k = \{\mathbf{x} : |u| > u_{\alpha/2}\}$$

$$\text{b.) } \mathcal{X}_k = \{\mathbf{x} : u > u_{\alpha}\}$$

$$\text{c.) } \mathcal{X}_k = \{\mathbf{x} : u < -u_{\alpha}\}$$

#### b.) Egymintás t-próba

$X_1, \dots, X_n \sim N(m, \sigma^2)$ , ahol  $\sigma, m$  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:  $T(\mathbf{X}) = t = \frac{\bar{X} - m_0}{s_n^*/\sqrt{n}} \stackrel{H_0 \text{ esetén}}{\sim} t_{n-1}$

A kritikus tartományok:

$$\text{a.) } \mathcal{X}_k = \{\mathbf{x} : |t| > t_{n-1, \alpha/2}\}$$

$$\text{b.) } \mathcal{X}_k = \{\mathbf{x} : t > t_{n-1, \alpha}\}$$

$$\text{c.) } \mathcal{X}_k = \{\mathbf{x} : t < -t_{n-1, \alpha}\}$$

### 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
$\sigma_1$ és $\sigma_2$ ismert	a.) kétmintás u-próba	egymintás u-próba a különbségekre
$\sigma_1$ és $\sigma_2$ ismeretlen	előzetes F-próba	
	b.) kétmintás t-próba	c.) Welch-próba
		egymintás t-próba a különbségekre

#### a.) kétmintás u-próba

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

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

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

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

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

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

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

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

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

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

## $\chi^2$ -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 próbatasztika:  $T_n = \sum_{i=1}^r \frac{(N_i - np_i)^2}{np_i} \stackrel{H_0 \text{ esetén}}{\sim} \chi_{r-1}^2$  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 próbatasztika:  $T_n \stackrel{H_0 \text{ esetén}}{\sim} \chi_{r-1-s}^2$  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.

*Illeszkedésvizsgálat "szemmel"*: **Q-Q plot** és **P-P plot**

Jelölje  $F$  az illesztett eloszlás eloszlásfüggvényét,  $x_k^*$  pedig a  $k$ . rendezett mintaelemet.

**Q-Q plot**: az illesztett eloszlás kvantiliseit vetjük össze a tapasztalati kvantilisekkel, azaz a következő pontokat ábrázoljuk:  $\left( F^{-1}\left(\frac{k}{n+1}\right), x_k^* \right)$ , ahol  $k = 1, \dots, n$ .

**P-P plot**: az illesztett eloszlás valószínűségeit vetjük össze a tapasztalati valószínűségekkel, azaz a következő pontokat ábrázoljuk:  $\left( \frac{k}{n+1}, F(x_k^*) \right)$ , ahol  $k = 1, \dots, n$ .

Mindkét ábránál be szokták húzni a 45 fokos egyenest és minél jobban rásimulnak a pontok az egyenesre, annál jobbnak tekinthető az illeszkedés.

### 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
<b>1. minta</b>					
Valószínűségek	$p_1$	$p_2$	...	$p_r$	1
Gyakoriságok	$N_1$	$N_2$	...	$N_r$	n
<b>2. minta</b>					
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 próbatast.:  $T_{n,m} = \sum_{i=1}^r \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{\frac{N_i + M_i}{n+m}} \stackrel{H_0 \text{ esetén}}{\sim} \chi_{r-1}^2$  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	$N_{11}$	...	$N_{1j}$	...	$N_{1s}$	$N_{1\bullet}$
	...	...	...	...	...	...	...
	i	$N_{i1}$	...	$N_{ij}$	...	$N_{is}$	$N_{i\bullet}$
	...	...	...	...	...	...	...
	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 próbatasztika:  $T_n = n \left( \sum_{i=1}^r \sum_{j=1}^s \frac{N_{ij}^2}{N_{i\bullet} N_{\bullet j}} - 1 \right) \xrightarrow{H_0 \text{ esetén}} \chi_{(r-1)(s-1)}^2$  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 próbatasztika  $T_n = n \cdot \frac{(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\bullet}N_{2\bullet}N_{\bullet 1}N_{\bullet 2}}$ -re egyszerűsödik, az aszimptotikus eloszlás pedig 1 szabadságfokú  $\chi^2$ .

### Regressziószámítás

Feladat:  $Y$  valószínűségi változót szeretnénk közelíteni  $X$  val. változó lineáris függvénye segítségével:

$$E[Y - (aX + b)]^2 \rightarrow \min_{a,b} \rightsquigarrow \text{Megoldása: } a_{opt} = \frac{Cov(X,Y)}{D^2(X)}$$

$$b_{opt} = EY - a_{opt}EX$$

Feladat (**lineáris regresszió**): Adottak  $(x_1, y_1), \dots, (x_n, y_n)$  pontok, ezekre szeretnénk egyenest illeszteni (neve: *regressziós egyenes*) legkisebb négyzetek módszerével.

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

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

Reziduukok:  $\hat{\varepsilon}_i = y_i - \hat{a}x_i - \hat{b} \quad i = 1, \dots, n$

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

A hibák szórásnégyzetének becslése:  $\hat{\sigma}^2 = \frac{RN\ddot{O}}{n-2}$

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.

### Többváltozós regresszióelemzés

Legyenek  $Y, X_1, \dots, X_p$  véges szórású valószínűségi változók, amik egy véletlen jelenség egy-egy jellemzői. A regresszióelemzés célja a bennünket különösen érdeklő  $Y$  valószínűségi változó "minél jobb" közelítése az  $X_1, \dots, X_p$  valószínűségi változók segítségével.

$Y$  elnevezései: eredményváltozó, függő változó, endogén változó

$X_i$ -k elnevezései: magyarázó változók, független változók, exogén változók

Általában megfigyeléseink vannak, amik az  $(Y, X_1, \dots, X_p)^T$  valószínűségi vektorváltozó realizációinak tekinthetők:

$$(y_i, x_{i,1}, \dots, x_{i,p})^T \quad i = 1, 2, \dots, n \quad \text{általában } n \gg p$$

Feltehetjük, hogy az  $y_i$  megfigyelések rendszerint mérési eredmények, amik sajnos pontatlanok. A mérési hibát  $\varepsilon_i$ -vel fogjuk jelölni, amiről természetes feltétel, hogy legyen 0 várható értékű és egy véges  $\sigma$  szórású valószínűségi változó.

A **lineáris modell**:  $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$ , ahol

- $\mathbf{y} = (y_1, \dots, y_n)^T$
- $\mathbf{X} = \begin{bmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{bmatrix}$
- $\mathbf{b} = (b_1, \dots, b_p)^T$
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$

Paraméterbecslés:  $\hat{\mathbf{b}} = (X^T X)^{-1} X^T \mathbf{y}$

Projekció az  $F := \text{Im}X$  altérre:  $P_F = X(X^T X)^{-1} X^T$

Becsült értékek:  $\hat{\mathbf{y}} := X\hat{\mathbf{b}}$

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

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

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

Determinációs együttható:  $R^2 = 1 - \frac{RN\ddot{O}}{N\ddot{O}} = \frac{N\ddot{O} - RN\ddot{O}}{N\ddot{O}} \rightsquigarrow$  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:  $R_{\text{adj}}^2 = 1 - \frac{n-1}{n-r-1} \frac{SSR}{N}$   $\rightsquigarrow$  minél nagyobb, annál jobb
- **Akaike-féle információs kritérium:**  $AIC = 2k - 2 \log \widehat{L}$ , ahol
  - $k$ : a becsülendő paraméterek száma, a regressziós modellben  $k = p + 1$
  - $\widehat{L}$  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:**  $BIC = \log n \cdot k - 2 \log \widehat{L}$   $\rightsquigarrow$  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

Attól függően, hogy az eredmény-, illetve a magyarázóváltozó diszkrét-e vagy folytonos, az alábbi statisztikai módszerek használandók a kapcsolat vizsgálatára:

		Az eredményváltozó	
		diszkrét	abszolút folytonos
A magyarázóváltozó	diszkrét	asszociáció $\chi^2$ -próba	vegyes kapcsolat $t$ -próba, ANOVA
	absz. folyt.	osztályozási eljárások, diszkriminancia analízis	korreláció regresszió