Segédanyag a Leíró és matematikai statisztika tantárgyhoz

2017. március 28.

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:
- időbeli ismérv: az egységek időbeli elhelyezésére szolgáló rendezőelvek
- területi ismérv: az egységek térbeli elhelyezésére szolgáló rendezőelvek

Statisztikai sor tágabb értelemben: a sokaság egyes jellemzőinek felsorolása. A statisztikai sorok fajtái:

- Csoportosító sor: a sokaság egy megkülönböztető ismérv szerinti osztályozásának eredménye; az adatok összegezhetők (van 'Összesen' sor)
- Összehasonlító sor: a sokaság *egy részének* a sokaságot egy megkülönböztető ismérv szerinti osztályozásának eredménye; az adatok nem összegezhetők
- Leíró sor: különböző fajta, gyakran eltérő mértékegységű statisztikai adatokat tartalmaz

Az ismérvek fajtája szerint beszélhetünk minőségi, mennyiségi, időbeli és területi sorokról. Például ha egy statisztikai sor tartalmazza az osztályteremben a hallgatókat nemek szerint, akkor ez minőségi csoportosító sor.

Statisztikai tábla tágabb értelemben: a statisztikai sorok összefüggő rendszere. A tábla dimenziószáma az a szám, amennyi statisztikai sorhoz egy-egy táblabeli adat tartozik. Általában 2, maximum 3 dimenziós táblákkal dolgozunk, ennél magasabb dimenziósat már nehéz áttekinteni.
A statisztikai táblák fajtái:

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

A statisztikai elemzések egyik legfontosabb eszközei a viszonyszámok. A **viszonyszám** 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. $\overline{V}$, ezt *összetett viszonyszám*nak hívják) és annak részeire is (jel. $V_1, \dots, V_m$). Ekkor a teljes sokaságra számolt viszonyszám kiszámítási lehetőségei:

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

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

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

Idősorok elemzése (alapok)

Idősorok fajtái:

- állapotidősor: a benne lévő adatok egy-egy adott időpontra vonatkoznak (pl. egy cég raktárkészlete adott napokon);
- tartamidősor: a benne lévő adatok időszakra vonatkoznak (pl. egy cég havi nyereségei).

Véges idősor: $Y_1, \dots, Y_n$, ahol Y_i-k valószínűségi változók. Ezek realizációját, konkrét értékeit jelöljük $y_1, \dots, y_n$-nel. Az idősor megfigyelt értékeiből számíthatunk dinamikus viszonyszámokat. A din. viszonyszámok fajtái:

- Bázisviszonyszámok: $b_t = \frac{y_t}{y_b}$, ahol $t = 1, \dots, n$; b fix, neve: bázisidőszak;
- Láncviszonyszámok: $l_t = \frac{y_t}{y_{t-1}}$, ahol $t = 2, \dots, n$.

Állítás. A bázisviszonyszámok idősorából ki lehet számítani a láncviszonyszámok idősorát és fordítva:

- láncból bázis: $b_t = l_2 \cdot l_3 \cdot \ldots \cdot l_t$ $(t = 1, \dots, n)$;
- bázisból lánc: $l_t = \frac{b_t}{b_{t-1}}$ $(t = 2, \dots, n)$.

Az idősor átlagos értékének kiszámítása:

- tartamidősor esetén sima számtani átlaggal: $\overline{y} = \frac{\sum_{t=1}^{n} y_t}{n}$
- állapotidősor esetén kronologikus átlaggal: $\overline{y}_k = \frac{\frac{1}{2}y_1 + \sum_{t=2}^{n-1} y_t + \frac{1}{2}y_n}{n-1}$

Az idősor átlagos változásának vizsgálata:

- a fejlődés átlagos mértéke: $\overline{d} = \frac{y_n - y_1}{n-1}$
- a fejlődés átlagos üteme: $\overline{l} = \sqrt[n-1]{\frac{y_n}{y_1}}$

Mennyiségi sorok elemzése

Mennyiségi sor készítése:

- Ha a mennyiségi ismérv diszkrét és viszonylag kevés ismérvérték van, akkor minden ismérvértéket felsorolunk.
- Ha a mennyiségi ismérv folytonos vagy sok ismérvérték van, akkor *osztályközös gyakorisági sort* készítünk. Jelölje n a sokaság elemszámát. Az osztályközök meghatározása nem egyértelmű, gyakran választják az osztályok számának a $k = \lfloor \log_2 n \rfloor$ értéket. Ha azonos hosszúságú (h) osztályközöket akarunk létrehozni, akkor $h = \frac{x_{\max} - x_{\min}}{k}$.

Standard jelölések osztályközös gyakoriságú mennyiségi soroknál:

- $x_{i,a}$: az i. osztályköz alsó határa;
- $x_{i,f}$: az i. osztályköz felső határa;
- x_i: az i. osztályközép, azaz $x_i = \frac{x_{i,a} + x_{i,f}}{2}$;
- f_i: gyakoriság az i. osztályközben;
- f_i': kumulált gyakoriság az i. osztályközben, azaz $f_i' = \sum_{k=1}^{i} f_i$;
- g_i: relatív gyakoriság az i. osztályközben, azaz $g_i = \frac{f_i}{\sum_i f_i}$;
- g_i': kumulált relatív gyakoriság az i. osztályközben;
- s_i: az i. osztályköz értékösszege: $z_i = x_i \cdot f_i$;
- s_i' az i. osztályköz kumulált értékösszege.
- z_i: az i. osztályköz relatív értékösszege: $z_i = \frac{s_i}{\sum_i s_i}$;
- z_i' az i. osztályköz kumulált relatív értékösszege.

Koncentráció: a sokasághoz tartozó teljes értékösszeg jelentős része a sokaság kevés egységére összpontosul.

Legyen a sokaság n elemű, a minket érdeklő ismérv szerint a különböző ismérvértékek $x_1, ..., x_k$, ezek gyakoriságai pedig legyenek f_j-k ($\sum_j f_j = n$).

Gini-együttható: $G = \frac{1}{n(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} f_i f_j |x_i - x_j|$.

Lorenz-görbe: a koncentráció mértékét szemléltető ábra. A vízszintes tengelyen a g_i' kumulált relatív gyakoriságok, a függőleges tengelyen a z_i' kumulált relatív értékösszegek szerepelnek, 0-től 100%-ig. Behúzzuk a 45 fokos egyenest. Végül megrajzoljuk a $(0,0), (g_1', z_1'), (g_2', z_2'), \ldots, (g_k', z_k'), (1,1)$ pontok összekötésével kapott töröttvonalat. Koncentrációs területnek hívjuk a töröttvonal és az átló által közbezárt területet.

Erős a koncentráció, ha a töröttvonal közel van a négyzet oldalaihoz. Gyenge a koncentráció, ha a töröttvonal közel van az átlóhoz.

A koncentráció mutatószámai:

- *Koncentrációs együttható*: $L = \frac{G}{2\overline{x}}$
  Ez nem más, mint a koncentrációs terület 2-szerese.
  Értéke 0 és 1 között van. Minél nagyobb, annál erősebb a koncentráció.
- *Herfindahl-index*: $HI = \sum_{i=1}^{k} z_i^2$
  Értéke $\frac{1}{k}$ és 1 közötti; minél nagyobb, annál erősebb a koncentráció.

Nevezetes diszkrét eloszlások:

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

Nevezetes abszolút folytonos eloszlások:

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

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

Fontos speciális kvantilisek: kvartilisek:

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

Definíció. **Módusz:** abszolút folytonos eloszlás esetén a sűrűségfüggvény maxi-

mumhelye(i), diszkrét eloszlás esetén pedig az eloszlás maximumhelye(i). Tehát

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

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

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

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

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

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

Minta: $X_1,\ldots,X_n$ valószínűségi változó sorozat, jel. $\mathbf{X}=(X_1,\ldots,X_n)^T$
A továbbiakban feltesszük, hogy függetlenek és azonos eloszlásúak – ezt röviden *i.i.d. mintá*nak hívjuk (independent, identically distributed).
Az elméleti értékeket nagy, a konkrét, realizált mintából számolt értékeket mindig kis betű fogja jelölni, azaz minta esetén $x_1,...,x_n$.

Statisztika: a minta valamely függvénye: $T:\mathbf{X}\mapsto ...$
Becslés: a minta eloszlásának ismeretlen paraméterét közelíti a minta segítségével.
Megj.: Minden becslés statisztika.

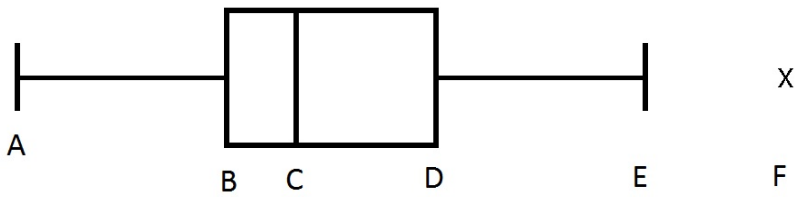
Néhány lényeges statisztika:

- *Rendezett minta*: $X_1^*\le ...\le X_n^*$ nem csökkenő sorrendbe tesszük a mintaelemeket
- *Terjedelem*: $R=X_n^*-X_1^*$ (R=range)
- *Mintaátlag*: $\overline{X}=\frac{\sum_{i=1}^n X_i}{n}$
- *Tapasztalati szórás*: $S_n=\sqrt{\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n}}$
  Értelmezése: az átlagtól való átlagos eltérés abszolút mértékegységben
- *Korrigált tapasztalati szórás*: $S_n^*=\sqrt{\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n-1}}$
- *Szórási együttható*: $V=\frac{S_n}{\overline{X}}$
  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 $I(X_i<x)=\begin{cases}1 & \text{ha } X_i<x\\ 0 & \text{ha } X_i\ge x\end{cases}$ $\rightsquigarrow$ karakterisztikus függvény
- *Tapasztalati z-kvantilis*: 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.
  Osztályközös gyakorisági sorban rendelkezésre álló minta esetén a következő becsést lehet használni: keressük meg kumulálással azt az osztályközt, ahol a q_z van, sorszám: $(n+1)z$. Jelölje j az osztályköz számát. Ezután $q_z=x_{j,a}+\frac{z\cdot(n+1)-f'_{j-1}}{f_j}h_j$
  - $x_{j,a}$: a kvantilist tartalmazó osztályköz alsó értéke;
  - h_j: a kvantilist tartalmazó osztályköz hossza;
  - f'_{j-1}: a kvantilist közvetlenül megelőző osztályköz osztályköz kumulált gyakorisága
  - f_j: a kvantilist tartalmazó osztályköz gyakorisága
- *Interkvartilis terjedelem*: $IQR=Q_3-Q_1$
- *Tapasztalati módusz*: a legtöbbször előforduló érték.
  Értelmezése: a minta tipikus, leggyakrabban előforduló értéke.
  Osztályközös gyakoriságok esetén interpolációra van szükség, ekkor a következő becslést lehet használni: Mo= $x_{mo,a}+\frac{d_a}{d_a+d_f}\cdot h_{mo}$, ahol
  - $x_{mo,a}$: a móduszt tartalmazó osztályköz alsó értéke;
  - h_{mo}: a móduszt tartalmazó osztályköz hossza;
  - d_a: a móduszt tartalmazó osztályköz gyakorisága mínusz a móduszt közvetlenül megelőző osztályköz gyakorisága
  - d_f: a móduszt tartalmazó osztályköz gyakorisága mínusz a móduszt közvetlenül követő osztályköz gyakorisága
- *Tapasztalati ferdeség*: $\frac{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^3}{S_n^3}$
- *Tapasztalati csúcsosság*: $\frac{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^4}{S_n^4}-3$

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

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
- Alkalmas osztályközös gyakorisági sor készítése
- Középértékek kiszámítása
  - Átlag (számtani vagy mértani – amelyiknek értelme van)
  - Helyzeti középértékek:
    - Módusz az osztályközös gyakorisági sorból
    - Medián
- Szóródási mutatók kiszámítása
  - Terjedelem
  - Interkvartilis terjedelem
  - Szórás
  - Relatív szórás
- Alakmutatók kiszámítása
  - Ferdeség
  - Csúcsosság
- Ábrák készítése:
  - Sűrűséghisztogram
  - Boxplot ábra
  - Lorenz-görbe (értékösszeg sor esetén)

Becsléselmélet

Paramétertér: Θ, ahol $\Theta \subseteq \mathbb{R}^p$ összefüggő és nyílt halmaz.

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. 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 $D^2_\vartheta(T_1(\mathbf{X})) \leq D^2_\vartheta(T_2(\mathbf{X}))$ 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.

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

Definíció. **Gyenge konzisztencia:** A $T_n(\mathbf{X})$ becsléssorozat $(n = 1, 2, ...)$ gyengén konzisztens becslése a $g(\vartheta)$-nak , ha $T_n(\mathbf{X}) \xrightarrow[n\to\infty]{\text{sztochasztikusan}} g(\vartheta) \quad \forall \vartheta \in \Theta$ esetén. Másképpen: $\forall \epsilon > 0$-ra $P_\vartheta(|T_n(\mathbf{X}) - g(\vartheta)| \geq \epsilon) \xrightarrow{n\to\infty} 0 \quad \forall \vartheta \in \Theta$ esetén.

Tétel. **Elégséges feltétel gyenge konzisztenciára.** Ha $E_\vartheta T_n(\mathbf{X}) \xrightarrow{n\to\infty} g(\vartheta)$ és $D^2_\vartheta T_n(\mathbf{X}) \xrightarrow{n\to\infty} 0$, akkor T_n becsléssorozat gyengén konzisztens becslése $g(\vartheta)$-nak.

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

Állítás.

- Az eloszlásfüggvény torzítatlan és erősen konzisztens becslése a tapasztalati eloszlásfüggvény.
- A várható érték torzítatlan és erősen konzisztens becslése a mintaátlag.
- A szórásnégyzet aszimptotikusan torzítatlan és erősen konzisztens becslése a tapasztalati szórásnégyzet.
- A szórásnégyzet torzítatlan és erősen konzisztens becslése a korrigált tapasztalati szórásnégyzet.

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.

Definíció. **Likelihood függvény:** Legyen $\mathbf{X} = (X_1, ..., X_n)$ 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:** $l(\vartheta, \mathbf{x}) = \log(L(\vartheta, \mathbf{x}))$.

Paraméterbecslési módszerek

- **Maximum likelihood módszer (ML-módszer):** Azt a paraméterértéket keressük, ahol a likelihood függvény 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.
  Ha ϑ 1 dimenziós, akkor az
  - elsőrendű feltétel: $\partial_\vartheta l(\vartheta, \mathbf{x}) = 0 \quad \rightsquigarrow \quad \hat{\vartheta}$
  - másodrendű feltétel: $\partial^2_\vartheta l(\vartheta, \mathbf{x}) < 0$

  Ha ϑ p dimenziós, akkor $\vartheta = (\vartheta_1, ..., \vartheta_p)$, az
  - elsőrendű feltétel:
    $\partial_{\vartheta_i} l(\vartheta, \mathbf{x}) = 0 \quad \rightsquigarrow \quad \hat{\vartheta}_i \ \ (i = 1, ..., p) \quad \rightsquigarrow \quad \hat{\vartheta} = (\hat{\vartheta}_1, ..., \hat{\vartheta}_p)$
  - másodrendű feltétel: $H(\vartheta_1, ..., \vartheta_p) = \left(\partial_{\vartheta_i}\partial_{\vartheta_j} l(\vartheta, \mathbf{x})\right)_{i,j=1,...,p}$ Hesse-mátrix negatív definit a $\vartheta = \hat{\vartheta}$ helyen
- **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 egyenletrendszert oldjuk meg:
  $$M_1 = m_1$$
  $$\vdots$$
  $$M_p = m_p$$
  Megjegyzés: $m_1 = \overline{x}$

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

Definíció. **χ^2-eloszlás:** Az X valószínűségi változó n szabadságfokú χ^2-eloszlást követ (jel.: $X \sim \chi^2_n$), ha $X = U_1^2 + ... + U_n^2$, ahol $U_i \sim N(0,1)$ minden 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 $Y_n \sim \chi^2_n$ függetlenek egymástól.

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

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

- 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
- $\chi^2_{n,\alpha}$: n szabadságfokú χ^2-eloszlás α-kvantilise
- $F^\alpha_{m,n}$: 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.

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

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

Konfidencia intervallum a valószínűségre (p) nagy minta esetén, ha normális eloszlással közelítünk: $\hat{p} \pm u_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Hipotézisvizsgálat

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

Paramétertér: $\Theta = \Theta_0 \cup^* \Theta_1 \longrightarrow$ "valóság"
Mintatér: $\mathcal{X} = \mathcal{X}_e \cup^* \mathcal{X}_k \longrightarrow$ "látszat" - 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:
$H_0: \ \vartheta \in \Theta_0 \quad \rightsquigarrow$ nullhipotézis
$H_1: \ \vartheta \in \Theta_1 \quad \rightsquigarrow$ 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 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: $\mathcal{X} \to \mathbb{R}$, $\quad \mathcal{X}_k = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) > c\}$ $\qquad c$ neve: kritikus érték
$\mathcal{X}_e = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) \leq c\}$

Döntés / "Valóság"	H_0-t	
	elfogadjuk ($\mathcal{X}_e$)	elutasítjuk ($\mathcal{X}_k$)
H_0 teljesül (Θ_0)	helyes döntés	elsőfajú hiba
H_0 nem teljesül (Θ_1)	másodfajú hiba	helyes döntés

P(elsőfajú hiba)=$\alpha(\vartheta)$=$\mathrm{P}_\vartheta(\mathcal{X}_k)$, ahol $\vartheta \in \Theta_0$
P(másodfajú hiba)=$\beta(\vartheta)$=$\mathrm{P}_\vartheta(\mathcal{X}_e)$, ahol $\vartheta \in \Theta_1$
Erőfüggvény: ψ: $\Theta_1 \to \mathbb{R}$, $\psi(\vartheta) = P_\vartheta(\mathcal{X}_k)$
Terjedelem: $\alpha = \sup\{\alpha(\vartheta)\colon \vartheta \in \Theta_0\}$

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: φ: $\mathcal{X} \to$[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 α terjedelem, ami esetén a próbastatisztika értéke egyenlő a kritikus értékkel : $\mathrm{T}(\mathbf{x}) = c_\alpha$.
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-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 hipotéziseket folytonos esetre írom fel. Diszkrétre a sűrűségfüggvény helyett a konkrét eloszlást kell írni.
$H_0:\ f = f_0$
$\underline{H_1:\ f = f_1}$
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 $\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$ 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.

Néhány konkrét próba – az α végig a próba terjedelmét jelöli, ami előre adott

1.) Egymintás próbák

a.) Egymintás u-próba

$X_1, \ldots, X_n \sim \mathrm{N}(m, \sigma^2)$, ahol σ ismert, m paraméter

a.) $H_0 : m = m_0$ / $\underline{H_1 : m \neq m_0}$ $\qquad$ b.) $H_0 : m = m_0$ / $\underline{H_1 : m > m_0}$ $\qquad$ c.) $H_0 : m = m_0$ / $\underline{H_1 : m < m_0}$

A próbastatisztika: $\mathrm{T}(\mathbf{X}) = u = \sqrt{n}\frac{\overline{X} - m_0}{\sigma} \overset{H_0 \text{ esetén}}{\sim} N(0,1)$
A kritikus tartományok:
a.) $\mathcal{X}_k = \{\mathbf{x} : |u| > u_{\alpha/2}\}$
b.) $\mathcal{X}_k = \{\mathbf{x} : u > u_\alpha\}$
c.) $\mathcal{X}_k = \{\mathbf{x} : u < -u_\alpha\}$

b.) Egymintás t-próba

$X_1, \ldots, X_n \sim \mathrm{N}(m, \sigma^2)$, ahol σ, m paraméter

a.) $H_0 : m = m_0$ / $\underline{H_1 : m \neq m_0}$ $\qquad$ b.) $H_0 : m = m_0$ / $\underline{H_1 : m > m_0}$ $\qquad$ c.) $H_0 : m = m_0$ / $\underline{H_1 : m < m_0}$

A próbastatisztika: $\mathrm{T}(\mathbf{X}) = t = \sqrt{n}\frac{\overline{X} - m_0}{s_n^*} \overset{H_0 \text{ esetén}}{\sim} \mathrm{t}_{n-1}$
A kritikus tartományok:
a.) $\mathcal{X}_k = \{\mathbf{x} : |t| > t_{n-1,\alpha/2}\}$
b.) $\mathcal{X}_k = \{\mathbf{x} : t > t_{n-1,\alpha}\}$
c.) $\mathcal{X}_k = \{\mathbf{x} : t < -t_{n-1,\alpha}\}$

2.) Kétmintás próbák

$X_1, \ldots, X_n \sim \mathrm{N}(m_1, \sigma_1^2)$
$Y_1, \ldots, Y_m \sim \mathrm{N}(m_2, \sigma_2^2)$
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	b.) kétmintás u-próba		egymintás u-próba a különbségekre
σ_1 és σ_2 ismeretlen	előzetes F-próba: $\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	egymintás t-próba a különbségekre
	c.) kétmintás t-próba	d.) Welch-próba	

a.) F-próba

$m_1, m_2, \sigma_1, \sigma_2$ paraméterek
$H_0 : \sigma_1 = \sigma_2$ és H_1: ami a szövegkörnyezetben értelmes
A próbastatisztika: $F = \begin{cases} \frac{(s_1^*)^2}{(s_2^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{n-1,m-1} & \text{ha } s_1^* > s_2^* \\ \frac{(s_2^*)^2}{(s_1^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{m-1,n-1} & \text{ha } s_2^* > s_1^* \end{cases}$

b.) kétmintás u-próba

m_1, m_2 paraméterek, σ_1, σ_2 ismert
$H_0 : m_1 = m_2$ és H_1: ami a szövegkörnyezetben értelmes

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

c.) kétmintás t-próba

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

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

A próbastatisztika: $t = \sqrt{\frac{nm}{n+m}} \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(n-1)(s_1^*)^2+(m-1)(s_2^*)^2}{n+m-2}}} \overset{H_0 \text{ esetén}}{\sim} \mathrm{t}_{n+m-2}$

d.) Welch-próba

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

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

A próbastatisztika: $t' = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(s_1^*)^2}{n}+\frac{(s_2^*)^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} \mathrm{t}_f$, ahol

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

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

χ^2-próbák

a.) Diszkrét illeszkedésvizsgálat

Feladat: adott egy $\mathbf{X} = (X_1, \ldots, X_n)$ 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: $\mathrm{P}(X_i = x_j) = p_j \quad j = 1, \ldots, r$. 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: $\mathbf{p}$=$(p_1, \ldots, p_r)$

H_1: nem ezek a valószínűségek

A próbastatisztika: $T_n = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ eloszlásban, ha $n \to \infty$

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

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 próbastatisztika: $T_n \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1-s}$ eloszlásban, ha $n \to \infty$.

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 a kisebb mintáknál legalább 3, közepeseknél legalább 5 elem szerepeljen az egyes cellákban. Amennyiben a cellákban túl alacsony a gyakoriságok száma, 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 $k = 1, ..., n$.

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 $k = 1, ..., n$.

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

b.) Diszkrét homogenitávizsgálat

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: $(p_1, \ldots, p_r) = (q_1, \ldots, q_r)$

H_1: nem ezek a valószínűségek

A próbastat.: $T_{n,m} = \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ eloszlásban, ha $n \to \infty$

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

c.) Függetlenségvizsgálat

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

$\mathrm{p}_{i,j}$=P(egy megfigyelés az (i,j) osztályba kerül)

$\mathrm{N}_{i,j}$=ennyi megfigyelés kerül az (i,j) osztályba

A mintavétel eredménye:

		2. szempont					
		1	...	j	...	s	Összesen
	1	N_{11}	...	N_{1j}	...	N_{1s}	$N_{1\bullet}$
	⋮	⋮		⋮		⋮	⋮
1. szempont	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 próbastatisztika: $T_n = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{N_{i,j}^2}{N_{i\bullet}N_{\bullet j}} - 1\right) \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{(r-1)(s-1)}$ eloszlásban, ha $n \to \infty$
A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\underline{x}) > \chi^2_{(r-1)(s-1),1-\alpha}\}$

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

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

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

Feladat (**lineáris regresszió**): Adottak $(x_1,y_1),\ldots,(x_n,y_n)$ pontok, ezekre szeretnénk egyenest illeszteni (neve: *regressziós egyenes*) legkisebb négyzetek módszerével.
A modell: $Y_i = aX_i + b + \varepsilon_i$, ahol $\mathrm{E}\varepsilon_i = 0$ és $\mathrm{D}^2\varepsilon_i = \sigma^2 < \infty$ $(i = 1,\ldots,n)$
Megoldás: $\hat{a} = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$, $\hat{b} = \overline{y} - \hat{a}\overline{x}$
Reziduumok: $\hat{\varepsilon}_i = y_i - \hat{a}x_i - \hat{b}$ (i=1,..., n)
Reziduális négyzetösszeg: RNÖ$=\sum\hat{\varepsilon}_i^2 = \sum(y_i-\overline{y})^2 - \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$
$\hat{\sigma}^2 = \frac{\text{RNÖ}}{n-2}$
Tapasztalati korrelációs együttható: $R = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2 \cdot \sum(y_i-\overline{y})^2}}$. 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.

Érték-, ár- és volumenindexek

Index vagy **indexszám**: közvetlenül nem összesíthető, de gazdaságilag összetartozó adatok átlagos változását mutató összetett viszonyszám.
Tegyük fel, hogy m különböző terméket értékesítünk két különböző időszakban, és az értékesítés árbevételét szeretnénk elemezni.
Jelölések:

- $q_{0,j}$: a j. termékből eladott mennyiség a bázisidőszakban
- $q_{1,j}$: a j. termékből eladott mennyiség a tárgyidőszakban
- $p_{0,j}$ ($p_{1,j}$): az j. termék egységára a bázis- (tárgy)időszakban
- $v_{0,j}$: a j. termék értékesítéséből származó árbevétel (tágabb értelemben *termelési érték*) a bázisidőszakban, számítása: $v_{0,j} = q_{0,j} \cdot p_{0,j}$
- $v_{1,j}$: a j. termék értékesítéséből származó árbevétel a tárgyidőszakban, számítása: $v_{1,j} = q_{1,j} \cdot p_{1,j}$
- Egyedi indexek: (mostantól a j indexeket lehagyjuk)
  - Egyedi volumenindexek: $i_{q,j} = \frac{q_{1,j}}{q_{0,j}} \rightsquigarrow i_q = \frac{q_1}{q_0}$
  - Egyedi árindexek: $i_{p,j} = \frac{p_{1,j}}{p_{0,j}} \rightsquigarrow i_p = \frac{p_1}{p_0}$
  - Egyedi értékindexek: $i_{v,j} = \frac{v_{1,j}}{v_{0,j}} = \frac{q_{1,j}\cdot p_{1,j}}{p_{1,j}\cdot p_{0,j}} \rightsquigarrow i_v = \frac{v_1}{v_0} = \frac{q_1 p_1}{q_0 p_0} = i_p \cdot i_q$
- Összetett indexek:

Index fajtája	Bázisidőszaki súlyozású vagy Laspeyres-féle	Tárgyidőszaki súlyozású vagy Paasche-féle	Fisher-féle
- Árindexek:	$I_p^0 = \frac{\sum q_0 p_1}{\sum q_0 p_0}$	$I_p^1 = \frac{\sum q_1 p_1}{\sum q_1 p_0}$	$I_p^F = \sqrt{I_p^0 \cdot I_p^1}$
- Volumenindexek:	$I_q^0 = \frac{\sum q_1 p_0}{\sum q_0 p_0}$	$I_q^1 = \frac{\sum q_1 p_1}{\sum q_0 p_1}$	$I_q^F = \sqrt{I_q^0 \cdot I_q^1}$
- Értékindex: $I_v = \frac{\sum q_1 p_1}{\sum q_0 p_0}$			

Néhány összefüggés:

- $I_v = I_q^0 \cdot I_p^1 = I_q^1 \cdot I_p^0 = \frac{\sum q_0 p_0 \cdot i_v}{\sum q_0 p_0} = \frac{q_1 p_1}{\sum \frac{\sum q_1 p_1}{i_v}}$
- $I_p^0 = \frac{\sum q_0 p_0 \cdot i_p}{\sum q_0 p_0} = \frac{\sum q_0 p_1}{\sum \frac{q_0 p_1}{i_p}}$
- $I_q^1 = \frac{\sum q_0 p_1 \cdot i_q}{\sum q_0 p_1} = \frac{\sum q_1 p_1}{\sum \frac{q_1 p_1}{i_q}}$

Az indexek képleteiben lévő osztások helyett különbségeket is lehet képezni, ekkor az I és i helyett K-t és k-t írunk. Például $K_p^0 = \sum q_0 p_1 - \sum q_0 p_0$.
