Statisztika 2 előadás

Baran Sándor

A tananyag elkészítését az EFOP-3.4.3-16-2016-00021 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

Irodalom

- Hunyadi László., Vita László: *Statisztika I.* Akadémiai Kiadó, Budapest, 2018. Online verzió (2019): `https://mersz.hu/hunyadi-vita-statisztika-i`
- Hunyadi László, Vita László: *Statisztika II.* Akadémiai Kiadó, Budapest, 2018. Online verzió (2019): `https://mersz.hu/hunyadi-vita-statisztika-ii`
- Keresztély Tibor, Sugár András, Szarvas Beatrix: *Statisztika közgazdászoknak. Példatár és feladatgyűjtemény.* Nemzeti Tankönyvkiadó, Budapest, 2005.

Tartalom

Az előző félév tartalmából ....

Hipotézisvizsgálati alapfogalmak:

- Nullhipotézis, ellenhipotézis
- Próbafüggvény
- Szignifikancia szint, kritikus tartomány
- Első- és másodfajú hiba
- p-érték.

Nevezetes eloszlások:

- Normális eloszlás
- Khi-négyzet eloszlás
- t-eloszlás
- F-eloszlás

Egymintás z-próba

$y_1, y_2, \ldots, y_n$: FAE minta $\mathcal{N}(\mu, \sigma^2)$ eloszlásból, σ **ismert**.

Nullhipotézis: $H_0 : \mu = \mu_0$;

Ellenhipotézis: $H_1 : \mu \neq \mu_0$; (kétoldali ellenhipotézis)

$H_1^b : \mu < \mu_0$; (bal oldali ellenhipotézis)

$H_1^j : \mu > \mu_0$. (jobb oldali ellenhipotézis)

Próbafüggvény: $z := \dfrac{\overline{y} - \mu_0}{\sigma/\sqrt{n}}$.

Ha H_0 teljesül, z eloszlása **standard normális**.

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$H_1 : \mu \neq \mu_0$ esetén $z \leq z_{\alpha/2} = -z_{1-\alpha/2}$ vagy $z \geq z_{1-\alpha/2}$, azaz $|z| \geq z_{1-\alpha/2}$;

$H_1^b : \mu < \mu_0$ esetén $z \leq z_\alpha = -z_{1-\alpha}$;

$H_1^j : \mu > \mu_0$ esetén $z \geq z_{1-\alpha}$.

z_p: a standard normális eloszlás p-kvantilise, $\Phi(z_p) = p$.

Egymintás t-próba

$y_1, y_2, \ldots, y_n$: FAE minta $\mathcal{N}(\mu, \sigma^2)$ eloszlásból, σ **nem ismert**.

Nullhipotézis: $H_0 : \mu = \mu_0$;

Ellenhipotézis: $H_1 : \mu \neq \mu_0$; (kétoldali ellenhipotézis)

$H_1^b : \mu < \mu_0$; (bal oldali ellenhipotézis)

$H_1^j : \mu > \mu_0$. (jobb oldali ellenhipotézis)

Próbafüggvény: $t := \dfrac{\overline{y} - \mu_0}{s/\sqrt{n}}, \qquad s^2 := \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$

Ha H_0 teljesül, t eloszlása $n-1$ szabadsági fokú ***t*-eloszlás** (t_{n-1}).

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$$\begin{aligned}
H_1 : \mu \neq \mu_0 \quad &\text{esetén} \quad |t| \geq t_{1-\alpha/2}(n-1);\\
H_1^b : \mu < \mu_0 \quad &\text{esetén} \quad t \leq t_{\alpha}(n-1) = -t_{1-\alpha}(n-1);\\
H_1^j : \mu > \mu_0 \quad &\text{esetén} \quad t \geq t_{1-\alpha}(n-1).
\end{aligned}$$

$t_p(n-1)$: az $n-1$ szabadsági fokú t-eloszlás p-kvantilise. Táblázatból megadható.

Példa

Egy gabonaraktárban 60 kg-os kiszerelésben búzát csomagolnak. A havi minőségellenőrzés során azt is meg akarták vizsgálni, hogy a raktárból kikerülő zsákokban tényleg 60 kg búza van-e, ezért lemértek tíz darab véletlenül kiválasztott zsákot. Eredményül a következőket kapták:

$$60.2,\ 63.4,\ 58.8,\ 63.6,\ 64.7,\ 62.5,\ 66.0,\ 59.1,\ 65.1,\ 62.0.$$

Hipotéziseit és az adatokra vonatkozó feltételeit pontosan megfogalmazva döntsön 5%-os szinten, a zsákok átlagos töltőtömege tényleg 60 kg-e.

Megoldás.

$$H_0 : \mu = 60 \text{ kg}; \qquad H_1 : \mu \neq 60 \text{ kg} \quad (\text{kétoldali ellenhipotézis}).$$

Feltételezzük, hogy a zsákok töltőtömege normális eloszlású.

$n = 10,\ \alpha = 0.05,\ \overline{y} = 62.54,\ s^2 = 6.2938,\ s = 2.5087.$

A próbafüggvény értéke: $t = \dfrac{\bar{y} - \mu_0}{s}\sqrt{n} = \dfrac{62.54 - 60}{2.5087}\sqrt{10} = 3.2017.$

Ha H_0 igaz, a próbafüggvény eloszlása t_9 eloszlás.

A kritikus tartomány: $|t| \geq t_{0.975}(9) = 2.262.$

A kapott érték beleesik, 5%-os szinten **elvetjük** H_0-t.

Aszimptotikus z-próba

$y_1, y_2, \ldots, y_n$: **nagy** FAE minta tetszőleges véges szórású és μ várható értékű eloszlásból.

Nullhipotézis: $H_0 : \mu = \mu_0$;

Ellenhipotézis:

$$H_1 : \mu \neq \mu_0; \qquad H_1^b : \mu < \mu_0; \qquad H_1^j : \mu > \mu_0.$$

Próbafüggvény: $z := \dfrac{\overline{y} - \mu_0}{s/\sqrt{n}}, \qquad s^2 := \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$

Ha H_0 teljesül, a központi határeloszlás tétel miatt z eloszlása **közel standard normális**. Ha a minta eloszlása szimmetrikus, már $n \approx 30$ elegendő.

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$$\begin{aligned} H_1 : \mu \neq \mu_0 &\quad \text{esetén} \quad |z| \geq z_{1-\alpha/2}; \\ H_1^b : \mu < \mu_0 &\quad \text{esetén} \quad z \leq z_\alpha = -z_{1-\alpha}; \\ H_1^j : \mu > \mu_0 &\quad \text{esetén} \quad z \geq z_{1-\alpha}. \end{aligned}$$

Khi-négyzet próba a szórásra

$y_1, y_2, \ldots, y_n$: FAE minta $\mathcal{N}(\mu, \sigma^2)$ eloszlásból.

Nullhipotézis: $H_0 : \sigma = \sigma_0$; (vagy $H_0 : \sigma^2 = \sigma_0^2$)

Ellenhipotézis:

$$H_1 : \sigma \neq \sigma_0; \qquad H_1^b : \sigma < \sigma_0; \qquad H_1^j : \sigma > \sigma_0.$$

Próbafüggvény: $\mathcal{X}^2 := \dfrac{(n-1)s^2}{\sigma_0^2}, \qquad s^2 := \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$

Ha H_0 teljesül, $\mathcal{X}^2$ eloszlása $n-1$ szabadsági fokú **khi-négyzet eloszlás** ($\mathcal{X}^2_{n-1}$).

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$H_1 : \sigma \neq \sigma_0$ esetén $\mathcal{X}^2 \leq \mathcal{X}^2_{\alpha/2}(n-1)$ vagy $\mathcal{X}^2 \geq \mathcal{X}^2_{1-\alpha/2}(n-1)$;

$H_1^b : \sigma < \sigma_0$ esetén $\mathcal{X}^2 \leq \mathcal{X}^2_{\alpha}(n-1)$;

$H_1^j : \sigma > \sigma_0$ esetén $\mathcal{X}^2 \geq \mathcal{X}^2_{1-\alpha}(n-1)$.

$\mathcal{X}^2_p(n-1)$: a $\mathcal{X}^2_{n-1}$ eloszlás p-kvantilise. Táblázatból megadható.

Példa

A Felsőkutyfalvi Kerékpárüzem kerékrészlegének vezetője arra gyanakszik, hogy az egyik beszállító által készített küllők hosszúsága igencsak változékony. Gyanújának ellenőrzése céljából az adott beszállító termékeiből véletlenszerűen kiválasztott 20 darabot és megmérte azok hosszát. A hossz szórásnégyzetének a minta alapján számolt torzítatlan becslése (azaz a minta korrigált tapasztalati szórásnégyzete) 1.0369 mm^2.
A beszállító állítása szerint a küllők hosszának szórása 0.75 mm.
A küllők hosszát normális eloszlásúnak tételezve fel ellenőrizze, megalapozott-e a részlegvezető gyanúja. Döntsön 5%-os szinten.

Megoldás.

$$H_0 : \sigma = 0.75 \text{ mm}; \qquad H_1 : \sigma > 0.75 \text{ mm} \qquad \text{(jobb oldali ellenhipotézis).}$$

$n = 20$, $\alpha = 0.05$, $s^2 = 1.0369$. A próbafüggvény értéke:

$$\mathcal{X}^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{19 \cdot 1.0369}{0.75^2} = 35.0242.$$

Ha H_0 igaz, a próbafüggvény eloszlása $\mathcal{X}^2_{19}$ eloszlás.

A kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}^2_{0.95}(19) = 30.144$. A kapott érték beleesik, így 5%-os szinten **elvetjük** H_0-t, ami alátámasztja a részlegvezető gyanúját.

Sokasági arányra irányuló nagymintás próba

Legyen adott egy esemény, aminek a valószínűsége P. Például feldobunk egy érmét és fejet dobunk, egy véletlenszerűen kiválasztott hallgató lány, stb.

n elemű minta: n darab független kísérlet az adott eseményre.

$p = k/n$: P torzítatlan és konzisztens becslőfüggvénye, k a vizsgált esemény bekövetkezéseinek száma.

Nullhipotézis: $H_0 : P = P_0$;

Ellenhipotézis:

$$H_1 : P \neq P_0; \qquad H_1^b : P < P_0; \qquad H_1^j : P > P_0.$$

Próbafüggvény: $z := \dfrac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$.

Ha a mintaelemszám nagy, azaz $\min\{nP_0, n(1-P_0)\} \geq 10$, és H_0 teljesül, akkor z eloszlása **közel standard normális**.

Adott α szignifikanciaszinthez tartozó kritikus tartományok ugyanazok, mint a z-próbánál.

Példa

Egy felmérés során a 35 év alatti fiatalok mobilinternet eléréssel való ellátottságát vizsgálták. A véletlenszerűen kiválasztott 1000 megkérdezett 56%-ának volt mobilnet elérése. 1%-os döntési szintet használva vizsgálja meg azt az állítást, miszerint a vizsgált célcsoportnak kevesebb mint 60%-a használ mobilinternetet.

Megoldás.

$$H_0 : P = 0.6; \qquad H_1 : P < 0.6 \quad \text{(bal oldali ellenhipotézis)}.$$

$n = 1000$, $\alpha = 0.01$, $p = 0.56$. A próbafüggvény értéke:

$$z = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0.56 - 0.6}{\sqrt{\frac{0.6 \cdot 0.4}{1000}}} = -2.5820.$$

Mivel $0.6 \cdot 1000 > 10$ és $0.4 \cdot 1000 > 10$, ha H_0 igaz, z eloszlása közel standard normális.

A kritikus tartomány: $z \leq z_{0.01} = -z_{0.99} = -2.326$. A kapott érték beleesik, így **elvetjük** H_0-t.

Alternatív megoldás.

p-érték: $\mathrm{P}(z \leq -2.5820) = 0.0049 < 0.01$. **Elvetjük** H_0-t.

Khi-négyzet próbák általános jellemzői

A próbák nem egy adott paraméterre, hanem a sokaság, vagy egyszerre vizsgált két sokaság **eloszlására** vonatkoznak.

- Illeszkedésvizsgálat: a hipotézis a sokaság **eloszlásának egészére** vonatkozik. Illeszkedik-e a minta egy előre megadott eloszlásra?
- Függetlenségvizsgálat: azt vizsgáljuk, hogy egy sokaság két ismérve független-e egymástól. Szoros kapcsolat az **asszociációval**.
- Homogenitásvizsgálat: két eloszlás egyezőségének vizsgálata.

Kizárólag **nagy minták** esetén használhatóak.

A próbafüggvények aszimptotikusan **khi-négyzet eloszlásúak**.

A kritikus tartomány mindig **jobb oldali**.

Illeszkedésvizsgálat

Probléma. 600 dobás alapján döntsük el egy dobókockáról, hogy az szabályos-e.

$C_1, C_2, \ldots, C_k$: a sokaság valamely ismérv szerinti osztályozása.
Nincs átfedés és lefedik az összes kimenetelt.

$P_1, P_2, \ldots, P_k$: diszkrét valószínűségi eloszlás.
$P_i > 0,\ i = 1, 2, \ldots, k,\ \sum_{i=1}^{k} P_i = 1.$

Nullhipotézis: $H_0 : \mathrm{P}(C_i) = P_i,\ i = 1, 2, \ldots, k;$

Ellenhipotézis: H_1 : valamely i esetén $\mathrm{P}(C_i) \neq P_i$.

n elemű minta: n darab független kísérlet az adott eseményekre.

f_i: a C_i osztály megfigyelt gyakorisága a mintában.

g_i: a C_i osztály megfigyelt relatív gyakorisága a mintában.

nP_i: a C_i osztály várt gyakorisága, ha H_0 igaz.

Tiszta és becsléses illeszkedésvizsgálat

$C_1, C_2, \ldots, C_k$:	osztályok.
$f_1, f_2, \ldots, f_k$:	megfigyelt gyakoriságok.
$g_1, g_2, \ldots, g_k$:	megfigyelt relatív gyakoriságok.
$nP_1, nP_2, \ldots, nP_k$:	várt gyakoriságok.

Próbafüggvény: $$\mathcal{X}^2 := \sum_{i=1}^{k} \frac{(f_i - nP_i)^2}{nP_i} = n\left(\sum_{i=1}^{k} \frac{g_i^2}{P_i} - 1\right).$$

Tiszta illeszkedésvizsgálat: a P_i valószínűségek adottak.

Becsléses illeszkedésvizsgálat: a P_i valószínűségek megadásához b darab paramétert kell becsülnünk a mintából.

Ha a mintaelemszám nagy, azaz $nP_i \geq 5,\ i = 1, 2, \ldots, k$, és H_0 teljesül, akkor $\mathcal{X}^2$ eloszlása közel $\nu = k - b - 1$ szabadsági fokú **khi-négyzet eloszlás**. Tiszta illeszkedésvizsgálat: $b = 0$.

H_0 teljesül: $\mathrm{P}(C_i) = P_i$, azaz $f_i \approx nP_i$, azaz $\mathcal{X}^2$ **kicsi**.

Adott α szignifikanciaszinthez tartozó kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}^2_{1-\alpha}(k - b - 1)$.

Példa. Tiszta illeszkedésvizsgálat, diszkrét eloszlás

Egy újonnan kifejlesztett müzli ötféle magot (A, B, C, D és E) tartalmaz, melyek százalékos megoszlása a terméken lévő tájékoztató szerint 35%, 25%, 20%, 10%, illetve 10%. Egy véletlenül kiválasztott zacskó adatai:

Összetevő	A	B	C	D	E
Szem (darab)	184	145	100	68	63

Döntsön 10%-os szinten, hogy a minta összetétele megfelel-e a csomagoláson feltüntetettnek.

Megoldás.

H_0 : az összetétel megfelel a csomagoláson feltüntetettnek;

H_1 : az összetétel nem felel meg a csomagoláson feltüntetettnek.

$n=560,\ \alpha=0.1,\ k=5,\ P_1=0.35,\ P_2=0.25,\ P_3=0.20,\ P_4=0.10,\ P_5=0.10.$

Megfigyelt gyakoriságok (f_i):	184	145	100	68	63
Várt gyakoriságok (nP_i):	196	140	112	56	56

A próbafüggvény értéke: $\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(f_i - nP_i)^2}{nP_i} = 5.6454.$

Ha H_0 igaz, a próbafüggvény aszimptotikus eloszlása $\mathcal{X}^2_\nu,\ \nu = k-1 = 4.$

A kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}^2_{0.9}(4) = 7.779$. A kapott érték nem esik bele, így 10%-os szinten **elfogadjuk** H_0-t.

Példa. Tiszta illeszkedésvizsgálat, folytonos eloszlás

Egy számítógép segítségével 12 darab, a $[-6, 6]$ intervallumon vett egyenletes eloszlásból származó véletlen számot generáltunk, majd ezt még 99 alkalommal megismételtük. A száz darab mintaátlag eloszlását az alábbi táblázatban összesítettük:

Intervallum	Megfigyelt gyakoriság
$(-\infty, -0.6745)$	26
$[-0.6745, 0)$	21
$[0, 0.6745)$	27
$[0.6745, \infty)$	26

a) Vizsgálja meg 5%-os szinten azt a hipotézist, hogy a mintaátlagok a négy felsorolt intervallum mindegyikébe azonos valószínűséggel esnek.

b) -0.6745, 0 és 0.6745 a standard normális eloszlás alsó kvartilise, mediánja, illetve felső kvartilise. Milyen kapcsolatban áll az előző pontban kapott eredmény a központi határeloszlás tétellel?

Példa. Tiszta illeszkedésvizsgálat, folytonos eloszlás

Megoldás.

a)

$$H_0 : P_\ell = 0.25,\ \ell = 1, 2, 3, 4;$$
$$H_1 : \exists \ell \in \{1, 2, 3, 4\},\ P_\ell \neq 0.25.$$

$n = 100,\ \alpha = 0.05,\ k = 4.$

Megfigyelt gyakoriságok (f_i):	26	21	27	26
Várt gyakoriságok (nP_i):	25	25	25	25

A próbafüggvény értéke: $\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(f_i - nP_i)^2}{nP_i} = 0.8800.$

Ha H_0 igaz, a próbafüggvény aszimptotikus eloszlása $\mathcal{X}^2_\nu,\ \nu = k - 1 = 3.$

A kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}^2_{0.95}(3) = 7.815.$ A kapott érték nem esik bele, így 5%-os szinten **elfogadjuk** H_0-t.

b) Egy 12 elemű, a $[-6, 6]$ intervallumon vett egyenletes eloszlásból származó minta esetén a mintaátlag várható értéke 0, szórása pedig 1. A khi-négyzet próba alapján a mintaátlagok eloszlása a központi határeloszlás tételből adódó standard normális eloszlás.

Példa. Becsléses illeszkedésvizsgálat, diszkrét eloszlás

Egy botanikus hallgató úgy gondolta, hogy egy bizonyos növényfajta a füves réteken véletlenszerűen szétszórt helyeken bukkan fel. Kutatásai során megszámolta a növény egy véletlenszerűen kiválasztott egy négyzetméteres négyzetben (kvadráns) előforduló egyedeinek a számát, majd e kísérletet többször is megismételte. Az így kapott megfigyeléseit az alábbi táblázatban összegezte:

A növények száma	0	1	2	3	4	5	6	legalább 7
Gyakoriság	10	25	43	34	21	15	2	0

a) Az adatokból számítsa ki a vizsgált növény egyedeinek egy négyzetméterre eső átlagos számát.

A szakkönyvek szerint a fenti jellegű megfigyelési eredmények Poisson eloszlással modellezhetők.

b) Döntsön 5%-os szinten, vajon a Poisson modell megfelelően illeszkedik-e a hallgató által kapott adatokra.

Megoldás.

a) Mintalelemszám: $n = 150$. Az egy négyzetméterre eső növények átlagos száma:

$$\overline{y} = \frac{\sum f_i y_i}{n} = \frac{10 \cdot 0 + 25 \cdot 1 + 43 \cdot 2 + 34 \cdot 3 + 21 \cdot 4 + 15 \cdot 5 + 2 \cdot 6}{150} = 2.56.$$

Példa. Becsléses illeszkedésvizsgálat, diszkrét eloszlás

b)

$$H_0 : \text{a minta Poisson eloszlásból származik;}$$
$$H_1 : \text{a minta nem Poisson eloszlásból származik.}$$

Ha Y Poisson eloszlású $\lambda > 0$ paraméterrel:

$$P_\ell(\lambda) := \mathrm{P}(Y = \ell) = \frac{\lambda^\ell}{\ell!}\mathrm{e}^{-\lambda}, \qquad \ell = 0, 1, 2, \ldots .$$

$\mathrm{E}(Y) = \lambda$. A λ becslése (ML, momentumok módszere): $\widehat{\lambda} = \overline{y} = 2.56$.

Becsült valószínűségek: $\widehat{P}_\ell = P_\ell(\widehat{\lambda}) = \dfrac{2.56^\ell}{\ell!}\mathrm{e}^{-2.56}, \qquad \ell = 0, 1, 2, \ldots .$

$n = 150$, $\alpha = 0.05$, $k = 8$, $b = 1$.

Y	0	1	2	3	4	5	6	> 6
f_i	10	25	43	34	21	15	2	0
$\widehat{P}_i$	0.0773	0.1979	0.2533	0.2162	0.1383	0.0708	0.0302	0.0159
$n\widehat{P}_i$	11.60	29.69	38	32.42	20.75	10.62	**4.53**	**2.39**

Az utolsó két kategóriában a várt gyakoriságok kicsik (< 5). Ezt a két kategóriát összevonjuk.

Példa. Becsléses illeszkedésvizsgálat, diszkrét eloszlás

H_0 : a minta Poisson eloszlásból származik;

H_1 : a minta nem Poisson eloszlásból származik.

$n = 150, \ \alpha = 0.05, \ k = 7, \ b = 1.$

Y	0	1	2	3	4	5	> 5
f_i	10	25	43	34	21	15	2
$n\widehat{P}_i$	11.60	29.69	38	32.42	20.75	10.62	6.92

A próbafüggvény értéke: $\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(f_i - n\widehat{P}_i)^2}{n\widehat{P}_i} = 6.9995.$

Ha H_0 igaz, a próbafüggvény aszimptotikus eloszlása $\mathcal{X}^2_\nu, \ \nu = k - b - 1 = 5.$

A kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}^2_{0.95}(5) = 11.07$. A kapott érték nem esik bele, így 5%-os szinten **elfogadjuk** H_0-t.
A minta **Poisson eloszlásból** származik.

Függetlenségvizsgálat

Probléma. Függetlenek-e egymástól a Gazdaságinformatikus BSc hallgatók Makroökonómia és Statisztika 1 jegyei?

X: ismérv. $C_1^X, C_2^X, \ldots, C_r^X$: X szerinti osztályok.

Y: ismérv. $C_1^Y, C_2^Y, \ldots, C_c^Y$: Y szerinti osztályok.

Nullhipotézis: $H_0 : X$ és Y függetlenek.

Ellenhipotézis: $H_1 : X$ és Y nem függetlenek.

Formális megfogalmazás:

$$H_0 : \mathrm{P}(C_i^X \cdot C_j^Y) = \mathrm{P}(C_i^X) \cdot \mathrm{P}(C_j^Y), \quad i = 1, 2, \ldots, r, \; j = 1, 2, \ldots, c.$$

n elemű minta: n darab független kísérlet az adott eseményekre.

n_{ij}: a $C_i^X \cdot C_j^Y$ megfigyelt gyakorisága. Azon mintaelemek száma, melyek mind C_i^X, mind pedig C_j^Y elemei.

Kontingencia tábla

Az X ismérv szerinti osztályok	Az Y ismérv szerinti osztályok						$\sum j$
	C_1^Y	C_2^Y	$\ldots$	C_j^Y	$\ldots$	C_c^Y	
C_1^X	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1c}	$n_{1.}$
C_2^X	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
C_i^X	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
C_r^X	n_{r1}	n_{r2}	$\ldots$	n_{rj}	$\ldots$	n_{rc}	$n_{r.}$
$\sum i$	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.j}$	$\ldots$	$n_{.c}$	n

n_{ij}: a $C_i^X \cdot C_j^Y$ megfigyelt gyakorisága.

$$\sum_{j=1}^{c} n_{ij} = n_{i.}, \quad \sum_{i=1}^{r} n_{ij} = n_{.j}, \qquad \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r} n_{i.} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = n.$$

$n_{i.}$: a C_i^X megfigyelt gyakorisága. $n_{.j}$: a C_j^Y megfigyelt gyakorisága.

Megfigyelt és várt gyakoriságok

Nullhipotézis: $H_0 : \mathrm{P}(C_i^X \cdot C_j^Y) = \mathrm{P}(C_i^X) \cdot \mathrm{P}(C_j^Y), \quad i = 1, 2, \ldots, r,\ j = 1, 2, \ldots, c.$

A $\mathrm{P}\left(C_i^X\right)$ és $\mathrm{P}\left(C_j^Y\right)$ valószínűségek általában **nem** ismertek.

Becslések

C_i^X megfigyelt gyakorisága: $n_{i\cdot}$; $\mathrm{P}\left(C_i^X\right) \approx \widehat{P}_i^X := \frac{n_{i\cdot}}{n}$. $\quad r-1$ darab paramétert becslünk.

C_j^Y megfigyelt gyakorisága: $n_{\cdot j}$; $\mathrm{P}\left(C_j^Y\right) \approx \widehat{P}_j^Y := \frac{n_{\cdot j}}{n}$. $\quad c-1$ darab paramétert becslünk.

$C_i^X \cdot C_j^Y$ megfigyelt gyakorisága: n_{ij}.

Ha H_0 igaz, $C_i^X \cdot C_j^Y$ várt gyakorisága:

$$n_{ij}^* := n\widehat{P}_i^X \cdot \widehat{P}_j^Y = n\frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}.$$

Próbafüggvény

Nullhipotézis: $H_0 : \mathrm{P}(C_i^X \cdot C_j^Y) = \mathrm{P}(C_i^X) \cdot \mathrm{P}(C_j^Y), \quad i = 1, 2, \ldots, r,\ j = 1, 2, \ldots, c.$

$C_i^X \cdot C_j^Y$ megfigyelt gyakorisága: n_{ij}.

$C_i^X \cdot C_j^Y$ várt gyakorisága: $n_{ij}^* := \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$.

Próbafüggvény:

$$\mathcal{X}^2 := \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*} = n \left(\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i\cdot} \cdot n_{\cdot j}} - 1 \right).$$

Ha a mintaelemszám nagy ($n_{ij}^* \geq 5,\ i=1, 2, \ldots, r,\ j=1, 2, \ldots, c$) és H_0 teljesül, akkor $\mathcal{X}^2$ eloszlása közel $\mathcal{X}_\nu^2$, ahol

$$\nu = rc - (r-1) - (c-1) - 1 = (r-1)(c-1).$$

H_0 teljesül: $n_{ij} \approx n_{ij}^*$, azaz $\mathcal{X}^2$ **kicsi**.

Adott α szignifikanciaszinthez tartozó kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}_{1-\alpha}^2\big((r-1)(c-1)\big)$.

Példa

Egy kutatócsoport azt vizsgálta, milyen szoros az összefüggés egy bizonyos betegség leolyásának súlyossága és a betegek életkora között. A vizsgálat során 200 beteg adatait gyűjtötték össze, majd azokat csoportosították a betegség súlyossági foka és a paciens életkora szerint. Eredményül az alábbi táblázatot kapták:

		Életkor			Összesen
		40 alatti	40–60	60 fölötti	
Lefolyás	enyhe	41	34	9	84
	közepes	25	25	12	62
	súlyos	6	33	15	54
Összesen		72	92	36	200

Hipotéziseit pontosan megfogalmazva döntsön 1%-os szinten, van-e összefüggés a betegek életkora és a betegség lefolyásának súlyossága között.

Példa

Megoldás.

$$H_0 : \text{nincs összefüggés;}$$
$$H_1 : \text{van összefüggés.}$$

$r = c = 3, \ \alpha = 0.01, \ n = 200.$:

Megfigyelt gyakoriság:	41	34	9	84
	25	25	12	62
	6	33	15	54
	72	92	36	200

Várt gyakoriság:	30.24	38.64	15.12	84
	22.32	28.52	11.16	62
	19.44	24.84	9.72	54
	72	92	36	200

A próbafüggvény értéke:

$$\mathcal{X}^2 = \frac{(41 - 30.24)^2}{30.24} + \frac{(34 - 38.64)^2}{38.64} + \ldots + \frac{(15 - 9.72)^2}{9.72} = 22.5230.$$

Ha H_0 igaz, a próbafüggvény aszimptotikus eloszlása $\mathcal{X}^2_\nu, \ \nu = (r-1)(c-1) = 4.$

A kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}^2_{0.99}(4) = 13.277.$ A kapott érték beleesik, így 1%-os szinten **elvetjük** H_0-t.

Homogenitásvizsgálat

Probléma. Azonos eloszlásúak-e a Gazdaságinformatikus BSc és a Gazdálkodási és menedzsment BA hallgatók Makroökonómia jegyei?

X és Y: két sokaság.

Nullhipotézis: H_0 : az X és Y eloszlása azonos;

Ellenhipotézis: H_1 : az X és Y eloszlása nem azonos.

$C_1, C_2, \ldots, C_k$: olyan osztályozás, ami mindkét sokaság esetén értelmezve van.

n_{X_i}: a C_i osztály megfigyelt gyakorisága az X sokaságra vett n_X elemű mintában.

n_{Y_i}: a C_i osztály megfigyelt gyakorisága az Y sokaságra vett n_Y elemű mintában.

$\frac{n_{X_i}}{n_X}$: a C_i valószínűségének becslése az X-re vett minta alapján.

$\frac{n_{Y_i}}{n_Y}$: a C_i valószínűségének becslése az Y-ra vett minta alapján.

Próbafüggvény

Nullhipotézis: H_0 : az X és Y sokaság eloszlása azonos.

$C_1, C_2, \ldots, C_k$: osztályozás.

$$P_X(C_i) \approx \frac{n_{X_i}}{n_X}, \qquad P_Y(C_i) \approx \frac{n_{Y_i}}{n_Y}, \quad i = 1, 2, \ldots, k.$$

Próbafüggvény:

$$\mathcal{X}^2 := n_X n_Y \sum_{i=1}^{k} \frac{1}{n_{X_i} + n_{Y_i}} \left(\frac{n_{X_i}}{n_X} - \frac{n_{Y_i}}{n_Y} \right)^2.$$

Ha a H_0 teljesül, akkor $\mathcal{X}^2$ eloszlása közel $\mathcal{X}^2_\nu$, ahol $\nu = k - 1$.

H_0 teljesül: $P_X(C_i) = P_Y(C_i)$, azaz $\frac{n_{X_i}}{n_X} \approx \frac{n_{Y_i}}{n_Y}$, azaz $\mathcal{X}^2$ **kicsi**.

Adott α szignifikanciaszinthez tartozó kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}^2_{1-\alpha}(k-1)$.

Példa

Az alábbi táblázat a magyar lakásállomány megoszlását (ezer darab) tartalmazza 1990 és 2022 január 1.-én.

	Év	Szobák száma			
		1	2	3 és több	Összesen
Lakások száma	1990	645	1681	1527	3853
	2022	458	1696	2365	4519

Hipotéziseit pontosan megfogalmazva döntsön 1%-os szinten, változott-e a lakásállomány megoszlása.

Megoldás.

X: egy véletlenszerűen választott lakás szobaszáma 1990-ben.

Y: egy véletlenszerűen választott lakás szobaszáma 2022-ben.

$$H_0: \text{a két évben a lakásállomány megoszlása azonos;}$$
$$H_1: \text{a két évben a lakásállomány megoszlása nem azonos.}$$

$n_X = 3853,\ n_Y = 4519,\ \alpha = 0.01,\ k = 3.$

Példa

H_0 : a két évben a lakásállomány megoszlása azonos;

H_1 : a két évben a lakásállomány megoszlása nem azonos.

Szobák száma	Lakások száma			Relatív gyakoriság		$\frac{1}{n_{X_i}+n_{Y_i}}\left(\frac{n_{X_i}}{n_X}-\frac{n_{Y_i}}{n_Y}\right)^2$
	1990	2022	Összesen	1990	2022	
	n_{X_i}	n_{Y_i}	$n_{X_i}+n_{Y_i}$	n_{X_i}/n_X	n_{Y_i}/n_Y	
1	645	458	1103	0.1674	0.1013	3.9555×10^{-6}
2	1681	1696	3377	0.4363	0.3753	1.1011×10^{-6}
3 és több	1527	2365	3892	0.3963	0.5233	4.1462×10^{-6}
Összesen	3853	4519	8372	1	1	9.2028×10^{-6}

A próbafüggvény értéke:

$$\mathcal{X}^2 = n_X n_Y \sum_{i=1}^{k} \frac{1}{n_{X_i}+n_{Y_i}}\left(\frac{n_{X_i}}{n_X}-\frac{n_{Y_i}}{n_Y}\right)^2 = 3583\cdot 4519\cdot 9.2028\times 10^{-6} = 160.236$$

Ha H_0 igaz, a próbafüggvény aszimptotikus eloszlása $\mathcal{X}^2_\nu$, $\nu = k-1 = 2$.

A kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}^2_{0.99}(2) = 9.210$. A kapott érték beleesik, így 1%-os szinten **elvetjük** H_0-t.

Kétmintás z-próba

$y_1, y_2, \ldots, y_{n_Y}$: FAE minta $\mathcal{N}(\mu_Y, \sigma_Y^2)$ eloszlásból;

$x_1, x_2, \ldots, x_{n_X}$: FAE minta $\mathcal{N}(\mu_X, \sigma_X^2)$ eloszlásból.

σ_Y, σ_X **ismert**, a minták egymástól függetlenek.

Nullhipotézis: $H_0 : \mu_Y - \mu_X = \delta_0$;

Ellenhipotézis: $H_1 : \mu_Y - \mu_X \neq \delta_0$; (kétoldali ellenhipotézis)

$H_1^b : \mu_Y - \mu_X < \delta_0$; (bal oldali ellenhipotézis)

$H_1^j : \mu_Y - \mu_X > \delta_0$. (jobb oldali ellenhipotézis)

Próbafüggvény: $z := \dfrac{\overline{y} - \overline{x} - \delta_0}{\sqrt{\frac{\sigma_Y^2}{n_Y} + \frac{\sigma_X^2}{n_X}}}$. Ha H_0 teljesül, z eloszlása **standard normális**.

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$H_1 : \mu_Y - \mu_X \neq \delta_0$ esetén $|z| \geq z_{1-\alpha/2}$;

$H_1^b : \mu_Y - \mu_X < \delta_0$ esetén $z \leq z_\alpha = -z_{1-\alpha}$;

$H_1^j : \mu_Y - \mu_X > \delta_0$ esetén $z \geq z_{1-\alpha}$.

Példa

Egy kiterjedt népegészségügyi vizsgálat során megállapították, hogy az egészséges felnőtt populáció esetén a diasztolés (alsó) vérnyomás értékek átlaga 84.8 higanymilliméter, szórása pedig 12.8 higanymilliméter. Az Alsóbezgenyei Atlétikai Klub hat véletlenszerűen kiválasztott versenyzőjénél a klub sportorvosa az alábbi diasztolés értékeket jegyezte fel:

$$79.2,\ 64.6,\ 86.8,\ 73.7,\ 74.9,\ 62.4.$$

Az Alsóbezgenyei Sakk Klub versenyzői szintén meglátogatták a fent említett doktort, aki az ő esetükben is feljegyezte öt véletlenszerűen kiválasztott sportoló diasztolés vérnyomás értékét, melyek az alábbiak:

$$84.6,\ 93.2,\ 104.6,\ 106.7,\ 76.3.$$

Hipotéziseit pontosan megfogalmazva döntsön 1%-os szinten, hogy a sakkozók átlagos diasztolés vérnyomása magasabb-e, mint az atlétáké. A sakkozók és az atléták diasztolés vérnyomásáról feltehetjük, hogy normális eloszlást követ, szórása pedig megegyezik a teljes népesség körében mért értékkel.

Példa

Megoldás.

Y: egy véletlenszerűen választott atléta diasztolés vérnyomása.

X: egy véletlenszerűen választott sakkozó diasztolés vérnyomása.

$$H_0 : \mu_Y - \mu_X = 0; \qquad \text{vagy} \qquad H_0 : \mu_Y = \mu_X.$$
$$H_1 : \mu_Y - \mu_X < 0; \qquad \text{vagy} \qquad H_1 : \mu_Y < \mu_X.$$

$n_X = 5, \; n_Y = 6, \; \alpha = 0.01, \; \delta_0 = 0, \; \sigma_X = \sigma_Y = 12.8, \; \overline{y} = 73.6, \; \overline{x} = 93.08.$

A próbafüggvény értéke:

$$z = \frac{\overline{y} - \overline{x} - \delta_0}{\sqrt{\frac{\sigma_Y^2}{n_Y} + \frac{\sigma_X^2}{n_X}}} = \frac{73.6 - 93.08}{\sqrt{\frac{12.8^2}{6} + \frac{12.8^2}{5}}} = -2.5133.$$

Ha H_0 teljesül, z eloszlása **standard normális**.

A kritikus tartomány: $z \leq z_{0.01} = -z_{0.99} = -2.326$. A kapott érték beleesik, így 1%-os szinten **elvetjük** H_0-t.

Alternatív megoldás.

p-érték: $\mathrm{P}(z \leq -2.5133) = 0.0060 < 0.01$. **Elvetjük** H_0-t.

Kétmintás t-próba

$y_1, y_2, \ldots, y_{n_Y}$: FAE minta $\mathcal{N}(\mu_Y, \sigma_Y^2)$ eloszlásból;
$x_1, x_2, \ldots, x_{n_X}$: FAE minta $\mathcal{N}(\mu_X, \sigma_X^2)$ eloszlásból.

σ_Y, σ_X **nem ismert**, a minták egymástól függetlenek.

Nullhipotézis: $H_0 : \mu_Y - \mu_X = \delta_0$;

Ellenhipotézis: $H_1 : \mu_Y - \mu_X \neq \delta_0$; (kétoldali ellenhipotézis)
$H_1^b : \mu_Y - \mu_X < \delta_0$; (bal oldali ellenhipotézis)
$H_1^j : \mu_Y - \mu_X > \delta_0$. (jobb oldali ellenhipotézis)

σ_Y^2 becslése: $s_Y^2 := \dfrac{1}{n_Y - 1} \sum_{i=1}^{n_Y} (y_i - \overline{y})^2$. σ_X^2 becslése: $s_X^2 := \dfrac{1}{n_X - 1} \sum_{i=1}^{n_X} (x_i - \overline{x})^2$.

A kétmintás t-próba esetei

A) **Megegyező szórások**: $\sigma_X = \sigma_Y = \sigma$.

σ^2 kombinált becslése: $s_C^2 := \dfrac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$.

Próbafüggvény: $t := \dfrac{\overline{y} - \overline{x} - \delta_0}{s_C\sqrt{\frac{1}{n_Y} + \frac{1}{n_X}}}$.

Ha H_0 teljesül, a próbafüggvény eloszlása t_ν, $\nu = n_X + n_Y - 2$.

B) **Nem megegyező szórások**: $\sigma_X \neq \sigma_Y$

Próbafüggvény: $t := \dfrac{\overline{y} - \overline{x} - \delta_0}{\sqrt{\frac{s_Y^2}{n_Y} + \frac{s_X^2}{n_X}}}$.

Ha H_0 teljesül, a próbafüggvény eloszlása t_ν, ahol

$$\nu = \frac{\left(s_X^2/n_X + s_Y^2/n_Y\right)^2}{\frac{1}{n_X-1}\left(s_X^2/n_X\right)^2 + \frac{1}{n_Y-1}\left(s_Y^2/n_Y\right)^2}.$$

Kritikus tartományok

Nullhipotézis: $\quad H_0 : \mu_Y - \mu_X = \delta_0.$

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$$\begin{aligned} H_1 : \mu_Y - \mu_X \neq \delta_0 &\quad \text{esetén} \quad |t| \geq t_{1-\alpha/2}(\nu); \\ H_1^b : \mu_Y - \mu_X < \delta_0 &\quad \text{esetén} \quad t \leq t_\alpha(\nu) = -t_{1-\alpha}(\nu); \\ H_1^j : \mu_Y - \mu_X > \delta_0 &\quad \text{esetén} \quad t \geq t_{1-\alpha}(\nu). \end{aligned}$$

A) $\sigma_X = \sigma_Y$ (előzetesen tudjuk, vagy próbával igazoljuk):

$$\nu = n_X + n_Y - 2.$$

B) $\sigma_X \neq \sigma_Y$:

$$\nu = \frac{\left(s_X^2/n_X + s_Y^2/n_Y\right)^2}{\frac{1}{n_X-1}\left(s_X^2/n_X\right)^2 + \frac{1}{n_Y-1}\left(s_Y^2/n_Y\right)^2}.$$

Nem feltétlenül egész szám! Programcsomagok (pl. SPSS, Matlab, R) kezelik, számolják a kvantiliseket. Táblázatnál egészre kerekítés.

F-próba a szórások egyenlőségére

$y_1, y_2, \ldots, y_{n_Y}$: FAE minta $\mathcal{N}(\mu_Y, \sigma_Y^2)$ eloszlásból;
$x_1, x_2, \ldots, x_{n_X}$: FAE minta $\mathcal{N}(\mu_X, \sigma_X^2)$ eloszlásból. A minták egymástól függetlenek.

Nullhipotézis: $H_0 : \sigma_Y = \sigma_X$; (vagy $H_0 : \sigma_Y^2 = \sigma_X^2$)

Ellenhipotézis: $H_1 : \sigma_Y \neq \sigma_X$; (kétoldali ellenhipotézis)
$H_1^b : \sigma_Y < \sigma_X$; (bal oldali ellenhipotézis)
$H_1^j : \sigma_Y > \sigma_X$. (jobb oldali ellenhipotézis)

Próbafüggvény: $F := \dfrac{s_Y^2}{s_X^2}$.

Ha H_0 teljesül, a próbafüggvény eloszlása $\nu_1 = n_Y - 1$, $\nu_2 = n_X - 1$ szabadsági fokú **F-eloszlás** (F_{n_Y-1, n_X-1}).

$F_p(\nu_1, \nu_2)$: az F_{ν_1,ν_2} eloszlás p-kvantilise.

$$F_{1-p}(\nu_1, \nu_2) = \frac{1}{F_p(\nu_2, \nu_1)}.$$

Kritikus tartományok

Nullhipotézis: $H_0 : \sigma_Y = \sigma_X$.

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$$H_1 : \sigma_Y \neq \sigma_X \quad \text{esetén} \quad F \leq \frac{1}{F_{1-\alpha/2}(n_X - 1, n_Y - 1)} \quad \text{vagy} \quad F \geq F_{1-\alpha/2}(n_Y - 1, n_X - 1);$$

$$H_1^b : \sigma_Y < \sigma_X \quad \text{esetén} \quad F \leq \frac{1}{F_{1-\alpha}(n_X - 1, n_Y - 1)};$$

$$H_1^j : \sigma_Y > \sigma_X \quad \text{esetén} \quad F \geq F_{1-\alpha}(n_Y - 1, n_X - 1).$$

Alternatív próbafüggvény: $F^* := \max\left\{F, \dfrac{1}{F}\right\} = \max\left\{\dfrac{s_Y^2}{s_X^2}, \dfrac{s_X^2}{s_Y^2}\right\} \geq 1.$

Ha H_0 teljesül, a próbafüggvény eloszlása F_{ν_1,ν_2}.

ν_1, ν_2: a számláló, illetve a nevező szabadsági foka.

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$$H_1 : \sigma_Y \neq \sigma_X \quad \text{esetén} \quad F^* \geq F_{1-\alpha/2}(\nu_1, \nu_2).$$

Példa

Kétfajta instant kávé oldódási idejét tesztelték, melyekből minden alkalommal azonos mennyiséget tettek 1 dl forrásban lévő vízbe. A kísérletek eredményeit az alábbi táblázat tartalmazza:

Kávé	Oldódási idő (másodperc)							
Mokka Makka	8.2	5.0	6.8	6.7	5.8	7.3	6.4	7.8
Koffe In	5.1	4.3	3.4	3.7	6.1	4.7		

a) Az oldódási időket normálisnak tételezve fel 5%-os szinten igazoljuk, hogy nincs különbség az oldódási idők szórása között.

b) Az a) pontbeli szinten vizsgáljuk meg azt az állítást, hogy a Mokka Makka kávé lassabban oldódik, mint a Koffe In.

Megoldás.

Y: a Mokka Makka oldódási ideje.

X: a Koffe In oldódási ideje.

Példa

a)

$$H_0 : \sigma_Y = \sigma_X;$$
$$H_1 : \sigma_Y \neq \sigma_X.$$

$n_Y = 8,\ n_X = 6,\ \alpha = 0.05,\ \overline{y} = 6.75,\ \overline{x} = 4.55,\ s_Y^2 = 1.0857,\ s_X^2 = 0.967.$

A próbafüggvény értéke (a nagyobb szórásnégyzet kerül a számlálóba):

$$F^* = \frac{s_Y^2}{s_X^2} = \frac{1.0857}{0.967} = 1.1228.$$

Ha H_0 igaz, a próbafüggvény eloszlása F-eloszlás $\nu_1 = 7$ és $\nu_2 = 5$ szabadsági fokokkal.

A kritikus tartomány: $F^* \geq F_{0.975}(7,5) = 6.853$. A kapott érték nem esik bele, így 5%-os szinten **elfogadjuk** H_0-t.

Példa

b)

$$H_0 : \mu_Y - \mu_X = 0; \qquad \text{vagy} \qquad H_0 : \mu_Y = \mu_X.$$
$$H_1 : \mu_Y - \mu_X > 0; \qquad \text{vagy} \qquad H_1 : \mu_Y > \mu_X.$$

$n_Y = 8,\ n_X = 6,\ \alpha = 0.05,\ \overline{y} = 6.75,\ \overline{x} = 4.55,\ s_Y^2 = 1.0857,\ s_X^2 = 0.967.$

A szórásnégyzet kombinált becslése:

$$s_C^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2} = \frac{5 \cdot 0.967 + 7 \cdot 1.0857}{12} = 1.0362.$$

A próbafüggvény értéke:

$$t = \frac{\overline{y} - \overline{x}}{\sqrt{s_C^2(\frac{1}{n_Y} + \frac{1}{n_X})}} = \frac{6.75 - 4.55}{\sqrt{1.0362(\frac{1}{8} + \frac{1}{6})}} = 4.0018.$$

Ha H_0 igaz, a próbasfüggvény eloszlása t-eloszlás $\nu = 12$ szabadsági fokkal.

A kritikus tartomány: $t \geq t_{0.95}(12) = 1.782$. A kapott érték beleesik, így 5%-os szinten **elvetjük** H_0-t.

Kétmintás aszimptotikus z-próba

$y_1, y_2, \ldots, y_{n_Y}$: **nagy** FAE minta tetszőleges véges szórású, μ_Y várható értékű eloszlásból;
$x_1, x_2, \ldots, x_{n_X}$: **nagy** FAE minta tetszőleges véges szórású, μ_X várható értékű eloszlásból.
A minták egymástól függetlenek.

Nullhipotézis: $H_0 : \mu_Y - \mu_X = \delta_0$;

Ellenhipotézis: $H_1 : \mu_Y - \mu_X \neq \delta_0$; (kétoldali ellenhipotézis)

$H_1^b : \mu_Y - \mu_X < \delta_0$; (bal oldali ellenhipotézis)

$H_1^j : \mu_Y - \mu_X > \delta_0$. (jobb oldali ellenhipotézis)

Próbafüggvény: $z := \dfrac{\overline{y} - \overline{x} - \delta_0}{\sqrt{\frac{s_Y^2}{n_Y} + \frac{s_X^2}{n_X}}}$.

Ha H_0 teljesül, z eloszlása **közel standard normális**.

Adott α szignifikanciaszinthez tartozó kritikus tartományok ugyanazok, mint a kétmintás z-próba esetén.

Páros mintás t-próba

$\binom{y_1}{x_1}, \binom{y_2}{x_2}, \ldots, \binom{y_n}{x_n}$: FAE minta $\binom{Y}{X}$ vektorra, $d_i = y_i - x_i$, $i = 1, 2, \ldots, n$, normális eloszlású, $\mathrm{E}(Y) = \mu_Y$, $\mathrm{E}(X) = \mu_X$. A két ismérv **nem feltétlenül független**!

Nullhipotézis: $H_0 : \mu_Y - \mu_X = \delta_0;$

Ellenhipotézis: $H_1 : \mu_Y - \mu_X \neq \delta_0;$ (kétoldali ellenhipotézis)

$H_1^b : \mu_Y - \mu_X < \delta_0;$ (bal oldali ellenhipotézis)

$H_1^j : \mu_Y - \mu_X > \delta_0.$ (jobb oldali ellenhipotézis)

$d_1, d_2, \ldots, d_n$: új minta $\mathcal{N}(\delta, \sigma_d^2)$ eloszlásból, $\delta = \mu_Y - \mu_X$.

Ekvivalens nullhipotézis: $H_0 : \delta = \delta_0;$

Ekvivalens ellenhipotézis: $H_1 : \delta \neq \delta_0;$

$H_1^b : \delta < \delta_0;$

$H_1^j : \delta > \delta_0.$

Egymintás t-próba a $d_1, d_2, \ldots, d_n$ mintával.

Példa

A Mindent Tudás Egyeteme másodéves gazdaságinformatikus hallgatói két zárthelyi dolgozatot írtak statisztikából. Az alábbi táblázat tíz véletlenszerűen kiválasztott hallgató eredményeit tartalmazza:

Hallgató	A	B	C	D	E	F	G	H	I	J
I. dolgozat (Y)	57	63	67	82	45	65	53	32	51	27
II. dolgozat (X)	53	62	63	80	46	64	44	28	50	29

A dolgozateredmények eltérését normális eloszlásúnak tételezve fel döntsön 5%-os szinten, van-e különbség a két dolgozat nehézségi foka között.

Megoldás. Hipotézisek: $H_0 : \mu_Y - \mu_X = 0; \qquad H_1 : \mu_Y - \mu_X \neq 0.$

Új minta ($d_i = y_i - x_i$): $4,\ 1,\ 4,\ 2,\ -1,\ 1,\ 9,\ 4,\ 1,\ -2.$

Az eredetivel ekvivalens hipotézisek: $H_0 : \delta = 0; \qquad H_1 : \delta \neq 0.$

$n = 10,\ \alpha = 0.05,\ \overline{d} = 2.3,\ s^2 = 9.7889,\ s = 3.1287.$

A próbafüggvény értéke: $t = \dfrac{\overline{d} - \delta_0}{s/\sqrt{n}} = \dfrac{2.3 - 0}{3.1287/\sqrt{10}} = 2.3247.$

Ha H_0 igaz, a próbafüggvény eloszlása t-eloszlás $\nu = 9$ szabadsági fokkal.

A kritikus tartomány: $|t| \geq t_{0.975}(9) = 2.262$. A kapott érték bele esik, így **elvetjük** H_0-t.

Sokasági arányra vonatkozó kétmintás próba

P_Y és P_X: két különböző esemény valószínűsége, vagy két sokasági arány.

p_Y: a P_Y valószínűségű esemény relatív gyakorisága n_Y darab független kísérletből.
$\mathrm{E}(p_Y) = P_Y, \ \mathrm{Var}(p_Y) = P_Y(1 - P_Y)/n_Y.$

p_X: a P_X valószínűségű esemény relatív gyakorisága n_X darab független kísérletből.
$\mathrm{E}(p_X) = P_X, \ \mathrm{Var}(p_X) = P_X(1 - P_X)/n_X.$

A kísérletsorozatok függetlenek, mindkét mintaelemszám **nagy**.

Nullhipotézis: $H_0 : P_Y - P_X = \varepsilon_0;$

Ellenhipotézis: $H_1 : P_Y - P_X \neq \varepsilon_0;$ (kétoldali ellenhipotézis)
$H_1^b : P_Y - P_X < \varepsilon_0;$ (bal oldali ellenhipotézis)
$H_1^j : P_Y - P_X > \varepsilon_0.$ (jobb oldali ellenhipotézis)

A próba esetei

A) $\varepsilon_0 \neq 0$.

Próbafüggvény: $z_{\varepsilon_0} := \dfrac{p_Y - p_X - \varepsilon_0}{\sqrt{\frac{p_Y(1-p_Y)}{n_Y} + \frac{p_X(1-p_X)}{n_X}}}$.

Ha H_0 teljesül és a mintalelemszámok nagyok, z_{ε_0} eloszlása **közel standard normális**.

B) $\varepsilon_0 = 0$. Ha H_0 teljesül, $P_Y = P_X$.

A közös valószínűség kombinált becslése: $\overline{p} := \frac{n_Y p_Y + n_X p_X}{n_Y + n_X}$.

Próbafüggvény: $z_0 := \dfrac{p_Y - p_X}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_Y} + \frac{1}{n_X}\right)}}$.

Ha H_0 teljesül és a mintalelemszámok nagyok, z_0 eloszlása **közel standard normális**.

Adott α szignifikanciaszinthez tartozó kritikus tartományok ugyanazok, mint a z-próba esetén.

Példa

A Tárki 2017. január 13-23. között lezajlott 999 fős reprezentatív mintán alapuló közvéleménykutatása alapján a pártot nem választó szavazásra jogosultak aránya 35%[1]. Ugyanezen csoport aránya a Medián 2017. január 22-27. közötti 1200 elemű reprezentatív mintás felmérése alapjány 31%[2]. Hipotéziseit pontosan megfogalmazva döntsön 5%-os szinten, van-e eltérés a két közvéleménykutató cég eredménye között.

Megoldás. P_Y, P_X: a pártot nem választók aránya a Tárki, illetve a Medián szerint.

$$H_0 : P_Y = P_X; \qquad H_1 : P_Y \neq P_X.$$

$n_X = 999$, $n_Y = 1200$, $\alpha = 0.05$, $p_Y = 0.35$, $p_X = 0.31$.

Kombinált becslés: $\overline{p} := \frac{n_Y p_Y + n_X p_X}{n_Y + n_X} = \frac{999 \cdot 0.35 + 1200 \cdot 0.31}{2199} = 0.3282.$

Próbafüggvény értéke: $z_0 = \frac{p_Y - p_X}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_Y} + \frac{1}{n_X}\right)}} = \frac{0.35 - 0.31}{\sqrt{0.3282 \cdot 0.6718\left(\frac{1}{999} + \frac{1}{1200}\right)}} = 1.9890.$

Ha H_0 igaz, a próbafüggvény eloszlása standard normális.

A kritikus tartomány: $|z_0| \geq z_{0.975} = 1.960$. A kapott érték beleesik, így **elvetjük** H_0-t.

[1] `www.tarki.hu/hu/news/2017/kitekint/20170130_valasztas.html`

[2] `http://www.median.hu/object.57507fe4-3ab4-4a2e-8ca4-4da6d8ee750b.ivy`

Több független minta

Adott M darab sokaság, amit egy adott különbség vagy arányskálán mérhető változó szempontjából vizsgálunk.

$\mu_j,\ \sigma_j^2$: a j-edik sokaság várható értéke, illetve varianciája, $j = 1, 2, \dots, M$.

$y_{1j}, y_{2j}, \dots, y_{n_j j}$: a j-edik sokaságra vett minta. Az egyes FAE minták egymástól is függetlenek.

μ_j és σ_j^2 torzítatlan becslései:

$$\overline{y}_j := \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}, \quad s_j^2 := \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2, \qquad j = 1, 2, \dots, M.$$

$\overline{y}_j$: j-edik részátlag.

A teljes minta elemszáma: $n = \sum_{j=1}^{M} n_j$.

A teljes minta főátlaga:

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{n_j} y_{ij} = \frac{1}{n}\sum_{j=1}^{M} n_j \overline{y}_j.$$

Egy szempontú szórásanalízis

y_{ij}: a j-edik minta i-edik eleme $(j = 1, 2, \ldots, M,\ i = 1, 2, \ldots, n_j)$.

y_{ij} eloszlása: $\mathcal{N}(\mu_j, \sigma_j^2)$.

Nullhipotézis: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_M = \mu$;

Ellenhipotézis: $H_1 : \exists\, j,\ \mu_j \neq \mu$.

Négyzetösszegek közötti összefüggés:

$$\sum_{j=1}^{M}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2 = \sum_{j=1}^{M}\sum_{i=1}^{n_j}(y_{ij} - \overline{y_j})^2 + \sum_{j=1}^{M} n_j(\overline{y_j} - \overline{y})^2.$$

$$SST = SSB + SSK.$$

SST a teljes, SSB a belső, SSK a külső négyzetösszeg.

Átlagos négyzetösszegek:

$$s_k^2 := \frac{SSK}{M-1} \quad \text{(külső)}, \qquad s_b^2 := \frac{SSB}{n-M} \quad \text{(belső)}.$$

Probafüggvény

Nullhipotézis: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_M = \mu.$

$$SSB := \sum_{j=1}^{M} \sum_{i=1}^{n_j} (y_{ij} - \overline{y_j})^2, \qquad SSK := \sum_{j=1}^{M} n_j (\overline{y_j} - \overline{y})^2.$$

A próbafüggvény:

$$F := \frac{SSK/(M-1)}{SSB/(n-M)} = \frac{s_k^2}{s_b^2}.$$

Ha az egyes (normális eloszlású) minták szórásai megegyeznek, azaz $\sigma_1 = \sigma_2 = \cdots = \sigma_M$ (tesztelhető), és H_0 teljesül, akkor a próbafüggvény eloszlása $M-1$ és $n-M$ szabadsági fokú **F-eloszlás**.

Ha H_0 igaz, SSK **kicsi**, SSB **nagy**.

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$$F \geq F_{1-\alpha}(M-1, n-M).$$

Varianciaanalízis-táblázat

Minta: $y_{ij} \sim \mathcal{N}(\mu_j, \sigma_j^2), \quad j = 1, 2, \ldots, M, \ i = 1, 2, \ldots, n_j.$

Feltétel: $\sigma_1 = \sigma_2 = \cdots = \sigma_M.$

Nullhipotézis: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_M = \mu;$

Ellenhipotézis: $H_1 : \exists j, \ \mu_j \neq \mu.$

$$SSB := \sum_{j=1}^{M} \sum_{i=1}^{n_j} (y_{ij} - \overline{y_j})^2, \quad SSK := \sum_{j=1}^{M} n_j (\overline{y_j} - \overline{y})^2, \quad SST = SSK + SSB.$$

Varianciaanalízis-táblázat:

A szóródás oka	Eltérés négyzetösszeg	Szabadsági fok	Átlagos négyzetösszeg	F	p-érték
Külső	SSK	$M-1$	$s_k^2 = \frac{SSK}{M-1}$	s_k^2 / s_b^2	p
Belső	SSB	$n-M$	$s_b^2 = \frac{SSB}{n-M}$	–	–
Teljes	SST	$n-1$	–	–	–

Több variancia egyezőségének vizsgálata, Bartlett-próba

y_{ij}: a j-edik minta i-edik eleme $(j = 1, 2, \ldots, M, \; i = 1, 2, \ldots, n_j)$.

y_{ij} eloszlása: $\mathcal{N}(\mu_j, \sigma_j^2)$.

Nullhipotézis: $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_M^2$;

Ellenhipotézis: H_1 : a varianciák nem egyeznek meg.

$$s_b^2 = \frac{1}{n-M}\sum_{j=1}^{M}\sum_{i=1}^{n_j}(y_{ij} - \overline{y_j})^2, \qquad s_j^2 := \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2.$$

A próbafüggvény:

$$B^2 := \frac{1}{c}\left(\nu \ln s_b^2 - \sum_{j=1}^{M}\nu_j \ln s_j^2\right), \quad c := 1 + \frac{1}{3(M-1)}\left(\sum_{j=1}^{M}\frac{1}{\nu_j} - \frac{1}{\nu}\right),$$

$$\nu := n - M, \quad \nu_j = n_j - 1, \quad j = 1, 2, \ldots M.$$

Ha H_0 igaz, a próbafüggvény aszimptotikus eloszlása $\mathcal{X}^2_{M-1}$.

Adott α szignifikanciaszinthez tartozó kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}^2_{1-\alpha}(M-1)$.

Példa

Egy vizsgálat során azt próbálták kideríteni, hogy a diákok tanulási hatékonysága függ-e a tanulási szokásaiktól. Ennek érdekében a kísérletben részvevők kaptak egy szöveget, amit háromféle módszerrel memorizálhattak: csak olvasással, olvasva és aláhúzva a fontosabb részeket, olvasva és kijegyzetelve a lényeges dolgokat. Egy hét elteltével ugyanezek a diákok írtak egy felmérőt, amiben a kapott szöveg tartalmáról kérdezték őket. A felmérők eredményeinek (pontszámainak) összesítését az alábbi táblázat tartalmazza:

Tanulási módszer	A felmérő eredménye						
Olvas	15	14	18	13	11	14	13	
Olvas és aláhúz	16	20	18	17	14			
Olvas és jegyzetel	18	17	23	16	19	22	20	25

a) Töltse ki a varianciaanalízis-táblázatot.

b) Hipotéziseit pontosan megfogalmazva döntsön 1%-os szinten, igaz-e hogy a tanulás módja befolyásolja annak hatékonyságát.

c) Mondjon legalább két, az adatokra vonatkozó feltételt, ami elengedhetetlen a varianciaanalízis végrehajtásához.

d) Hipotéziseit pontosan megfogalmazva döntsön 10%-os szinten, teljesül-e a szórások egyenlőségére vonatkozó feltétel.

Példa

Megoldás.

a)

Módszer	Minta-elemszám	Összeg $\left(\sum y\right)$	Négyzetösszeg $\left(\sum y^2\right)$	Átlag	Variancia
Olvas	7	98	1400	14	4.6667
Olvas és aláhúz	5	85	1465	17	5
Olvas és jegyzetel	8	160	3268	20	9.7143
Teljes	20	343	6133	17.15	13.1868

$$SST = 6133 - \frac{343^2}{20} = 250.55; \quad SSK = \left(\frac{98^2}{7} + \frac{85^2}{5} + \frac{160^2}{8}\right) - \frac{343^2}{20} = 134.55;$$

$$SSB = SST - SSK = 116.$$

A szóródás oka	Eltérés négyzetösszeg	Szabadsági fok	Átlagos négyzetösszeg	F
Külső	$SSK = 134.55$	$M - 1 = 2$	$s_k^2 = \frac{SSK}{M-1} = 67.275$	$\frac{s_k^2}{s_b^2} = 9.8593$
Belső	$SSB = 116$	$n - M = 17$	$s_b^2 = \frac{SSB}{n-M} = 6.8235$	–
Teljes	$SST = 250.55$	$n - 1 = 19$	–	–

Példa

b) A hipotézisek:

$$H_0 : \text{nincs különbség az átlagos pontszámok között;}$$
$$H_1 : \text{van különbség.}$$

$M = 3,\ n = 20,\ \alpha = 0.01,\ F = 9.8593.$

Ha H_0 igaz, a próbafüggvény eloszlása F-eloszlás $M - 1 = 2$ és $n - M = 17$ szabadsági fokokkal.

A kritikus tartomány: $F \geq F_{0.99}(2, 17) = 6.1121$. A kapott érték beleesik,így **elvetjük** H_0-t.

c) Feltételek a varianciaanalízis végrehajthatóságához:

- A diákokat véletlenszerűen választjuk.
- A felmérők pontszámai mindhárom csoportban normális eloszlásúak.
- Az egyes csoportok szórásai megegyeznek.

Példa

d) A hipotézisek:

$$H_0: \text{minták varianciái megegyeznek;}$$
$$H_1: \text{a varianciák nem egyeznek meg.}$$

$$n = 20,\ M = 3,\ \alpha = 0.1;\ \nu = n - M = 17,\ \nu_1 = n_1 - 1 = 6,\ \nu_2 = n_2 - 1 = 4,$$
$$\nu_3 = n_3 - 1 = 7,\ s_b^2 = 6.8235,\ s_1^2 = 4.6667,\ s_2^2 = 5;\ s_3^2 = 9.7143.$$

$$c = 1 + \frac{1}{3(M-1)}\left(\sum_{j=1}^{M} \frac{1}{\nu_j} - \frac{1}{\nu}\right) = 1 + \frac{1}{3 \cdot 2}\left(\frac{1}{6} + \frac{1}{4} + \frac{1}{7} - \frac{1}{20}\right) = 1.0849.$$

A próbafüggvény értéke:

$$B^2 = \frac{1}{c}\left(\nu \ln s_b^2 - \sum_{j=1}^{M} \nu_j \ln s_j^2\right) = \frac{17 \ln 6.8235 - 6 \ln 4.6667 - 4 \ln 5 - 7 \ln 9.7143}{1.0849} = 0.9686.$$

Ha H_0 igaz, a próbafüggvény aszimptotikus eloszlása $\mathcal{X}_2^2$.

A kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}_{0.90}^2(2) = 4.605$. A kapott érték nem esik bele, így 10%-os szinten **elfogadjuk** H_0-t.

Binomiális próba

Legyen adott egy esemény, aminek a valószínűsége P. Például feldobunk egy érmét és fejet dobunk, egy véletlenszerűen kiválasztott hallgató lány, stb.

n elemű minta: n darab független kísérlet az adott eseményre.

Y: a vizsgált esemény bekövetkezéseinek száma.

Nullhipotézis: $H_0 : P = P_0$;

Ellenhipotézis:

$$H_1 : P \neq P_0; \qquad H_1^b : P < P_0; \qquad H_1^j : P > P_0.$$

Ha H_0 teljesül, akkor Y **binomiális eloszlású** n és P_0 paraméterekkel $(\mathcal{B}(n, P_0))$:

$$\mathrm{P}(Y = k) = \binom{n}{k} P_0^k (1 - P_0)^{n-k}, \quad k = 0, 1, 2, \ldots, n.$$

$$\mathrm{E}(Y) = nP_0, \qquad \mathrm{Var}(Y) = nP_0(1 - P_0).$$

Kritikus tartomány

Nullhipotézis: $H_0 : P = P_0$.

Próbafüggvény: a vizsgált esemény bekövetkezéseinek Y száma n darab független kísérletből.

Adott α szignifikanciaszinthez esetén legyen $c_a(\alpha)$ illetve $c_f(\alpha)$ a legnagyobb, illetve a legkisebb érték, melyre

$$\sum_{k=0}^{c_a(\alpha)} \mathrm{P}(Y = k) = \sum_{k=0}^{c_a(\alpha)} \binom{n}{k} P_0^k (1 - P_0)^{n-k} \leq \alpha,$$

$$\sum_{k=c_f(\alpha)}^{n} \mathrm{P}(Y = k) = \sum_{k=c_f(\alpha)}^{n} \binom{n}{k} P_0^k (1 - P_0)^{n-k} \leq \alpha.$$

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$H_1 : P \neq P_0$ esetén $Y \leq c_a(\alpha/2)$ vagy $Y \geq c_f(\alpha/2)$;

$H_1^b : P < P_0$ esetén $Y \leq c_a(\alpha)$;

$H_1^j : P > P_0$ esetén $Y \geq c_f(\alpha)$.

Folytonossági korrekció

Nullhipotézis: $H_0 : P = P_0$.

Y: a vizsgált esemény bekövetkezéseinek száma n darab független kísérletből.

Nagy minta: $\min\{nP_0, n(1-P_0)\} \geq 10$. Próbafüggvény:

$$z := \frac{Y - nP_0}{\sqrt{nP_0(1-P_0)}} = \frac{Y/n - P_0}{\sqrt{P_0(1-P_0)/n}}.$$

Ha H_0 teljesül, akkor z eloszlása **közel standard normális**.

Közelítés: $\mathrm{P}(Y = k) \approx \Phi\left(\frac{k + 1/2 - nP_0}{\sqrt{nP_0(1-P_0)}}\right) - \Phi\left(\frac{k - 1/2 - nP_0}{\sqrt{nP_0(1-P_0)}}\right).$

Ha a nagy mintára vonatkozó feltétel éppen csak teljesül, folytonossági korrekció szükséges. A próbastatisztika változik.

Bal oldali alternatíva: $$z^a := \frac{Y - nP_0 + 1/2}{\sqrt{nP_0(1-P_0)}}.$$

Jobb oldali alternatíva: $$z^f := \frac{Y - nP_0 - 1/2}{\sqrt{nP_0(1-P_0)}}.$$

Előjel próba

$y_1, y_2, \ldots, y_n$: FAE minta tetszőleges **folytonos** eloszlású Y véletlen változóra.

Y mediánja Me, azaz $\mathrm{P}(Y < Me) = \mathrm{P}(Y > Me) = \frac{1}{2}$.

Nullhipotézis: $H_0 : Me = Me_0$;

Ellenhipotézis:

$$H_1 : Me \neq Me_0; \qquad H_1^b : Me < Me_0; \qquad H_1^j : Me > Me_0.$$

Ekvivalens átfogalmazás a $P := \mathrm{P}(Y > Me_0)$ jelöléssel.

Nullhipotézis: $H_0 : P = \frac{1}{2}$;

Ellenhipotézis:

$$H_1 : P \neq \frac{1}{2}; \qquad H_1^b : P < \frac{1}{2}; \qquad H_1^j : P > \frac{1}{2}.$$

Binomiális próba a $P_0 = \frac{1}{2}$ esetre.

$y_i - Me_0$ értékek között vizsgáljuk például a pozitív előjelűek arányát.

Elméletben nem lehetnek 0 különbségek. A gyakorlatban vannak, elhagyjuk azokat.

Példa

Az alábbi adatok 12 *Turbo tudás* módszerrel felkészített hallgató vizsgapontszámait tartalmazzák (a maximális pontszám 50):

$$36 \quad 26 \quad 30 \quad 34 \quad 42 \quad 24 \quad 30 \quad 45 \quad 32 \quad 19 \quad 35 \quad 38.$$

Közismert, hogy a hagyományos módszerrel tanulók körében a pontok mediánja 30. Az előjel próba segítségével döntsön 10%-os szinten, hogy az új módszerrel megszerzett pontok magasabb medián értékkel bírnak-e.

Megoldás. A hipotézisek: $H_0 : Me = 30; \qquad H_0 : Me > 30.$

Az előjelek (érték - 30 előjele):

$$+ \; - \; 0 \; + \; + \; - \; 0 \; + \; + \; - \; + \; + \, .$$

A próbafüggvény értéke (a + jelek száma): $B = 7$.

Elhagyva a 0 különbségeket, ha H_0 igaz, B eloszlása $\mathcal{B}(10, 0.5)$.

Döntési szint: $\alpha = 0.1$

p-érték: $p = \mathrm{P}(B \geq 7) = 0.172 > 0.1$. **Elfogadjuk** H_0-t.

Páros mintás előjel próba

Probléma. Javítja-e egy gazdaságinformatikus hallgató általános közérzetét, ha a Statisztika 2 vizsga előtti este elfogyaszt egy pohár (1 dl) villányi cabernet savignont?

Válaszok: javítja („+”); rontja („-”); nem változik („0”).

A minta sorrendi skálán értelmezett elempárokból áll.

Azon párok P arányát vizsgáljuk, ahol a pár első tagja valamilyen értelemben megelőzi a másodikat.

Nullhipotézis: $H_0 : P = \frac{1}{2}$;

Ellenhipotézis: $H_1 : P \neq \frac{1}{2}$; $H_1^b : P < \frac{1}{2}$; $H_1^j : P > \frac{1}{2}$.

Speciális eset:

$\begin{pmatrix} y_1 \\ x_1 \end{pmatrix}, \begin{pmatrix} y_2 \\ x_2 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ x_n \end{pmatrix}$: FAE minta a folytonos $\begin{pmatrix} Y \\ X \end{pmatrix}$ vektorra.

Me_{Y-X}: az $Y - X$ különbség mediánja.

Nullhipotézis: $H_0 : Me_{Y-X} = 0$.

Példa

A *Roncsautó* című autós szaklap összehasonlította az azonos árfekvésű Skoda Sztrapacska és Lada Borscs 34 közös jellemzőjét. Az eredményt táblázatos formában is közölték, ahol a ,,+" jelentette, hogy az adott jellemző tekintetében a Skoda Sztrapacska bizonyult jobbnak, a ,,-", hogy a Lada Borscs, a ,,0" pedig, hogy nincs különbség a két autó között. A táblázat összesítése:

21 darab ,,+ ", 9 darab ,,-" és 4 darab ,,0".

5%-os szintet alkalmazva vizsgálja meg a cseh autógyár állítását miszerint a Skoda Sztrapacska a jobb autó.

Megoldás. A hipotézisek:

H_0 : nincs különbség a ,,+" és ,,-" jelek száma között;

H_1 : több a ,,+" mint a ,,-", azaz a Skoda a jobb.

A próbafüggvény értéke (a + jelek száma): $B = 21$.

Elhagyva a 0 különbségeket, ha H_0 igaz, B eloszlása $\mathcal{B}(30, 0.5)$.

Döntési szint: $\alpha = 0.05$

p-érték: $p = \mathrm{P}(B \geq 21) = 0.021 < 0.05$. **Elvetjük** H_0-t.

Sorozatpróba

Probléma. A teremben ülő hölgyek és urak sorrendje véletlenszerű-e.

Adott egy megfigyeléssorozat egy csupán két értéket felvevő változóra.

Cél: annak ellenőrzése, hogy a minta elemei véletlenszerű sorrendben követik-e egymást.

Hipotézisek:

$$H_0: \text{ az egyes értékek mintabeli sorrendje véletlen;}$$
$$H_1: \text{ az egyes értékek mintabeli sorrendje nem véletlen.}$$

A sorozatpróba alkalmazásának feltételei:

- A mintaelemek sorrendje egyértelműen értelmezhető.
- A mintaelemek mindegyike két osztály (például X és Y) valamelyikébe besorolható.

Próbafüggvény

Adott egy minta, melynek elemei két osztályba sorolhatóak.

Nullhipotézis:

$$H_0 : \text{ az egyes osztályok mintabeli sorrendje véletlen.}$$

Átkódolás:

- Az n darab mintaelem mindegyikét besoroljuk az X vagy az Y osztályba.
- A mintaelemek helyére beírjuk a megfelelő osztályokat. Ez egy n_X darab X és n_Y darab Y jelből álló átkódolt sorozatot eredményez $(n_x + n_Y = n)$.
- Az átkódolt mintában megszámoljuk a sorozatok r számát. Próbafüggvény: r.

Sorozat: megszakítás nélkül csak X-ből, vagy csak Y-ból álló jelszakasz.

Példa.

$$YXYXXYXXYXYYYYYYYY.$$

$n = 20,\ n_X = 6,\ n_Y = 12,\ r = 9.$

Kritikus tartomány

Adott egy minta, elemei két osztályba (X és Y) sorolhatóak.

$$H_0: \text{ az egyes osztályok mintabeli sorrendje véletlen.}$$

r: a sorozatok (összefüggő X vagy Y jelsorozat) száma.

Problémás esetek:

- **Túl kevés sorozat**: a mintaelemek csoportosulnak, a sorrend nem véletlen.
- **Túl sok sorozat**: a mintaelemek sorrendje valamilyen szabályszerűséget követ.

Kétoldali ellenhipotézis, alsó és felső kritikus tartomány.

Kis minta: $n_X \leq 20$, $n_Y \leq 20$.

Adott α szignifikanciaszinthez tartozó alsó és felső kritikus értékek táblázatból adhatóak meg.

Nagymintás eset

Adott egy minta, elemei két osztályba (X és Y) sorolhatóak.

$$H_0: \text{ az egyes osztályok mintabeli sorrendje véletlen.}$$

r: a sorozatok (összefüggő X vagy Y jelsorozat) száma.

Ha H_0 teljesül:

$$\mathrm{E}(r) = \frac{2n_X n_Y}{n_X + n_Y} + 1, \qquad \mathrm{Var}(r) = \frac{2n_X n_Y(2n_X n_Y - n_X - n_Y)}{(n_X + n_Y)^2(n_X + n_Y - 1)}.$$

Próbafüggvény:

$$z := \frac{r - \mathrm{E}(r)}{\sqrt{\mathrm{Var}(r)}}.$$

Ha H_0 teljesül és a mintalelemszámok nagyok, z eloszlása **közel standard normális**.

Adott α szignifikanciaszinthez tartozó kétoldali kritikus tartomány ugyanaz, mint a z-próba esetén.

Homogenitásvizsgálat (Wald-Wolfowitz próba)

$y_1, y_2, \ldots, y_{n_Y}$: FAE minta tetszőleges folytonos eloszlásból; eloszlásfüggv. $F(x) = \mathrm{P}(Y < x)$.

$x_1, x_2, \ldots, x_{n_X}$: FAE minta tetszőleges folytonos eloszlásból; eloszlásfüggv. $G(x) = \mathrm{P}(X < x)$.

A minták egymástól függetlenek.

Nullhipotézis: $H_0 : G(x) = F(x)$ (a két minta eloszlása azonos).

A próba végrehajtása:

- Egyesítjük a két mintát és elemeit rangsorba állítjuk.
- Az egyesített minta elemeit osztályozzuk aszerint, melyik mintából származnak. Elkészítjük az átkódolt mintát.

Ha H_0 teljesül, akkor az X és Y sokaságokhoz tartozó mintaelemek **véletlenszerűen** követik egymást.

SPSS: egyező mintaeleemek esetén kiszámolja a sorozatok minimális és maximális számát.

A próbát mindkét értékkel végrehajtja. Előfordulhat, hogy a próbák ellentmondó eredményt adnak, ekkor **nem tudunk dönteni**.

Rangösszegpróba (Mann-Whitney próba)

$y_1, y_2, \ldots, y_{n_Y}$: FAE minta tetszőleges folytonos eloszlású Y véletlen változóra.
Eloszlásfüggvénye: $F(x) = \mathrm{P}(Y < x)$; mediánja: Me_Y.

$x_1, x_2, \ldots, x_{n_X}$: FAE minta tetszőleges folytonos eloszlású X véletlen változóra.
Eloszlásfüggvénye: $G(x) = \mathrm{P}(X < x)$; mediánja: Me_X.

A minták egymástól függetlenek.

Nullhipotézis: $H_0 : G(x) = F(x)$.

Ha $G(x) = F(x)$, akkor $\mathrm{P}(X > Y) = \frac{1}{2}$ és $Me_X = Me_Y$.

A) Nullhipotézis: $H_0 : \mathrm{P}(X > Y) = \frac{1}{2}$;

Ellenhipotézis: $H_1 : \mathrm{P}(X > Y) \neq \frac{1}{2}$; $H_1^b : \mathrm{P}(X > Y) < \frac{1}{2}$; $H_1^j : \mathrm{P}(X > Y) > \frac{1}{2}$.

B) Nullhipotézis: $H_0 : Me_X = Me_Y$;

Ellenhipotézis: $H_1 : Me_X \neq Me_Y$; $H_1^b : Me_X < Me_Y$; $H_1^j : Me_X > Me_Y$.

Próbafüggvény

Adott n_Y elemű minta Y-ra és n_X elemű minta X-re.

Nullhipotézis: $H_0 : \mathrm{P}(X > Y) = \frac{1}{2}$.

- Egyesítjük a két mintát és a kapott $n_Y + n_X$ elemet rangsorba állítjuk.
- Minden egyes mintaelemhez hozzárendeljük a rangját, azaz a sorszámát. Egyenlő mintaelemek esetén az azonos értékek rangjainak átlagát vesszük (kapcsolt rangok).
- Meghatározzuk az X sokaságból való mintaelemek rangjainak R_X összegét. Wilcoxon-féle W rangösszeg.

SPSS: mindkét minta rangösszegét kiszámolja. Azt tekinti X sokaságnak, amelyiknek kisebb az átlagos rangja.

$$\frac{n_X(n_X+1)}{2} \leq R_X \leq n_X \cdot n_Y + \frac{n_X(n_X+1)}{2}.$$

Próbafüggvény (Mann-Whitney U): $U_X := R_X - \dfrac{n_X(n_X+1)}{2}$.

U_X: Y sokaságbeli mintaelem hányszor kisebb, mint X-beli.

Példa

A Csajágóröcsögei Vegyipari Kombinát gépkezelői közül néhányat továbbképzésre küldtek annak érdekében, hogy munkájuk során kevesebb hibát vétsenek. A tanfolyam eredményességét vizsgálandó 6, a tanfolyamot már elvégzett, és 12 még előtte álló gépkezelőnek ugyanazt a feladatot adták és feljegyezték a végrehajtás során vétett hibáik szá mát.

Tanfolyam után	11	9	4	7	6	2						
Tanfolyam előtt	3	17	12	13	21	6	1	15	19	16	14	10

Hipotéziseit pontosan megfogalmazva egy alkalmas nemparaméteres próba segítségével döntsön 5%-os szinten, volt-e haszna a tanfolyamnak.

Megoldás. Y: tanfolyam előtti pontszám; X: tanfolyam utáni pontszám.

A hipotézisek: $H_0 : \mathrm{P}(X > Y) = \frac{1}{2}$; $\quad H_1 : \mathrm{P}(X > Y) < \frac{1}{2}$.

Az egyesített minta rangsora (aláhúzás: a tanfolyamot elvégzők adatai):

$$1,\ \underline{2},\ 3,\ \underline{4},\ \underline{6},\ 6,\ \underline{7},\ \underline{9},\ 10,\ \underline{11},\ 12,\ 13,\ 14,\ 15,\ 16,\ 17,\ 19,\ 21.$$

Az aláhúzott elemek rangjai: 2, 4, 5.5, 7, 8, 10. A rangösszeg: $R_X = 36.5$.

A próbafüggvény ($n_X = 6$, $n_Y = 12$): $U_X = R_X - \frac{n_X(n_X+1)}{2} = 36.5 - 21 = 15.5$.

Kritikus tartomány

Adott n_Y elemű minta Y-ra és n_X elemű minta X-re.

Nullhipotézis: $H_0 : \mathrm{P}(X > Y) = \frac{1}{2}$.

R_X: az X sokaságból való mintaelemek rangösszege.

Próbafüggvény: $0 \leq U_X := R_X - \dfrac{n_X(n_X+1)}{2} \leq n_X \cdot n_Y$.

Ha $\mathrm{P}(X > Y) < \frac{1}{2}$, akkor az X sokaságból vett elemek a rangsor elején állnak, R_X **kicsi**.

Kis minta: $n_Y \leq 20$, $n_X \leq 20$.

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$H_1 : \mathrm{P}(X > Y) \neq \frac{1}{2}$ esetén $U_X \leq c_a(\alpha/2)$ vagy $U_X \geq c_f(\alpha/2)$;

$H_1^b : \mathrm{P}(X > Y) < \frac{1}{2}$ esetén $U_X \leq c_a(\alpha)$;

$H_1^j : \mathrm{P}(X > Y) > \frac{1}{2}$ esetén $U_X \geq c_f(\alpha)$.

$c_a(\alpha)$: alsó kritikus érték, táblázatból. $c_f(\alpha)$: felső kritikus érték, $c_f(\alpha) = n_X \cdot n_Y - c_a(\alpha)$.

Példa

Megoldás. Y: tanfolyam előtti pontszám; X: tanfolyam utáni pontszám.

A hipotézisek:

$$H_0 : \mathrm{P}(X > Y) = \frac{1}{2}; \qquad H_1 : \mathrm{P}(X > Y) < \frac{1}{2}.$$

$n_X = 6, \; n_Y = 12, \; \alpha = 0.05, \; R_X = 36.5.$

A próbafüggvény értéke:

$$U_X = R_X - \frac{n_X(n_X + 1)}{2} = 36.5 - 21 = 15.5.$$

A kritikus tartomány: $U_X \leq U_{0.95}(6, 12) = 17.$

A kapott érték beleesik, így **elvetjük** H_0-t. Volt haszna a tanfolyamnak.

Nagymintás próba

Adott n_Y elemű minta Y-ra és n_X elemű minta X-re.

Nullhipotézis: $H_0 : \mathrm{P}(X > Y) = \frac{1}{2}$.

R_X: az X sokaságból való mintaelemek rangösszege.

$$0 \leq U_X := R_X - \frac{n_X(n_X+1)}{2} \leq n_X \cdot n_Y.$$

Ha H_0 teljesül:

$$\mathrm{E}(U_X) = \frac{n_X n_Y}{2}, \qquad \mathrm{Var}(U_X) = \frac{n_X n_Y(n_X + n_Y + 1)}{12}.$$

Próbafüggvény:

$$z := \frac{U_X - \frac{n_X n_Y}{2}}{\sqrt{\frac{n_X n_Y(n_X+n_Y+1)}{12}}}.$$

Ha H_0 teljesül és a mintalelemszámok nagyok, z eloszlása **közel standard normális**.

Adott α szignifikanciaszinthez tartozó kritikus tartományok ugyanazok, mint a z-próba esetén.

Kruskal-Wallis próba

Adott M darab sokaság, amit egy adott különbség vagy arányskálán mérhető változó szempontjából vizsgálunk.

$y_{1j}, y_{2j}, \ldots, y_{n_j j}$: FAE minta a j-edik sokaságot reprezentáló Y_j folytonos változóra.

Az egyes minták egymástól függetlenek.

$F_j(x) = \mathrm{P}(Y_j < x)$: az Y_j eloszlásfüggvénye, $j = 1, 2, \ldots, M$.

Nullhipotézis: $H_0 : F_1(x) = F_2(x) = \cdots = F_M(x)$;

Ellenhipotézis: H_1 : a minták nem azonos eloszlásúak.

- Egyesítjük az M mintát és a kapott $n = \sum_{j=1}^{M} n_j$ elemet rangsorba állítjuk.
- Minden egyes mintaelemhez hozzárendeljük a rangját, azaz a sorszámát. Egyenlő mintaelemek esetén az azonos értékek rangjainak átlagát vesszük.
- Meghatározzuk az Y_j sokaságból való mintaelemek rangjainak R_j összegét és $\overline{R}_j := R_j / n_j$ átlagos rangját.

Próbafüggvény

y_{ij}: a j-edik minta i-edik eleme $(j = 1, 2, \ldots, M,\ i = 1, 2, \ldots, n_j)$.

$F_j(x)$: a j-edik minta (közös) eloszlásfüggvénye.

$\overline{R}_j$: a j-edik minta átlagos rangja.

Nullhipotézis: $\quad H_0 : F_1(x) = F_2(x) = \cdots = F_M(x)$.

Próbafüggvény: $H := \dfrac{12}{n(n+1)} \displaystyle\sum_{j=1}^{M} n_j \left(\overline{R}_j - \dfrac{n+1}{2} \right)^2$.

Ha H_0 teljesül és a minták nagyok $(n_j \geq 5,\ j = 1, 2, \ldots, M)$, akkor H eloszlása közel $\mathcal{X}^2_{M-1}$.

Az egyesített minta rangösszege: $R := 1 + 2 + \ldots + n = \frac{n(n+1)}{2}$.

Az egyesített minta átlagos rangja: $\overline{R} := R/n = \frac{n+1}{2}$.

H_0 teljesül: $\overline{R}_j \approx \frac{n+1}{2}$, azaz H **kicsi**.

Adott α szignifikanciaszinthez tartozó kritikus tartomány: $H \geq \mathcal{X}^2_{1-\alpha}(M-1)$.

Példa

Egy vizsgálat során azt próbálták kideríteni, hogy a diákok tanulási hatékonysága függ-e a tanulási szokásaiktól. Ennek érdekében a kísérletben részvevők kaptak egy szöveget, amit háromféle módszerrel memorizálhattak: csak olvasással, olvasva és aláhúzva a fontosabb részeket, olvasva és kijegyzetelve a lényeges dolgokat. Egy hét elteltével ugyanezek a diákok írtak egy felmérőt, amiben a kapott szöveg tartalmáról kérdezték őket. A felmérők eredményeinek (pontszámainak) összesítését az alábbi táblázat tartalmazza:

Tanulási módszer	A felmérő eredménye							
Olvas	15	14	18	13	11	14	13	
Olvas és aláhúz	16	20	18	17	14			
Olvas és jegyzetel	18	17	23	16	19	22	20	25

Hipotéziseit pontosan megfogalmazva, alkalmas nemparaméteres próba segítségével döntsön 1%-os szinten, igaz-e hogy a tanulás módja befolyásolja annak hatékonyságát.

Megoldás: Hipotézisek:

$$H_0: \text{ a három minta azonos eloszlásból származik;}$$
$$H_1: \text{ a három minta nem azonos eloszlásból származik.}$$

Megoldás

Az egyesített minta elemei és rangjaik:

Mintaelem	11	13	13	14	14	14	15	16	16	17
Rang	1	2.5	2.5	5	5	5	7	8.5	8.5	10.5

Mintaelem	17	18	18	18	19	20	20	22	23	25
Rang	10.5	13	13	13	15	16.5	16.5	18	19	20

$n_1 = 7, \quad R_1 = 36, \quad \overline{R_1} = 5.1429; \quad n_2 = 5, \quad R_2 = 53.5, \quad \overline{R_2} = 10.7;$

$n_3 = 8, \quad R_3 = 120.5, \quad \overline{R_3} = 15.0625; \quad \boldsymbol{n = 20}, \quad \boldsymbol{R = 210}, \quad \boldsymbol{\overline{R} = 10.5}.$

$M = 3$, $\alpha = 0.01$. A próbafüggvény értéke:

$$H := \frac{12}{n(n+1)} \sum_{j=1}^{M} n_j \left(\overline{R}_j - \frac{n+1}{2} \right)^2 = \frac{12}{20 \cdot 21} \Big(7 \cdot (5.1429 - 10.5)^2$$

$$+ 5 \cdot (10.7 - 10.5)^2 + 8 \cdot (15.0625 - 10.5)^2 \Big) = 10.5035.$$

Ha H_0 igaz, a próbafüggvény aszimptotikus eloszlása $\mathcal{X}_2^2$.

A kritikus tartomány: $\mathcal{X}^2 \geq \mathcal{X}_{0.99}^2(2) = 9.210$. A kapott érték beleesik, így 1%-os szinten **elvetjük** H_0-t.

Empirikus eloszlásfüggvény

$y_1, y_2, \dots, y_n$: FAE minta tetszőleges folytonos eloszlású Y véletlen változóra; eloszlásfüggvénye: $F(x) = \mathrm{P}(Y < x)$.

$y_1^*, y_2^*, \dots, y_n^*$: rangsor.

A minta empirikus eloszlásfügvénye:

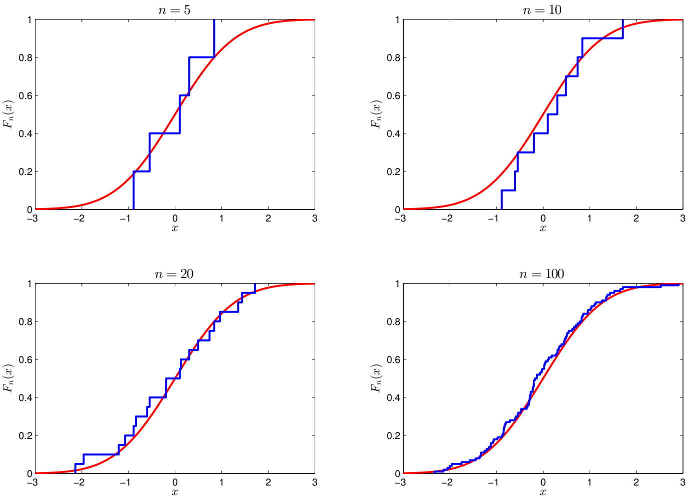
$$F_n(x) := \begin{cases} 0, & \text{ha } x \le y_1^*; \\ \frac{k}{n}, & \text{ha } y_k^* < x \le y_{k+1}^*, \quad k = 1, 2, \dots, n-1; \\ 1, & \text{ha } x \ge y_n^*. \end{cases}$$

Tétel. (A matematikai statisztika alaptétele) *Ha $y_1, y_2, \dots, y_n$ FAE minta egy $F(x)$ eloszlásfüggvényű eloszlásból és $F_n(x)$ a minta empirikus eloszlásfüggvénye, akkor*

$$\mathrm{P}\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} \left| F_n(x) - F(x) \right| \to 0 \right) = 1.$$

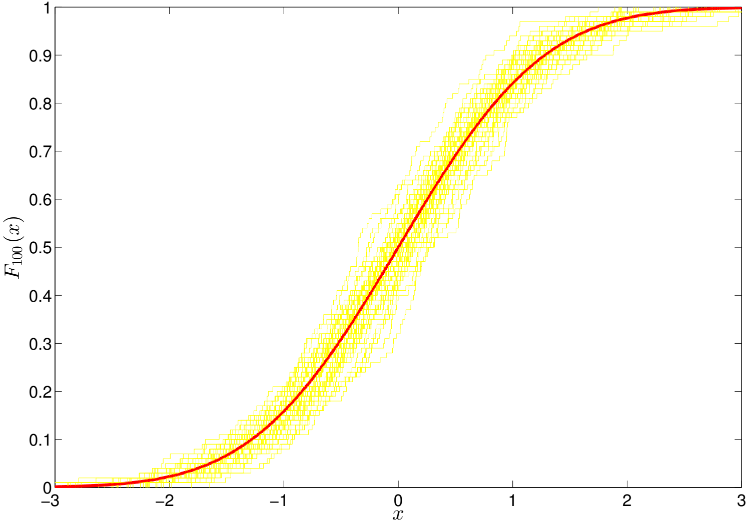
Ha a mintaelemszám nagy, $F_n(x)$ **jól közelíti** az $F(x)$ eloszlásfüggvényt.

Konvergencia



A standard normális eloszlás eloszlásfüggvénye és különböző mintaelemszámokhoz tartozó empirikus eloszlásfüggvények

Illeszkedés



A standard normális eloszlás eloszlásfüggvénye és 50 darab 100 elemű minta empirikus eloszlásfüggvénye

Egymintás Kolmogorov-Szmirnov próba

$y_1, y_2, \ldots, y_n$: FAE minta tetszőleges folytonos eloszlású Y véletlen változóra; eloszlásfüggvénye: $F(x) = \mathrm{P}(Y < x)$.

$F_0(x)$: tetszőleges folytonos eloszlásfüggvény.

Nullhipotézis: $H_0 : F(x) = F_0(x)$;

Ellenhipotézis: $H_1 : F(x) \neq F_0(x)$.

$F_n(x)$: a minta empirikus eloszlásfüggvénye.

Próbafüggvény:

$$D_n := \sup_{-\infty < x < \infty} \left| F_n(x) - F_0(x) \right|.$$

D_n kiszámításához csak a mintaelemek helyén kell nézni az eltéréseket.

Ha H_0 igaz, D_n **kicsi**, azaz a kritikus tartomány jobboldali.

A kritikus tartomány meghatározásához ismernünk kell a próbafüggvény (aszimptotikus) eloszlását.

Kritikus tartomány

$y_1, y_2, \ldots, y_n$: FAE minta az Y véletlen változóra, eloszlásfüggvénye: $F(x) = \mathrm{P}(Y < x)$.

Nullhipotézis: $H_0 : F(x) = F_0(x)$.

Próbafüggvény: $D_n := \sup\limits_{-\infty<x<\infty} \big|F_n(x) - F_0(x)\big|$.

Tétel. *Ha H_0 teljesül, akkor $\sqrt{n}D_n$ eloszlása megközelítőleg a Kolmogorov-eloszlás, aminek eloszlásfüggvénye*

$$K(z) := \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k \mathrm{e}^{-2k^2z^2}, & z > 0; \\ 0, & z \leq 0, \end{cases}$$

azaz

$$\lim_{n\to\infty} \mathrm{P}(\sqrt{n}D_n < z) = K(z).$$

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$$\sqrt{n}D_n \geq K_{1-\alpha}.$$

$K_{1-\alpha}$ értéke táblázatból, programcsomagok számolják.

Két független mintás Kolmogorov-Szmirnov próba

$y_1, y_2, \ldots, y_{n_Y}$: FAE minta tetszőleges folytonos eloszlású Y véletlen változóra; eloszlásfüggvénye: $F(x) = \mathrm{P}(Y < x)$.

$x_1, x_2, \ldots, x_{n_X}$: FAE minta tetszőleges folytonos eloszlású X véletlen változóra; eloszlásfüggvénye: $G(x) = \mathrm{P}(X < x)$.

A minták egymástól függetlenek.

Nullhipotézis: $H_0 : F(x) = G(x)$;

Ellenhipotézis: $H_1 : F(x) \neq G(x)$.

$F_{n_Y}(x)$: az Y minta empirikus eloszlásfüggvénye.

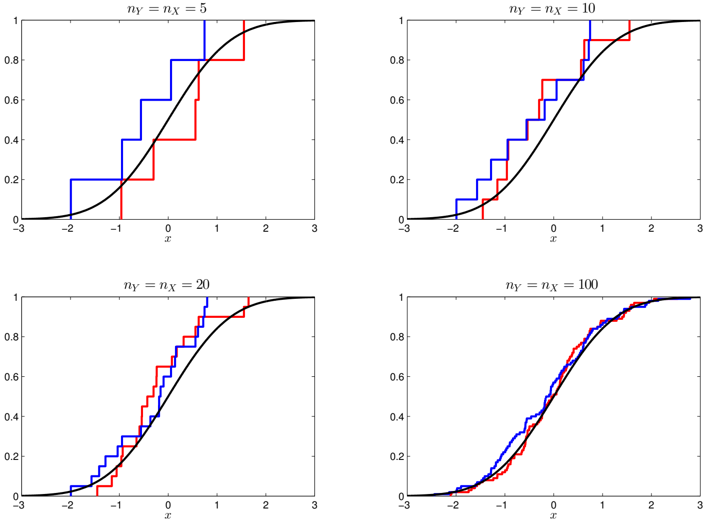
$G_{n_X}(x)$: az X minta empirikus eloszlásfüggvénye.

Próbafüggvény: $$D_{n_Y,n_X} := \sup_{-\infty<x<\infty} \left| F_{n_Y}(x) - G_{n_X}(x) \right|.$$

Nagy minták esetén $F_{n_Y}(x) \approx F(x)$ és $G_{n_X}(x) \approx G(x)$.

Ha H_0 igaz, $F_{n_Y}(x) \approx G_{n_X}(x)$, azaz D_{n_Y,n_X} **kicsi**.

Empirikus eloszlásfüggvények eltérése



Két standard normális eloszlású minta empirikus eloszlásfüggvénye különböző mintaelemszámok esetén.

Kritikus tartomány

$y_1, y_2, \ldots, y_{n_Y}$: FAE minta az Y változóra, eloszlásfüggvénye: $F(x) = \mathrm{P}(Y < x)$.
$x_1, x_2, \ldots, x_{n_X}$: FAE minta az X változóra, eloszlásfüggvénye: $G(x) = \mathrm{P}(X < x)$.

A minták egymástól függetlenek.

Nullhipotézis: $H_0 : F(x) = G(x)$.

Próbafüggvény: $D_{n_Y,n_X} := \sup\limits_{-\infty<x<\infty} \left|F_{n_Y}(x) - G_{n_X}(x)\right|$.

D_{n_Y,n_X} kiszámításához csak a mintaelemek helyén kell nézni az eltéréseket.

Ha H_0 teljesül, akkor $\sqrt{\frac{n_Y n_X}{n_Y+n_X}} D_{n_Y,n_X}$ eloszlása megközelítőleg a Kolmogorov-eloszlás, azaz

$$\lim_{n_Y,n_X\to\infty} \mathrm{P}\left(\sqrt{\frac{n_Y n_X}{n_Y + n_X}} D_{n_Y,n_X} < z\right) = K(z).$$

Adott α szignifikanciaszinthez tartozó kritikus tartomány:

$$\sqrt{\frac{n_Y n_X}{n_Y + n_X}} D_{n_Y,n_X} \geq K_{1-\alpha}.$$

Grafikus illeszkedésvizsgálat: PP plot, QQ plot

$y_1, y_2, \ldots, y_n$: FAE minta tetszőleges folytonos eloszlású Y véletlen változóra.
Eloszlásfüggvénye: $F(x) = \mathrm{P}(Y < x)$. Rangsor: $y_1^*, y_2^*, \ldots, y_n^*$.

$F_0(x)$: tetszőleges folytonos eloszlásfüggvény.

Nullhipotézis: $H_0 : F(x) = F_0(x)$

PP plot (Probability-probability plot) Ábrázoljuk az

$$\big(F_n(y_j^*), F_0(y_j^*)\big) \in [0,1] \times [0,1], \quad j = 1, 2, \ldots, n,$$

pontokat. $F_n(x)$: a minta empirikus eloszlásfüggvénye.

H_0 igaz: $F_n(y_j^*) \approx F_0(y_j^*)$, a pontok az **átló közelében vannak**.

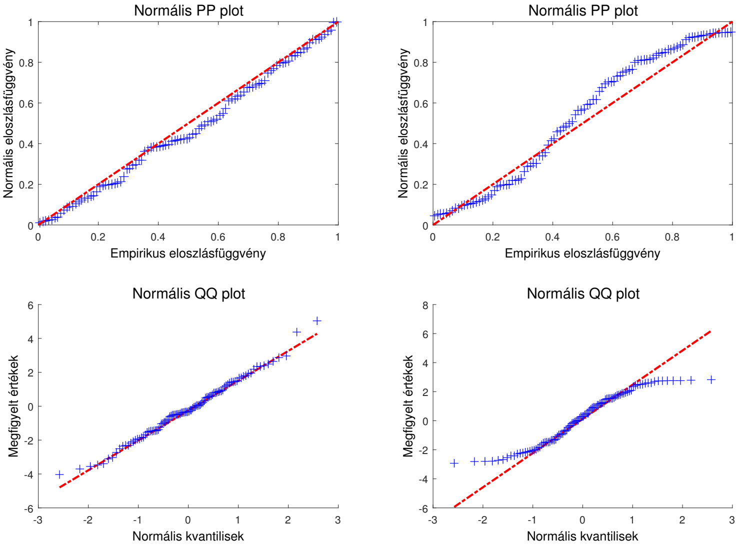
QQ plot (Quantile-quantile plot) Ábrázoljuk az

$$\Big(F_0^{-1}\big((j-0.5)/n\big), y_j^*\Big), \quad j = 1, 2, \ldots, n,$$

pontokat, azaz az elméleti és empirikus kvantiliseket.

H_0 igaz: a pontok egy **egyenes közelében vannak**.

Empirikus eloszlásfüggvények eltérése



Normalitás tesztelése 100 elemű $\mathcal{N}(0,3)$ (*bal oldal*) és $\mathcal{U}(-3,3)$ (*jobb oldal*) minta esetén.

Grafikus homogenitásvizsgálat: kétmintás QQ plot

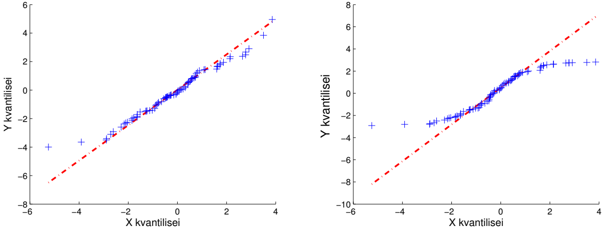
$y_1, y_2, \ldots, y_{n_Y}$: FAE minta az Y változóra, eloszlásfüggvénye: $F(x) = \mathrm{P}(Y < x)$.

$x_1, x_2, \ldots, x_{n_X}$: FAE minta az X változóra, eloszlásfüggvénye: $G(x) = \mathrm{P}(X < x)$.

A minták folytonos eloszlásúak és egymástól függetlenek.

Nullhipotézis: $H_0 : F(x) = G(x)$.

Ábrázoljuk a két minta egymásnak megfelelő $\min\{n_Y, n_X\}$ kvantilisét. Ha H_0 igaz, a pontok egy **egyenes közelében vannak**.



Egy 80 elemű $\mathcal{N}(0,3)$ és egy-egy 100 elemű $\mathcal{N}(0,3)$ (*bal oldal*) és $\mathcal{U}(-3,3)$ (*jobb oldal*) minta homogenitásának vizsgálata.

Regressziós modellek

Regresszió: a változók közötti kapcsolat elemzésének elterjedt eszköze.

Alapesetben azt vizsgáljuk, hogy egy kitüntetett változó, az eredményváltozó (vagy függő változó) hogyan függ egy vagy több magyarázó (vagy független) változótól.

Az eredményváltozó (Y) és a magyarázó változók ($X_1, X_2, \dots, X_k$) között **sztochasztikus kapcsolatot** tételezünk fel.

Általános modell:

$$Y = f(X_1, X_2, \dots, X_j, \dots, X_k, \varepsilon).$$

ε: valószínűségi változó (maradékváltozó).

Lineáris regressziós modell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j + \cdots + \beta_k X_k + \varepsilon.$$

Kétváltozós lineáris regressziós modell:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Megfigyelt adatok

Lineáris modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$

n darab megfigyelés a függő és a független változókra.

A mintára felírt modell:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad i = 1, 2, \ldots, n.$$

Mátrix-vektor elrendezés:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

A megfigyelt adatokra felírt regresszió:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Paraméterbecslés a kétváltozós modellben

Kétváltozós regressziós modell a megfigyelt adatokra:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

β_1: meredekség (slope); β_0: tengelymetszet (intercept).

Legkisebb négyzetes becslés: keressük a β_0 és β_1 paraméterek azon $\widehat{\beta}_0$ és $\widehat{\beta}_1$ becsléseit, melyekre a

$$Q(\beta_0, \beta_1) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

minimális.

Normálegyenletek:

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \qquad \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Becsült paraméterek

Kétváltozós regressziós modell a megfigyelt adatokra:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

A normálegyenletek megoldása:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n d_{x_i} d_{y_i}}{\sum_{i=1}^n d_{x_i}^2} = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^n x_i^2 - n(\overline{x})^2}, \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x},$$

$$d_{x_i} := x_i - \overline{x} = d_x, \qquad d_{y_i} := y_i - \overline{y} = d_y, \quad i = 1, 2, \ldots, n.$$

$\widehat{\beta}_0, \widehat{\beta}_1$: az adott megfigyelésekből számított becsült regressziós együtthatók.

$\widehat{\beta}_1$: a magyarázó változó egységnyi növekedése átlagosan hány egységnyi növekedéssel/csökkenéssel jár együtt az eredményváltozóban.

$\widehat{\beta}_0$: a modell szerint mekkora lesz az eredményváltozó értéke, ha a magyarázó változó 0 értéket vesz fel.

A regressziós egyenes egyenlete: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x$.

Példa

Néhány alsó középkategóriás személygépkocsi vegyes fogyasztása és CO_2 kibocsátása.

	Kia cee'd 1.4 CVVT	Citroën C4 1.4 Vti	Ford Focus 1.6 Ti-VCT	Honda Civic 1.4i
Teljesítmény (LE)	100	95	105	100
Fogyasztás (l/100km)	6.0	6.1	5.9	5.4
CO_2 (g/km)	139	140	136	128
	Mazda 3 1.6 MZR	**Opel Astra 1.4 Ecotec**	**Renault Mégane 1.6**	**Volkwagen Golf 1.2 TSI**
Teljesítmény (LE)	105	100	100	105
Fogyasztás (l/100km)	6.5	5.5	6.7	5.7
CO_2 (g/km)	149	129	155	134

Forrás: Az Autó, 2012/9.

Írja fel a CO_2 kibocsátást és a fogyasztást összekötő regressziós egyenes egyenletét.

Megoldás. X: fogyasztás; Y: CO_2 kibocsátás.

$$\bar{x} = 5.975, \quad d_x: \; 0.025,\; 0.125,\; -0.075,\; -0.575,\; 0.525,\; -0.475,\; 0.725,\; -0.275;$$

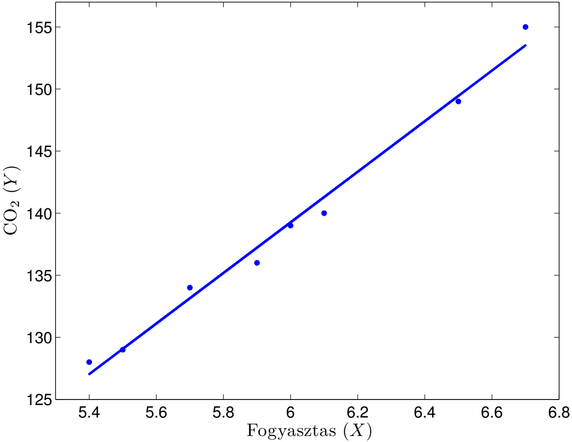
$$\bar{y} = 138.75, \quad d_y: \; 0.25,\; 1.25,\; -2.75,\; -10.75,\; 10.25,\; -9.75,\; 16.25,\; -4.75.$$

Példa

Paraméterbecslések:

$$\widehat{\beta}_1 = \frac{\sum d_x d_y}{\sum d_x^2} = \frac{29.65}{1.455} = 20.3780, \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = 138.75 - 20.3780 \cdot 5.975 = 16.9914.$$

A regressziós egyenes egyenlete: $\widehat{y} = 16.9914 + 20.3780 \cdot x.$



Elaszticitás

Kétváltozós regressziós modell: $Y = \beta_0 + \beta_1 X + \varepsilon$.

Rugalmasság (elaszticitás): olyan mutató, ami megmondja, a magyarázó változó 1%-os növekedése az eredményváltozó hány %-os növekedésével/csökkenésével jár együtt.

Ívrugalmasság: $\mathrm{EL}(Y, X) := \dfrac{\Delta Y}{Y} : \dfrac{\Delta X}{X} = \dfrac{\Delta Y}{\Delta X} \cdot \dfrac{X}{Y}$.

ΔY, ΔX: az Y és az X változók elmozdulásai.

Pontrugalmasság: $\mathrm{El}(Y, X) := \dfrac{dY}{dX} \cdot \dfrac{X}{Y}$.

Az ívrugalmassság határértéke végtelen kicsi elmozdulások esetén.

$\widehat{\beta}_0$, $\widehat{\beta}_1$: becsült regressziós együtthatók.

Kétváltozós lineáris modell rugalmassága:

$$\mathrm{El}(\widehat{y}, x) = \frac{\widehat{\beta}_1 x}{\widehat{y}} = \frac{\widehat{\beta}_1 x}{\widehat{\beta}_0 + \widehat{\beta}_1 x}.$$

Mindig egy adott x pontban van értelmezve.

Becsült regressziós függvényértékek, maradékok

Kétváltozós regressziós modell a megfigyelt adatokra:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

A regressziós egyenes egyenlete: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_0 \cdot x.$

Az x_i pontban a modell által előrejelzett érték: $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_0 \cdot x_i.$

Az x_i pontban a reziduum (maradék): $e_i := y_i - \widehat{y}_i.$

A kis e_i maradékok jó illeszkedést jeleznek, de $\sum_{i=1}^{n} e_i = 0.$

Reziduális négyzetösszeg: $SSE := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$

Reziduális szórás: $s_e^* := \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2}.$

Illeszkedési mutatók. Minél kisebbek, annál jobb a lineáris regressziós modell illeszkedése.

Programcsomagok a reziduális szórás **korrigált** verzióját számolják.

Korrelációs mérőszámok

Kétváltozós regressziós modell a megfigyelt adatokra:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

A minta (becsült) kovarianciája és a varianciák:

$$\mathrm{Cov}(x, y) := \frac{\sum d_x d_y}{n}, \qquad \mathrm{Var}(x) = s_x^{*2} := \frac{\sum d_x^2}{n}, \qquad \mathrm{Var}(y) = s_y^{*2} := \frac{\sum d_y^2}{n}.$$

Programcsomagok a **korrigált** varianciákat számolják.

A minta korrelációs együtthatója:

$$r := \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\,\mathrm{Var}(y)}} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}, \qquad -1 \le r \le 1.$$

$r = \pm 1$: szoros, közel lineáris függvényszerű kapcsolat.

$r = 0$: korrelálatlanság, a lineáris kapcsolat hiánya.

Determinációs együttható

Kétváltozós regressziós modell a megfigyelt adatokra:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

Előrejelzett értékek: $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_i$

Négyzetes eltérések:

$$SST := \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 =: SSR + SSE.$$

Determinációs együttható:

$$R^2 := \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \qquad 0 \leq R^2 \leq 1.$$

$R^2 \times 100\%$ azt mutatja meg, hogy az y adatokban meglévő variancia (bizonytalanság) hány százaléka szüntethető meg a regressziós modellel.

Megadja a modell **magyarázó erejét**.

A mutatók és a paraméterbecslések kapcsolata

Korreláció és becsült meredekség:

$$r=\frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}=\frac{\sum d_x d_y}{\sum d_x^2}\sqrt{\frac{\sum d_x^2}{\sum d_y^2}}=\widehat{\beta}_1\cdot\frac{s_x^*}{s_y^*}.$$

Korreláció és determinációs együttható:

$$R^2=\frac{SSR}{SST}=\frac{\sum(\widehat{y}_i-\overline{y})^2}{\sum d_y^2}=\frac{\sum\left(\widehat{\beta}_0+\widehat{\beta}_1 x_i-\widehat{\beta}_0-\widehat{\beta}_1\overline{x}\right)^2}{\sum d_y^2}=\widehat{\beta}_1^2\frac{\sum d_x^2}{\sum d_y^2}=r^2.$$

Csak a kétváltozós esetben igaz! Nem következik, hogy $R=r$!

További összefüggések:

$$SSR=\widehat{\beta}_1^2\sum d_x^2,\qquad SST=\sum d_y^2,$$
$$SSR=R^2\cdot SST,\qquad SSE=\left(1-R^2\right)\cdot SST.$$

Példa

X: fogyasztás; Y: CO_2 kibocsátás.

A regressziós egyenes egyenlete: $\widehat{y} = 16.9914 + 20.3780 \cdot x$.

$$\overline{x} = 5.975, \quad \overline{y} = 138.75; \quad \sum d_x d_y = 29.65, \quad \sum d_x^2 = 1.455, \quad \sum d_y^2 = 611.5.$$

Elaszticitás az $\overline{x} = 5.975$ pontban:

$$\mathrm{El}(\overline{\widehat{y}}, \overline{x}) = \frac{\widehat{\beta}_1 \overline{x}}{\widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}} = \frac{20.3780 \cdot 5.975}{16.9914 + 20.3780 \cdot 5.975} = 0.8775.$$

Becsült korreláció:

$$r := \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}} = \frac{29.65}{\sqrt{1.455 \cdot 611.5}} = 0.9940.$$

Nagyon erős lineáris kapcsolat a két változó között.

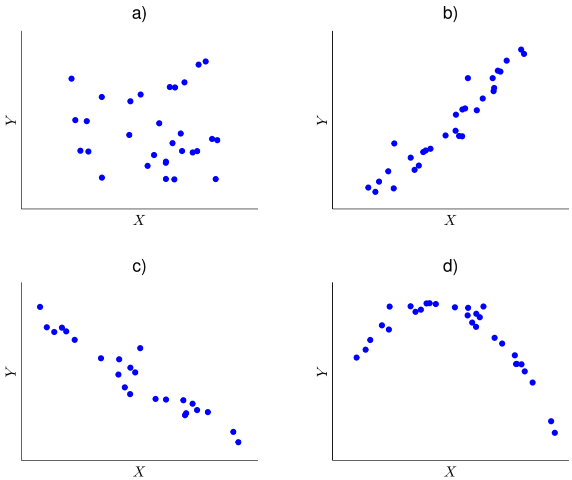
Determinációs együttható:

$$R^2 = r^2 = 0.9881.$$

Nagyon jó a lineáris modell illeszkedése. A regressziós modell az y varianciájának 98.81%-át magyarázza.

Pontdiagramok

A lineáris modell nem mindig megfelelő két változó kapcsolatának leírására. A kapcsolat jellegére a pontdiagram utalhat.



a) X és Y független.

b) X és Y között *pozitív irányú* (lineáris) kapcsolat.

c) X és Y között *negatív irányú* (lineáris) kapcsolat.

d) X és Y között *nemlineáris* kapcsolat.

Nemlineáris regressziós modellek

Valódi nemlineáris modell: nincs olyan transzformáció, amivel lineáris alakra hozható.

Példa. A amplitúdójú, ω frekvenciájú és φ fáziseltolású periodikus függvény additív hibával:

$$Y = A\sin(\omega x + \varphi) + \varepsilon.$$

Paraméterbecslés: általában legkisebb négyzetes becslés. Numerikus optimalizálást igényel. Problémát jelenthet, ha a minimalizálandó függvénynek lokális minimumhelyei is vannak.

Linearizálható modell: alkalmas transzformációval lineárissá alakítható. A transzformált modell paraméterei a lineáris modellre ismertetett módszerrel becsülhetőek, majd ezekből előállíthatóak az eredeti modell paramétereinek becslései.

Példa. S-görbe:

$$Y = \mathrm{e}^{\beta_0 + \beta_1 / X} \cdot \nu, \qquad \text{linearizáltja} \qquad \ln Y = \beta_0 + \beta_1 \cdot \frac{1}{X} + \ln \nu.$$

Exponenciális, hatványkitevős és polinomiális regresszió

Exponenciális regresszió multiplikatív maradékkal:

$$Y = \beta_0 \cdot \beta_1^X \cdot \nu, \qquad \text{linearizáltja} \qquad \ln Y = \ln \beta_0 + X \cdot \ln \beta_1 + \ln \nu.$$

Hatványkitevős regresszió multiplikatív maradékkal:

$$Y = \beta_0 \cdot X^{\beta_1} \cdot \nu, \qquad \text{linearizáltja} \qquad \ln Y = \ln \beta_0 + \beta_1 \ln X + \ln \nu.$$

Elaszticitás:

$$\mathrm{El}(\widehat{y}, x) = \frac{d\widehat{y}}{dx} \cdot \frac{x}{\widehat{y}} = \frac{\widehat{\beta}_0 \cdot \widehat{\beta}_1 \cdot x^{\widehat{\beta}_1 - 1} \cdot x}{\widehat{\beta}_0 \cdot x^{\widehat{\beta}_1}} = \widehat{\beta}_1.$$

Polinomiális regresszió additív maradékkal:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_\ell X^\ell + \varepsilon.$$

$X_i := X^i, \ i = 1, 2, \ldots, \ell,$ jelöléssel többváltozós lineáris modell.

Az SPSS kétváltozós modelljei

1.	Linear:	$Y = \beta_0 + \beta_1 X + \varepsilon;$
2.	Logarithmic:	$Y = \beta_0 + \beta_1 \ln X + \varepsilon;$
3.	Inverse:	$Y = \beta_0 + \beta_1 / X + \varepsilon;$
4.	Quadratic:	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon;$
5.	Cubic:	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon;$
6.	Compound:	$Y = \beta_0 \cdot \beta_1^X \cdot \nu;$ (exponenciális regresszió)
7.	Power:	$Y = \beta_0 \cdot X^{\beta_1} \cdot \nu;$ (hatványkitevős regresszió)
8.	S:	$Y = \mathrm{e}^{\beta_0 + \beta_1 / X} \cdot \nu;$
9.	Growth:	$Y = \mathrm{e}^{\beta_0 + \beta_1 X} \cdot \nu;$
10.	Exponential:	$Y = \beta_0 \cdot \mathrm{e}^{\beta_1 X} \cdot \nu;$
11.	Logistic:	$Y = \left(\frac{1}{u} + \beta_0 \cdot \beta_1^X \right)^{-1} \cdot \nu,$ u adott konstans.

Mindegyik modell linearizálható.

Paraméterbecslés a többváltozós modellben

Többváltozós lineáris modell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$$

A megfigyelt adatokra felírt regresszió:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

Normálegyenletek: $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y}$.

Ha $\boldsymbol{X}^\top \boldsymbol{X}$ invertálható, a megoldás ($\boldsymbol{\beta}$ becslése):

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

A paraméterek értelmezése

Regressziós modell: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Paraméterbecslés: $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Az egyértelmű megoldás létezésének feltétele: $\boldsymbol{X}^\top \boldsymbol{X}$ rangja $k+1$.

Pontosan akkor teljesül, ha $\boldsymbol{X}$ oszlopai **lineárisan függetlenek**.

Legalább **háromszor akkora** minta szükséges, mint ahány paramétert becslünk.

Az $\boldsymbol{y}$ becslése a modell alapján: $\widehat{\boldsymbol{y}} = (\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n)^\top = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

A regressziós hipersík egyenlete: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots \widehat{\beta}_k x_k$.

$\widehat{\beta}_j$ jelentése: az x_j egységnyi növekedése $\widehat{y}$ mekkora változásával jár együtt, ha a többi magyarázó változót rögzítjük $(j = 1, 2, \ldots, k)$.

Rugalmasság, reziduális variancia

Többváltozós lineáris modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$

A paraméterek becslése: $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k)^\top.$

Elaszticitás (rugalmasság):

$$\mathrm{El}(\widehat{y}, x_j) = \frac{\widehat{\beta}_j x_j}{\widehat{y}} = \frac{\widehat{\beta}_j x_j}{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k}.$$

Az x_j magyarázó változó 1%-os változása az eredményváltozó hány százalékos növekedésével/csökkenésével jár együtt, ha az összes többi változó fixen marad. Mindig egy adott $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)^\top$ pontban van értelmezve.

Reziduumok vektora: $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^\top = \boldsymbol{y} - \widehat{\boldsymbol{y}}.$

Reziduális variancia:

$$s_e^{*2} := \frac{\boldsymbol{e}^\top \boldsymbol{e}}{n} = \frac{1}{n} \sum_{i=1}^{n} e_i^2.$$

Korrelációs mátrix

Kérdés. Mennyire szoros a kapcsolat a függő változó és a magyarázó változók között, valamint a magyarázó változók összefüggenek-e egymással?

Korrelációs mátrix:

$$\boldsymbol{R} := \begin{bmatrix} 1 & r_{y1} & r_{y2} & \cdots & r_{yk} \\ r_{1y} & 1 & r_{12} & \cdots & r_{1k} \\ r_{2y} & r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{ky} & r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$

ahol

$$r_{yj} := r(y, x_j), \qquad r_{j\ell} := r(x_j, x_\ell),$$

az (Y, X_j) és (X_j, X_ℓ) párokra vett minták kétváltozós lineáris korrelációs együtthatói.

Szimmetrikus $(k+1) \times (k+1)$ dimenziós mátrix.

Parciális korreláció

Probléma. A lineáris korrelációk figyelembe veszik a változók közötti közvetett kapcsolatokat, például, ha y és x_j, valamint x_j és x_ℓ korrelálnak, akkor ez megjelenik az y és x_ℓ minták $r(y, x_\ell)$ korrelációs együtthatójában.

Közvetlen kapcsolat: a két változó kapcsolatából kiszűrjük mindazt a hatást, ami más változók közvetítésével realizálódik.

Az y és x_j közötti $r_{yj.1,2,\ldots,j-1,j+1,\ldots,k}$ parciális korrelációs együttható azt mutatja, hogy milyen szoros és milyen irányú a sztochasztikus kapcsolat y eredményváltozó és az x_j magyarázó változó között akkor, ha csak a közvetlen kapcsolatot tekintjük, és kiiktatjuk az $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ változókon keresztül érvényesülő közvetett hatásokat.

Korrelációs mátrix: $\boldsymbol{R}$; inverze: $\boldsymbol{R}^{-1} = \left[q_{ij}\right]$.

y és x_j parciális korrelációs együtthatója:

$$r_{yj.1,2,\ldots,j-1,j+1,\ldots,k} = \frac{-q_{yj}}{\sqrt{q_{yy} \cdot q_{jj}}}.$$

Többszörös determinációs együttható

Többváltozós regressziós modell a megfigyelt adatokra:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

Előrejelzett értékek: $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}.$

Négyzetes eltérések:

$$SST := \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 =: SSR + SSE.$$

Többszörös determinációs együttható:

$$R^2 = R^2_{y.1,2,\ldots,k} := SSR/SST = 1 - SSE/SST, \qquad 0 \le R^2 \le 1.$$

Megadja a modell **magyarázó erejét**.

Kapcsolat az $\boldsymbol{R} = \big[r_{ij}\big]$ korrelációs mátrixszal:

$$R^2 = 1 - 1/q_{yy}, \qquad \text{ahol} \qquad \boldsymbol{R}^{-1} = \big[q_{ij}\big].$$

Többszörös korrelációs együttható: $R = \sqrt{R^2}$.

A standard lineáris modell feltételrendszere

$X_1, X_2, \ldots, X_k$ ($\boldsymbol{X}$): magyarázó változók (független változók);

Y: eredményváltozó (függő változó); ε: maradékváltozó.

- A változók közötti kapcsolat **lineáris**, azaz az eredményváltozónak a magyarázó változókra vonatkozó feltételes várható értéke (adott $X_1, X_2, \ldots, X_k$ esetén vett értéke) a magyarázó változók lineáris függvénye:

$$\mathrm{E}(Y|\boldsymbol{X}) = \mathrm{E}(Y|X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$

- A magyarázó változók **nem valószínűségi változók**.
- A magyarázó változók megfigyelt értékei lineárisan független rendszert alkotnak.
- Az ε maradéknak a magyarázó változókra vonatkozó feltételes eloszlása **normális eloszlás** 0 várható értékkel és állandó σ^2 varianciával: $\varepsilon \mid X_1, X_2, \ldots, X_k \sim \mathcal{N}(0, \sigma^2)$.
- A maradékváltozó különböző magyarázó változókhoz tartozó értékei korrelálatlanok:
$$\mathrm{Cov}\left(\varepsilon|X_j, \varepsilon|X_\ell\right) = 0, \quad \text{ha} \quad j \neq \ell.$$

A becsült paramétervektor tulajdonságai

Regressziós modell: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

A $\boldsymbol{\varepsilon}$ maradékvektor ε_i, $i=1,2,\dots,n$, komponensei függetlenek és

$$\varepsilon_i \sim \mathcal{N}(0,\sigma^2), \qquad \text{azaz} \qquad \mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}, \quad \mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}\big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big) = \sigma^2\boldsymbol{E}_n.$$

$\boldsymbol{E}_n$: $n \times n$ dimenziós egységmátrix.

Paraméterbecslés:

$$\widehat{\boldsymbol{\beta}} = \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\big(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\big) = \boldsymbol{\beta} + \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}.$$

Várható érték vektor és kovarianciamátrix:

$$\mathrm{E}\big(\widehat{\boldsymbol{\beta}}\big) = \boldsymbol{\beta} + \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\mathrm{E}\big(\boldsymbol{\varepsilon}\big) = \boldsymbol{\beta}. \quad (\textbf{torzítatlan}\ \text{becslés});$$

$$\begin{aligned}\mathrm{Cov}\,\big(\widehat{\boldsymbol{\beta}}\big) &= \mathrm{E}\Big(\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big)^\top\Big) = \mathrm{E}\Big(\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\Big)\\ &= \sigma^2\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{E}_n\boldsymbol{X}\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1} = \sigma^2\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}.\end{aligned}$$

A paraméterbecslések standard hibája

Regressziós modell: $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$.

ε hibavektor: n-dimenziós normális, $\mathrm{E}(\varepsilon) = \mathbf{0}$, $\mathrm{Cov}(\varepsilon) = \sigma^2 \boldsymbol{E}_n$.

$\widehat{\boldsymbol{\beta}}$ paraméterbecslés vektor: eloszlása $(k+1)$-dimenziós normális,

$$\mathrm{E}\big(\widehat{\boldsymbol{\beta}}\big) = \boldsymbol{\beta}, \quad \mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2 \big(\boldsymbol{X}^\top \boldsymbol{X}\big)^{-1}.$$

$\mathrm{Var}\left(\widehat{\beta}_j\right)$: $\sigma^2 \big(\boldsymbol{X}^\top \boldsymbol{X}\big)^{-1}$ átlójának $(j+1)$-edik eleme, $j = 0, 1, \ldots, k$.

σ^2 nem ismert. Torzítatlan becslése: korrigált reziduális variancia:

$$\widehat{\sigma^2} = s_e^2 := \frac{\boldsymbol{e}^\top \boldsymbol{e}}{n-k-1} = \frac{1}{n-k-1} \sum_{i=1}^{n} e_i^2.$$

$\widehat{\boldsymbol{\beta}}$ becsült kovarianciamátrixa: $\mathrm{cov}\left(\widehat{\boldsymbol{\beta}}\right) := s_e^2 \big(\boldsymbol{X}^\top \boldsymbol{X}\big)^{-1}$.

$\widehat{\beta}_j$ becsült varianciája: $s_e^2 \big(\boldsymbol{X}^\top \boldsymbol{X}\big)^{-1}$ átlójának $(j+1)$-edik eleme. Jelölése: $\mathrm{var}\left(\widehat{\beta}_j\right)$.

$\widehat{\beta}_j$ standard hibája: $s_{\widehat{\beta}_j} := \sqrt{\mathrm{var}\left(\widehat{\beta}_j\right)}$, a $\widehat{\beta}_j$ szórásának becslése.

A paraméterbecslések eloszlása

Regressziós modell: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

$\boldsymbol{\varepsilon}$ hibavektor: n-dimenziós normális, $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{E}_n$.

$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_k)$: eloszlása $(k+1)$-dimenziós normális,

$$\mathrm{E}\big(\widehat{\boldsymbol{\beta}}\big) = \boldsymbol{\beta}, \quad \mathrm{Cov}\,\big(\widehat{\boldsymbol{\beta}}\big) = \sigma^2 \big(\boldsymbol{X}^\top \boldsymbol{X}\big)^{-1}.$$

Az egyes becslések eloszlásai:

$$\widehat{\beta}_j \sim \mathcal{N}\Big(\beta_j, \mathrm{Var}\,\big(\widehat{\beta}_j\big)\Big), \qquad j = 0, 1, \dots, k.$$

$$\frac{\widehat{\beta}_j - \beta_j}{s_{\widehat{\beta}_j}} \sim t_{n-k-1}, \qquad j = 0, 1, \dots, k.$$

$1-\alpha$ megbízhatóságú konfidencia intervallum a β_j együtthatóra:

$$\mathrm{Int}_{\,1-\alpha}\big(\beta_j\big) = \widehat{\beta}_j \pm t_{1-\alpha/2}(n-k-1) \cdot s_{\widehat{\beta}_j}.$$

$t_p(\nu)$: a ν szabadsági fokú t-eloszlás p-kvantilise.

Speciális eset: kétváltozós lineáris regresszió

Kétváltozós ($k = 1$) regressziós modell a megfigyelt adatokra:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, 2, \ldots, n.$$

Korrigált rezuduális variancia: $\widehat{\sigma^2} = s_e^2 := \dfrac{\boldsymbol{e}^\top \boldsymbol{e}}{n-2} = \dfrac{1}{n-2} \sum_{i=1}^{n} e_i^2.$

A paraméterek becslései:

$$\widehat{\beta}_1 = \frac{\sum d_x d_y}{\sum d_x^2}, \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}, \qquad d_x := x - \overline{x},\ d_y := y - \overline{y}.$$

Jellemző	Várható érték	Elméleti variancia	Becsült standard hiba	Eloszlás
$\widehat{\beta}_0$	β_0	$\sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum d_x^2} \right)$	$s_e \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum d_x^2}}$	$\frac{\widehat{\beta}_0 - \beta_0}{s_{\widehat{\beta}_0}} \sim t_{n-2}$
$\widehat{\beta}_1$	β_1	$\frac{\sigma^2}{\sum d_x^2}$	$\frac{s_e}{\sqrt{\sum d_x^2}}$	$\frac{\widehat{\beta}_1 - \beta_1}{s_{\widehat{\beta}_1}} \sim t_{n-2}$
s_e^2	σ^2	$\frac{2\sigma^4}{n-2}$	$s_e^2 \sqrt{\frac{2}{n-2}}$	$\frac{(n-2)s_e^2}{\sigma_2} \sim \mathcal{X}_{n-2}^2$

Konfidencia intervallum az átlagra

Többváltozós modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$.

Mátrix-vektor alak a megfigyelt adatokra: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Paraméterbecslés vektor: $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k)$.

Előrejelzett érték az $\boldsymbol{x}_* = (1, x_{1*}, x_{2*}, \ldots, x_{k*})^\top$ helyen:

$$\widehat{y}_* = \boldsymbol{x}_*^\top \widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1*} + \widehat{\beta}_2 x_{2*} + \cdots + \widehat{\beta}_k x_{k*}.$$

Becslés az $\boldsymbol{x}_*$ helyhez tartozó Y_* egyedi érték $\mathrm{E}(Y_*)$ várható értékére (átlagbecslés).
Varianciája és becsült varianciája:

$$\mathrm{Var}(\widehat{y}_*) = \sigma^2 \boldsymbol{x}_*^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}_*, \qquad \mathrm{var}(\widehat{y}_*) = s_e^2 \boldsymbol{x}_*^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}_*$$

Standard hiba általánosan és a kétváltozós ($k = 1$) esetben:

$$s_{\widehat{y}_*} = s_e \sqrt{\boldsymbol{x}_*^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}_*} \qquad \text{és} \qquad s_{\widehat{y}_*} = s_e \sqrt{\frac{1}{n} + \frac{(x_* - \overline{x})^2}{\sum d_x^2}}.$$

$1 - \alpha$ megbízhatóságú konfidencia intervallum az átlagra:

$$\mathrm{Int}_{\,1-\alpha}\big(\mathrm{E}(Y_*)\big) = \widehat{y}_* \pm t_{1-\alpha/2}(n-k-1) \cdot s_{\widehat{y}_*}.$$

Konfidencia intervallum az egyedi értékre

Többváltozós modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$

Y_*: az $\boldsymbol{x}_* = (1, x_{1*}, x_{2*}, \ldots, x_{k*})^\top$ helyhez tartozó egyedi érték.

E(Y_*) becslése: $\widehat{y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1*} + \widehat{\beta}_2 x_{2*} + \cdots + \widehat{\beta}_k x_{k*}.$

Y_* becslése: $\widehat{Y}_* = \widehat{y}_* + \varepsilon.$

Varianciák:

$$\mathrm{Var}(\widehat{y}_*) = \sigma^2 \boldsymbol{x}_*^\top \big(\boldsymbol{X}^\top \boldsymbol{X}\big)^{-1} \boldsymbol{x}_*, \qquad \mathrm{Var}(\widehat{Y}_*) = \sigma^2 \Big(\boldsymbol{x}_*^\top \big(\boldsymbol{X}^\top \boldsymbol{X}\big)^{-1} \boldsymbol{x}_* + 1\Big).$$

Standard hiba általánosan és a kétváltozós $(k = 1)$ esetben:

$$s_{\widehat{Y}_*} = s_e \sqrt{\boldsymbol{x}_*^\top \big(\boldsymbol{X}^\top \boldsymbol{X}\big)^{-1} \boldsymbol{x}_* + 1} \qquad \text{és} \qquad s_{\widehat{Y}_*} = s_e \sqrt{\frac{1}{n} + \frac{(x_* - \overline{x})^2}{\sum d_x^2} + 1}.$$

$1 - \alpha$ megbízhatóságú konfidencia intervallum az egyedi értékre:

$$\mathrm{Int}_{1-\alpha}\big(Y_*\big) = \widehat{y}_* \pm t_{1-\alpha/2}(n - k - 1) \cdot s_{\widehat{Y}_*}.$$

Példa

$$\begin{array}{lllllllllll} \text{Fogyasztás } (X) & : & 6.0, & 6.1, & 5.9, & 5.4, & 6.5, & 5.5, & 6.7, & 5.7; & \bar{x} = 5.975; \\ CO_2 \text{ kibocsátás } (Y) & : & 139, & 140, & 136, & 128, & 149, & 129, & 155, & 134; & \bar{y} = 138.75. \end{array}$$

A regressziós egyenes egyenlete: $\widehat{y} = 16.9914 + 20.3780 \cdot x$.

$$\begin{array}{lllllllll} \widehat{y}: & 139.26, & 141.30, & 137.22, & 127.03, & 149.45, & 129.07, & 153.52, & 133.15; \\ e: & -0.26, & -1.30, & -1.22, & 0.97, & -0.45, & -0.07, & 1.48, & 0.85. \end{array}$$

Négyzetösszegek:

$$\sum d_x^2 = 1.455, \quad SST = \sum d_y^2 = 611.5, \quad SSE = \sum e^2 = 7.2921, \quad SSR = 604.2079.$$

Modellilleszkedés:

$$R^2 = \frac{SSR}{SST} = \frac{604.2079}{611.5} = 0.9881, \qquad \widehat{\sigma^2} = s_e^2 = \frac{SSE}{n-k-1} = \frac{7.2921}{6} = 1.2153.$$

Paraméterbecslések:

$$\widehat{\beta}_0 = 16.9914, \quad s_{\widehat{\beta}_0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum d_x^2}} = \sqrt{1.2153 \cdot \left(\frac{1}{8} + \frac{5.975^2}{1.455}\right)} = 5.4747;$$

$$\widehat{\beta}_1 = 20.3780, \quad s_{\widehat{\beta}_1} = \frac{s_e}{\sqrt{\sum d_X^2}} = \sqrt{\frac{1.2153}{1.455}} = 0.9139.$$

Példa

X: fogyasztás (l/100km); Y: CO_2 kibocsátás (g/km). $n = 8$, $k = 1$.

Paraméterbecslések:

$$\widehat{\beta}_0 = 16.9914, \quad s_{\widehat{\beta}_0} = 5.4747; \qquad \widehat{\beta}_1 = 20.3780, \quad s_{\widehat{\beta}_1} = 0.9139.$$

95% megbízhatóságú konfidenciaintervallumok a paraméterekre:

$$\mathrm{Int}_{\,1-\alpha}\big(\widehat{\beta}_j\big) = \widehat{\beta}_j \pm t_{1-\alpha/2}(n-k-1)\, s_{\widehat{\beta}_0} = \widehat{\beta}_j \pm t_{0.975}(6)\, s_{\widehat{\beta}_0}, \quad j = 0, 1;$$

$$\mathrm{Int}_{\,95}\big(\widehat{\beta}_0\big) = 16.9914 \pm 2.4469 \cdot 5.4747; \quad \mathrm{Int}_{\,95}\big(\widehat{\beta}_0\big) = \big(3.5953, 30.3875\big);$$
$$\mathrm{Int}_{\,95}\big(\widehat{\beta}_1\big) = 20.3780 \pm 2.4469 \cdot 0.9139; \quad \mathrm{Int}_{\,95}\big(\widehat{\beta}_1\big) = \big(18.1417, 22.6143\big).$$

A regressziós egyenes egyenlete: $\widehat{y} = 16.9914 + 20.3780 \cdot x$.

Új megfigyelés: Skoda Octavia Combi 1.5 TSI (150 LE)

- fogyasztás: 4.9 l/100km;
- CO_2 kibocsátás:: 113 g/km.

Becsült CO_2 kibocsátás: $\widehat{y}_* = 16.9914 + 20.3780 \cdot 4.9 = 116.8436$.

Példa

X: fogyasztás (l/100km); Y: CO_2 kibocsátás (g/km).

A regressziós egyenes egyenlete: $\widehat{y} = 16.9914 + 20.3780 \cdot x$.

$$\overline{x} = 5.975, \quad \sum d_x^2 = 1.455; \quad s_e^2 = 1.2153; \quad s_e = 1.1024$$

Új megfigyelés: $x_* = 4.9$ l/100km; $y_* = 113$ g/km; $\widehat{y}_* = 116.8436$ g/km.

A becslés standard hibája:

$$s_{\widehat{y}_*} = s_e \sqrt{\frac{1}{n} + \frac{(x_* - \overline{x})^2}{\sum d_x^2}} = 1.1024 \cdot \sqrt{\frac{1}{8} + \frac{(4.9 - 5.975)^2}{1.455}} = 1.0570.$$

95% megbízhatóságú konfidenciaintervallum az átlagra:

$$\mathrm{Int}_{1-\alpha}\big(\mathrm{E}(Y_*)\big) = \widehat{y}_* \pm t_{1-\alpha/2}(n-k-1)\, s_{\widehat{y}_*} = \widehat{y}_* \pm t_{0.975}(6)\, s_{\widehat{y}_*};$$

$$\mathrm{Int}_{0.95}\big(\mathrm{E}(Y_*)\big) = 116.8436 \pm 2.4469 \cdot 1.0570;$$

$$\mathrm{Int}_{0.95}\big(\mathrm{E}(Y_*)\big) = \big(114.2573, 119.4299\big).$$

Példa

X: fogyasztás (l/100km); Y: CO_2 kibocsátás (g/km).

A regressziós egyenes egyenlete: $\widehat{y} = 16.9914 + 20.3780 \cdot x$.

$$\overline{x} = 5.975, \quad \sum d_x^2 = 1.455; \quad s_e^2 = 1.2153; \quad s_e = 1.1024$$

Új megfigyelés: $x_* = 4.9$ l/100km; $y_* = 113$ g/km; $\widehat{y}_* = 116.8436$ g/km.

95% megbízhatóságú konfidenciaintervallum az átlagra:

$$\mathrm{Int}_{\,0.975}\big(\mathrm{E}(Y_*)\big) = \big(114.2573, 119.4299\big).$$

Az egyedi érték standard hibája:

$$s_{\widehat{Y}_*} = s_e\sqrt{\frac{1}{n} + \frac{(x_* - \overline{x})^2}{\sum d_x^2} + 1} = 1.1024 \cdot \sqrt{\frac{1}{8} + \frac{(4.9 - 5.975)^2}{1.455} + 1} = 1.5272.$$

95% megbízhatóságú konfidenciaintervallum az egyedi értékre:

$$\mathrm{Int}_{\,1-\alpha}(Y_*) = \widehat{y}_* \pm t_{1-\alpha/2}(n-k-1)\, s_{\widehat{y}_*} = \widehat{y}_* \pm t_{0.975}(6)\, s_{\widehat{Y}_*};$$

$$\mathrm{Int}_{\,0.95}(Y_*) = 116.8436 \pm 2.4469 \cdot 1.5272;$$

$$\mathrm{Int}_{\,0.95}(Y_*) = \big(113.1066, 120.5806\big).$$

Hipotézisvizsgálat a regressziós modellben

Többváltozós lineáris regressziós modell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$$

A hipotézisvizsgálat által megválaszolandó kérdések:

- Az egyes magyarázó változók ténylegesen jó magyarázó változók-e a modellben?
  Az egyes paraméterek esetén szeparáltan teszteljük a $\beta_j = 0$ nullhipotézist.
- A magyarázó változók együttesen kielégítő módon magyarázzák-e az eredményváltozót?
  A $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ hipotézist teszteljük.
- A becslések tükrében helytállóak voltak-e a modellfeltételek?
  Megvizsgáljuk a regressziós maradékok normalitását, korrelálatlanságát és azt, hogy a varianciájuk konstans-e.

A paraméterek szeparált tesztelése

Többváltozós lineáris regressziós modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon, \ \ \varepsilon \sim \mathcal{N}(0, \sigma^2)$.

$\widehat{\beta}_j$: β_j becslése n megfigyelés esetén. $s_{\widehat{\beta}_j}$: $\widehat{\beta}_j$ standard hibája.

Hipotézisek:

$$H_0 : \beta_j = 0; \qquad H_1 : \beta_j \neq 0.$$

Próbafüggvény (parciális t-próba): $t := \widehat{\beta}_j / s_{\widehat{\beta}_j}$.

Ha H_0 teljesül, t eloszlása $n - k - 1$ szabadsági fokú ***t*-eloszlás**.

Adott α szinthez tartozó kritikus tartomány: $|t| \geq t_{1-\alpha/2}(n - k - 1)$.

Ha valamelyik $j \in \{1, 2, \ldots, k\}$ esetén elfogadjuk a nullhipotézist, akkor az X_j magyarázó változó kihagyható a regressziós modellből.

Ha a $\beta_0 = 0$ hipotézist fogadjuk el, akkor nem kell konstans.

Újra kell számolni a paraméterek becsléseit, a konfidencia intervallumokat és újra el kell végezni a hipotézisek vizsgálatát.

A modell egészének tesztelése

Többváltozós lineáris regressziós modell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Hipotézisek:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0; \qquad H_1 : \exists \beta_j \neq 0.$$

$y_i, \widehat{y}_i,\ i = 1, 2, \ldots n$: megfigyelt és előrejelzett értékek.

$$SST := \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 =: SSR + SSE.$$

Próbafüggvény (globális F-próba):

$$F := \frac{SSR/k}{SSE/(n-k-1)}.$$

Ha H_0 teljesül:

$$\frac{SSR}{\sigma^2} \sim \mathcal{X}_k^2, \qquad \frac{SSE}{\sigma^2} \sim \mathcal{X}_{n-k-1}^2, \qquad \frac{SST}{\sigma^2} \sim \mathcal{X}_{n-1}^2, \qquad F \sim F_{k,n-k-1}.$$

Varianciaanalízis-táblázat

Modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Hipotézisek: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0; \qquad H_1 : \exists \beta_j \neq 0$.

$$SSR := \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2, \qquad SSE := \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2, \qquad SST = SSR + SSE.$$

Varianciaanalízis-táblázat:

A variancia forrása	Négyzet-összeg	Szabadsági fok	Átlagos négyzetösszeg	F-érték	p-érték
Regresszió	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$	p
Maradék	SSE	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$	–	–
Teljes	SST	$n-1$	–	–	–

Ha H_0 teljesül, $F \sim F_{k,n-k-1}$, valamint $\widehat{y}_i \approx \overline{y}$, azaz SSR **kicsi**.

Adott α szignifikanciaszinthez tartozó kritikus tartomány: $F \geq F_{1-\alpha}(k, n-k-1)$.

H_0 igaz: a modell **teljes egészében rossz**, egyik magyarázó változót sem érdemes megtartani.

Összefüggés a többszörös determinációs együtthatóval

Globális F-próba próbastatisztikája:

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{n-k-1}{k} \cdot \frac{SSR}{SSE}.$$

Többszörös determinációs együttható:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}.$$

Kapcsolat:

$$F = \frac{n-k-1}{k} \cdot \frac{SSR}{SST - SSR} = \frac{n-k-1}{k} \cdot \frac{SSR/SST}{1 - SSR/SST} = \frac{n-k-1}{k} \cdot \frac{R^2}{1 - R^2}.$$

Speciális eset: kétváltozós lineáris regresszió

Kétváltozós ($k = 1$) modell: $Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$.

A meredekségre vonatkozó parciális t-próba hipotézisei:

$$H_0 : \beta_1 = 0; \qquad H_1 : \beta_1 \neq 0.$$

Próbafüggvény:

$$t = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}}, \qquad \text{ahol} \qquad \widehat{\beta}_1 = \sqrt{\frac{SSR}{\sum d_x^2}}, \quad s_{\widehat{\beta}_1} = \sqrt{\frac{s_e^2}{\sum d_x^2}} = \sqrt{\frac{SSE/(n-2)}{\sum d_x^2}}.$$

Ha H_0 igaz, $t \sim t_{n-2}$. Kritikus tartomány: $|t| \geq t_{1-\alpha/2}(n-2)$.

A globális F-próba hipotézisei: $H_0 : \beta_1 = 0; \quad H_1 : \beta_1 \neq 0$.

Próbafüggvény:

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{SSR}{SSE/(n-2)} = \frac{\widehat{\beta}_1^2}{s_{\widehat{\beta}_1}^2} = t^2.$$

Ha H_0 igaz, $F \sim F_{1,n-2}$. Kritikus tartomány: $F \geq F_{1-\alpha}(1, n-2)$.

Példa

X: fogyasztás (l/100km); Y: CO_2 kibocsátás (g/km). $n = 8,\ k = 1.$

Paraméterbecslések:

$$\widehat{\beta}_0 = 16.9914, \quad s_{\widehat{\beta}_0} = 5.4747; \qquad \widehat{\beta}_1 = 20.3780, \quad s_{\widehat{\beta}_1} = 0.9139.$$

Parciális t-próbák

A próbafüggvény: $t = \widehat{\beta}_j / s_{\widehat{\beta}_j},\ j = 0, 1;$ eloszlása H_0 teljesülése esetén: t_{n-2}.

Paraméter	Nullhipotézis	Ellenhipotézis	Próbastatisztika	p-érték
Konstans	$H_0 : \beta_0 = 0$	$H_1 : \beta_0 \neq 0$	$\frac{\widehat{\beta}_0}{s_{\widehat{\beta}_0}} = \frac{16.9914}{5.4747} = 3.1036$	0.0210
Meredekség	$H_0 : \beta_1 = 0$	$H_1 : \beta_1 \neq 0$	$\frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}} = \frac{20.3780}{0.9139} = 22.2968$	0.0000

5%-os szinten a kritikus tartomány: $|t| \geq t_{0.975}(6) = 2.4469.$

5%-os szinten mindkét nullhipotézist **elvetjük**. Mindkét paraméter **szignifikáns**.

Példa

X: fogyasztás (l/100km); Y: CO_2 kibocsátás (g/km). $n = 8$, $k = 1$.

Négyzetösszegek: $SST = 611.5$, $SSE = 7.2921$, $SSR = 604.2079$.

Globális F-próba

Hipotézisek: $H_0 : \beta_1 = 0$; $H_1 : \beta_1 \neq 0$.

A próbafüggvény: $F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$.

A próbafüggvény eloszlása H_0 teljesülése esetén: $F_{1,n-2}$.

Varianciaanalízis-táblázat:

A variancia forrása	Négyzet-összeg	Szabadsági fok	Átlagos négyzetösszeg	F-érték	p-érték
Regresszió	604.21	1	$\frac{604.21}{1} = 604.21$	$\frac{604.21}{1.2153} = 497.14$	0.0000
Maradék	7.2921	6	$\frac{7.2921}{6} = 1.2153$	–	–
Teljes	611.5	7	–	–	–

5%-os szinten a kritikus tartomány: $F \geq F_{0.95}(1,6) = t^2_{0.975}(6) = 5.9874$.

5%-os szinten **elvetjük** a nullhipotézist. **Van értelme** a két változó közötti lineáris regressziós modellnek.

A regressziós maradékok vizsgálata

Többváltozós regressziós modell a megfigyelt adatokra:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Előrejelzett értékek: $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}$.

Maradékok: $e_i = y_i - \widehat{y}_i, \ i = 1, 2, \ldots, n.$

A modellfeltételek alapján a maradékváltozók normális eloszlásúak azonos szórással és korrelálatlanok.

Normalitásvizsgálat

- Grafikus módszerek:
  - PP plot, QQ plot vizsgálata;
  - a hisztogram vizsgálata.
- Hipotézisvizsgálat:
  - Kolmogorov-Szmirnov próba;
  - Lillieforst próba;
  - Jarque-Bera próba.

A maradékok korrelálatlansága, Durbin-Watson teszt

ε_t: a regressziós modellben a t-edik megfigyeléshez tartozó maradékváltozó. Feltesszük, hogy a megfigyelések (és így a maradékok) sorrendje kötött. Becslése: $e_t = y_t - \widehat{y}_t,\ t = 1, 2, \ldots, n$.

Elsőrendű autokorreláció a maradékokra:

$$\varepsilon_t = \varrho\varepsilon_{t-1} + \eta_t, \qquad \eta_t \sim \mathcal{N}\left(0, \sigma^2\right), \text{ korrelálatlanok.}$$

$\varrho = 0$: az eredeti maradékváltozók is **normálisak azonos szórással** és **korrelálatlanok**.

Hipotézisek: $H_0 : \varrho = 0; \quad H_1 : \varrho \neq 0.$

ϱ becslése: $\widehat{\varrho} := \dfrac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}.$

Az e_t értékek becsültek, így a $\widehat{\varrho}$ eloszlása nem kezelhető. Nem alkalmas próbafüggvénynek.

Próbafüggvény

Elsőrendű autokorreláció a maradékokra: $\varepsilon_t = \varrho\varepsilon_{t-1} + \eta_t$.

Hipotézisek: $H_0 : \varrho = 0$; $H_1 : \varrho \neq 0$.

Próbafüggvény: $d := \dfrac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_t^2}$. $d \approx 2(1 - \widehat{\varrho})$.

Ha H_0 igaz, $d \in [0, 4]$ és a d eloszlása szimmetrikus a 2 pontra.

d_L, d_U: a próba kritikus értékei. Táblázatból, α, n, k függvényei.

Alkalmazása:

- $d < d_L$: elfogadjuk, hogy a maradékok **pozitív** autokorrelációval bírnak;
- $d_L < d < d_U$: **nem lehet dönteni**, semleges zóna;
- $d_U < d < 4 - d_U$: **elfogadjuk** H_0-at, miszerint nincs elsőrendű autokorreláció;
- $4 - d_U < d < 4 - d_L$: **nem lehet dönteni**, semleges zóna;
- $d > 4 - d_L$: elfogadjuk, hogy a maradékok **negatív** autokorrelációval bírnak.

Autokorreláció kiszűrése

Kétváltozós modell autokorreláló maradékokkal:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \qquad \varepsilon_t = \varrho\varepsilon_{t-1} + \eta_t.$$

η_t: korrelálatlan, $\mathcal{N}\big(0, \sigma^2\big)$ eloszlású sorozat.

Modelltranszformáció:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \varepsilon_t; \\ y_{t-1} &= \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}; \\ \varrho y_{t-1} &= \varrho\beta_0 + \varrho\beta_1 x_{t-1} + \varrho\varepsilon_{t-1}. \end{aligned}$$

Az első és utolsó sort kivonva egymásból:

$$y_t - \varrho y_{t-1} = (1-\varrho)\beta_0 + \beta_1\big(x_t - \varrho x_{t-1}\big) + \varepsilon_t - \varrho\varepsilon_{t-1}.$$

A transzformált modell maradékai már eleget tesznek a regressziós modell feltételeinek.

Ismerni kell a ϱ értékét. A mintából becsülendő.

Autokorrelációs függvény

ε_t: a regressziós modellben a t-edik megfigyeléshez tartozó maradékváltozó, $t = 1, 2, \ldots, n$.

A maradékok *h-lépéses autokorrelációi*:

$$\varrho(h) := \frac{\mathrm{E}(\varepsilon_t \varepsilon_{t-h})}{\mathrm{Var}(\varepsilon_t)}, \qquad h = 0, 1, \ldots, n-1.$$

Ha teljesülnek a regressziós modell feltételei:

$$\varrho(0) = 1, \quad \varrho(h) = 0, \qquad h = 1, 2, \ldots, n-1.$$

$e_t = y_t - \widehat{y}_t$: a regressziós modell becsült maradékai.

A maradékok *h-lépéses becsült autokorrelációi*:

$$\widehat{\varrho}(h) := \frac{\sum_{t=h+1}^{n} e_t e_{t-h}}{\sum_{t=1}^{n} e_t^2}, \qquad h = 0, 1, \ldots, n-1.$$

Autokorrelációs függvény (ACF): a becsült autokorrelációk diagramja. Alkalmas a korrelálatlanság grafikus vizsgálatára.

A maradékok szórásának vizsgálata

Homoszkedasztikus modell: a maradékváltozók azonos szórásúak. Ellenkező esetben a modell heteroszkedasztikus.

A heteroszkedaszticitás jellege többféle lehet:

- A maradékok két csoportot alkotnak, csoportokon belül homoszkedasztikusak.

  Például a változók időfüggőek és valamely időpontban a változók közötti kapcsolat megváltozik. Áttérés a piacgazdaságra: megnövekszik a jövedelmek szórása.

- Több homoszkedasztikus csoport (csoportos heteroszkedaszticitás).

  Például rétegzett mintavétel jövedelemkategóriánként. Regresszió a kategóriaátlagokra: a szórás függ a mintanagyságtól.

- A variancia valamely magyarázó változó értékével együtt változik (funkcionális heteroszkedaszticitás).

  Például a jövedelem növekedésével nő az élelmiszerekre költött összeg szórása.

Goldfeld-Quandt próba

A megfigyeléseket valamely X magyarázó változó szerint sorba rendezve a nagy, illetve a kis változóértékekhez tartozó maradékok varianciáinak egyenlőségét vizsgáljuk. Homoszkedasztikus esetben a varianciák megegyeznek.

1. Kiválasztunk egy az X **kis** értékeihez tartozó n_1 elemű mintát az eredmény- és a magyarázó változókból. Az adatokra egy lineáris modellt illesztünk.

σ_1^2: az illesztett regressziós modell maradékainak varianciája.

$e_{1,1}, e_{1,2}, \ldots, e_{1,n_1}$: az illesztett modell becsült maradékai.

2. Kiválasztunk egy az X **nagy** értékeihez tartozó n_2 elemű mintát az eredmény- és a magyarázó változókból. Az adatokra egy lineáris modellt illesztünk.

σ_2^2: az illesztett regressziós modell maradékainak varianciája.

$e_{2,1}, e_{2,2}, \ldots, e_{2,n_2}$: az illesztett modell becsült maradékai.

3. F-próbával teszteljük a következő hipotéziseket:

$$H_0 : \sigma_1^2 = \sigma_2^2; \qquad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Próbafüggvény

σ_1^2, σ_2^2: az X magyarázó változó kis és nagy értékeihez tartozó mintákra illesztett k-változós regressziós modellek maradékainak varianciái.

$e_{1,1}, e_{1,2}, \dots, e_{1,n_1}$ és $e_{2,1}, e_{2,2}, \dots, e_{2,n_2}$: az illesztett modellek becsült maradékai.

Hipotézisek: $H_0 : \sigma_1^2 = \sigma_2^2;\qquad H_1 : \sigma_1^2 \neq \sigma_2^2.$

F-próba a varianciák egyezésére. Próbafüggvény:

$$F^* := \max\left\{F, \frac{1}{F}\right\}, \qquad \text{ahol} \qquad F := \frac{\left(\sum_{i=1}^{n_1} e_{1,i}^2\right)(n_2 - k - 1)}{\left(\sum_{i=1}^{n_2} e_{2,i}^2\right)(n_1 - k - 1)}.$$

Ha H_0 teljesül: $F \sim F_{n_1-k-1,n_2-k-1},\ 1/F \sim F_{n_2-k-1,n_1-k-1}$.

Adott α szinthez tartozó kritikus tartomány:

$$F^* \geq F_{1-\alpha/2}(\nu_1, \nu_2).$$

$\nu_1,\ \nu_2$: a számláló, illetve a nevező szabadsági foka.

A próba végrehajtása

Gyakorlati javaslatok:

- Ha semmi nem szól ellene, akkor az $n_1 = n_2$ választás javasolt. A próbastatisztika a két négyzetösszeg hányadosa.

- A megfigyeléseket nem két, hanem három részre osztják. A középső résznek megfelelő ℓ darab megfigyelést kihagyják.

  Előnye: jobban szétvállnak az X magyarázó változó kis és a nagy értékei.

  Hátránya: ℓ választása szubjektív, nincs kritérium a nagyságára.

  Javaslat: ha $n_1 = n_2$, akkor $\ell \leq n/3$.

- Mivel n_1, n_2 és ℓ választása szubjektív, célszerű több kombinációt is kipróbálni és megvizsgálni, hogy a teszt kellően robusztus-e a feltételek változására.

Modellválasztás

Többváltozós regressziós modell a megfigyelt adatokra:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

Többszörös determinációs együttható:

$$R^2 := \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \qquad 0 \leq R^2 \leq 1.$$

A magyarázó változók számának növelésével az R^2 értéke **nem csökken**, a gyakorlatban mindig **nő**.

A magyarázó változók számának növelésével:

- megnő a veszélye annak, hogy a magyarázó változók összefüggenek (multikollinearitás);
- csökken a maradékok $n - k - 1$ szabadsági foka.

Szabadságfokkal korrigált (adjusted) R^2:

$$R_a^2 := 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right).$$

Bünteti a magyarázó változók számának növelését.

Változószelekció, forward eljárás

Modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$

Algoritmus:

1. A magyarázó változók közül kiválasztjuk a függő változóval legjobban korrelálót és ezzel felírjuk a kétváltozós regressziós modellt.
2. Globális F-próbával megvizsgáljuk, értelmes-e a kapott modell.
3. Amennyiben igen, egyenként megvizsgáljuk a kimaradt változókkal való bővítés lehetőségét. Kipróbáljuk, egy adott változóval való modellbővítés esetén az új változó szignifikáns-e (parciális t-próba).

   SPSS: F-change ($F = t^2$; t: a parciális t-próba próbafüggvénye) szignifikanciája egy adott érték alatt van-e.
4. Ha nem találunk olyan változót, ami szignifikáns, megállunk.
5. Ha vannak bevihető (szignifikáns) változók, akkor azok közül a legnagyobb abszolút értékű parciális korrelációs együtthatóval bíróval bővítjük a modellt és visszatérünk a 2. ponthoz.

Változószelekció, backward eljárás

Modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$

Algoritmus:

1. Az összes magyarázó változót felhasználva felírjuk a lineáris regressziós modellt.
2. Globális F-próbával megvizsgáljuk, értelmes-e a kapott modell.
3. Amennyiben igen, egyenként megvizsgáljuk a magyarázó változókat, van-e közöttük nem szignifikáns (parciális t-próba). Ha nincs, megállunk.

   SPSS: F-change szignifikanciája egy adott érték felett van-e.
4. A nem szignifikáns változók közül a legkisebb abszolút értékű parciális korrelációs együtthatóval bírót kihagyjuk a modellből és visszatérünk a 2. ponthoz.

Idősorok

Cél: jelenségek időbeli alakulásának, lefolyásának vizsgálata, modellezése és előrejelzése a mo- dellek segítségével.

Példák: az OTP részvények napi záróárolyama; Magyarország népessége minden év első napján; az éves bruttó átlagbér alakulása; az export havi teljes értékösszege.

Tartam (flow) idősor: az idősor értékei egy időszakra vonatkoznak. Például az éves bruttó átlagbér alakulása; az export havi teljes értékösszege.

Állapot (stock) idősor: az idősor értékei egy időpontra vonatkoznak. Például az OTP részvények napi záróárolyama; Magyarország népessége minden év első napján.

Y_t: az idősor értéke a t időpontban. Valószínűségi változó.

$\ldots, Y_{-T}, Y_{-T+1}, \ldots, Y_t, \ldots, Y_T, \ldots$: elméleti idősor.

$y_1, y_2, \ldots, y_n$: megfigyelt idősor. Az elméleti idősor egy részének megfigyelt értékeiből áll.

Dekompozíciós idősormodellek

Additív dekompozíciós idősormodell:

$$Y = \widehat{Y} + S + C + \varepsilon.$$

Multiplikatív dekompozíciós idősormodell:

$$Y = \widehat{Y} \cdot S^* \cdot C^* \cdot \nu.$$

A komponensek:

- $\widehat{Y}$: hosszú távú alapirányzat, trend;
- S, S^*: a szabályos rövid távú (többnyire havi vagy negyedéves) ingadozást leíró szezonális komponens;
- C, C^*: a szabálytalan hosszabb távú ingadozásokat leíró ciklikus komponens;
- ε, ν: a zavaró hatásokat leíró véletlen összetevők, hibatagok, $\mathrm{E}(\varepsilon) = 0$, $\mathrm{E}(\nu) = 1$.

Trendszámítás

Dekompozíciós modell:

$$Y = \widehat{Y} + \varepsilon, \qquad \text{vagy} \qquad Y = \widehat{Y} \cdot \nu.$$

Kétféle megközelítés:

- A trend valamilyen analiktikusan jól leírható függvény szerint alakul vagy ilyennel jól közelíthető. A cél ennek a függvénynek a becslése. Ez az analitikus trendszámítás.

  Trend meghatározása: kétváltozós regresszió, ahol a független változó az idő.

  Előrejelzésre is alkalmas.

- A trendről nem feltételezzük, hogy analitikusan leírható. Becslését csupán az idősor megfigyelt értékeinek különféle átlagaiból állítjuk elő. Ez a mozgóátlagolású trendszámítás.

  Előrejelzésre nem alkalmas.

Lineáris trendszámítás

A lineáris trendszámítás alapmodellje:

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t, \qquad \mathrm{E}\varepsilon_t = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2, \qquad -T \le t \le T.$$

A megfigyelt idősorra felírt modell:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t, \qquad t = 1, 2, \ldots, n.$$

Paraméterbecslés: legkisebb négyzetek módszere. Normálegyenletek:

$$\beta_0 n + \beta_1 \sum_{t=1}^{n} t = \sum_{t=1}^{n} y_t, \qquad \beta_0 \sum_{t=1}^{n} t + \beta_1 \sum_{t=1}^{n} t^2 = \sum_{t=1}^{n} t y_t.$$

A normálegyenletek megoldása:

$$\widehat{\beta}_1 = 6 \cdot \frac{2 \sum_{t=1}^{n} t y_t - n(n+1)\overline{y}}{n(n^2-1)}, \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \frac{n+1}{2}.$$

$\widehat{\beta}_0$: a $t = 0$ időpillanathoz tartozó trendérték.

$\widehat{\beta}_1$: a trendfüggvény meredeksége. Egy időegység alatt mekkora az idősorban az egy időszakra jutó változás.

Maradékok, előrejelzések

Lineáris trend modell: $y_t = \beta_0 + \beta_1 t + \varepsilon_t, \qquad t = 1, 2, \ldots, n.$

$\widehat{\beta}_0, \widehat{\beta}_1$: a β_0 és β_1 paraméterek legkisebb négyzetes becslései.

A trendfüggvény becsült értékei a megfigyelt időszakban:

$$\widehat{y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t, \qquad t = 1, 2, \ldots, n.$$

A véletlen komponens tapasztalati értékei (reziduumok):

$$e_t = y_t - \widehat{y}_t, \qquad t = 1, 2, \ldots, n.$$

Reziduális variancia:

$$s_e^{*2} = \frac{1}{n} \sum_{t=1}^{n} e_t^2.$$

Illeszkedési mutató, minél kisebb, annál jobban írja le a lineáris trend az idősor megfigyelt értékeit. A lineáris regressziós modell többi mutatója (R, R^2, R_a^2) ugyancsak használható.

A becsült trend segítségével adott előrejelzések:

$$\widehat{y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t, \qquad t = n+1, n+2, \ldots.$$

Példa

Az alábbi táblázat a vodka (0.5ℓ, palack) éves fogyasztói átlagárát (Ft) adja meg a 2011-2022 időszakban (forrás: KSH).

Év:	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Ár:	1700	1810	1870	1880	2220	2160	2540	2580	2740	2790	2890	3140

Jellemezze az idősort lineáris trenddel, becsülje meg a trend paramétereit, számolja ki a trendfüggvény értékeit a megfigyelési időszakra, készítsen előrejelzést a 2023-as évre és ábrázolja az elmondottakat.

Megoldás. y_t: az éves fogyasztói átlagár a t időpontban.

Időparaméter: $t = 1, 2, \ldots, 12$ ($t = 1$: a 2011. év).

Összegek: $n = 12$, $\sum t = 78$, $\sum t^2 = 650$, $\sum y_t = 28320$, $\sum ty_t = 202960$.

Normálegyenletek:

$$12\beta_0 + 78\beta_1 = 28320; \qquad 78\beta_0 + 650\beta_1 = 202960.$$

Paraméterbecslések (a normálegyenletek megoldásai):

$$\widehat{\beta}_1 = 6 \cdot \frac{2\sum ty_t - n(n+1)\overline{y}}{n(n^2-1)} = 6 \cdot \frac{2 \cdot 202960 - 13 \cdot 28320}{12 \cdot 143} = 132.0280;$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \frac{n+1}{2} = \frac{28320}{12} - 132.0280 \cdot \frac{13}{2} = 1501.8182.$$

Példa

y_t: a vodka éves fogyasztói átlagára (Ft) a 2011-2022 időszakban.

Időparaméter: $t = 1, 2, \ldots, 12$ $(t = 1$: a 2011. év$)$.

A becsült lineáris trendfüggvény: $\widehat{y}_t = 1501.8182 + 132.0280 \cdot t$.

A megfigyelt értékek (y_t), a trendértékek becslései $(\widehat{y}_t)$ és a maradékok (e_t) a vizsgált időszakban:

Év	2011	2012	2013	2014	2015	2016
y_t :	1700	1810	1870	1880	2220	2160
$\widehat{y}_t$:	1633.8	1765.9	1897.9	2029.9	2162.0	2294.0
e_t :	66.1538	44.1259	-27.9021	-149.9301	58.0420	-133.9860

Év	2017	2018	2019	2020	2021	2022
y_t :	2540	2580	2740	2790	2890	3140
$\widehat{y}_t$:	2426.0	2558.0	2690.1	2822.1	2954.1	3086.2
e_t :	113.9860	21.9580	49.9301	-32.0979	-64.1259	53.8462

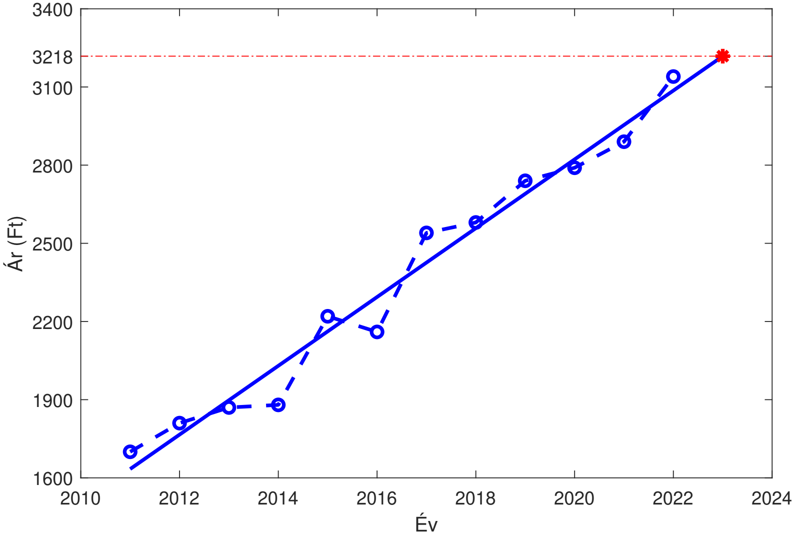
Reziduális négyzetösszeg: $\sum e_t^2 = 74911.8881$.

Reziduális variancia: $s_e^{*2} = \frac{1}{n} \sum e_t^2 = \frac{74911.8881}{12} = 6242.6573$.

Előrejelzés a 2023. évre $(t = 13)$: $\widehat{y}_{13} = 1501.8182 + 132.0280 \cdot 13 = 3218.1818$.

Példa

y_t: a vodka éves fogyasztói átlagára (Ft) a 2011-2022 időszakban.



Nemlineáris trendszámítás, exponenciális modell

Az exponenciális trend modellje a megfigyelt idősorra:

$$y_t = \beta_0 \beta_1^t \nu_t, \qquad \mathrm{E}(\nu_t) = 1, \qquad t = 1, 2, \ldots, n.$$

Mindkét oldal logaritmusát véve visszavezethető a lineáris modellre:

$$v_t = \gamma_0 + \gamma_1 t + \eta_t, \qquad t = 1, 2, \ldots, n,$$

ahol

$$v_t = \log y_t, \quad \gamma_0 = \log \beta_0, \quad \gamma_1 = \log \beta_1 \quad \eta_t = \log \nu_t.$$

$\widehat{\gamma}_0, \ \widehat{\gamma}_1$: a γ_0 és γ_1 legkisebb négyzetes becslései.

β_0 és β_1 becslései:

$$\widehat{\beta}_0 = \exp\left(\widehat{\gamma}_0\right), \qquad \widehat{\beta}_1 = \exp\left(\widehat{\gamma}_1\right).$$

A becslések torzítottak, a gyakorlatban a torzítás elhanyagolható.

A trendfüggvény becsült értékei a megfigyelt időszakban:

$$\widehat{y}_t = \widehat{\beta}_0 \widehat{\beta}_1^t, \qquad t = 1, 2, \ldots, n.$$

Illeszkedési mutatók

Exponenciális trend modell: $y_t = \beta_0 \beta_1^t \nu_t, \qquad t = 1, 2, \ldots, n.$

Becsült trend: $\widehat{y}_t = \widehat{\beta}_0 \widehat{\beta}_1^t, \qquad t = 1, 2, \ldots, n.$

$\widehat{\beta}_0$: a $t = 0$ időpillanathoz tartozó trendérték.

$\widehat{\beta}_1$: a trendfüggvény növekedési üteme.

$$\frac{\widehat{y}_{t+1}}{\widehat{y}_t} = \frac{\widehat{\beta}_0 \widehat{\beta}_1^{t+1}}{\widehat{\beta}_0 \widehat{\beta}_1^t} = \widehat{\beta}_1.$$

Maradékok becslése: $u_t = y_t / \widehat{y}_t$.

Reziduális variancia: $s_u^{*2} = \frac{1}{n} \sum_{t=1}^{n} (u_t - 1)^2$.

Leginkább az összehasonlításoknál használt illeszkedési mutató:

$$s_e^{*2} = \frac{1}{n} \sum_{t=1}^{n} (y_t - \widehat{y}_t)^2.$$

Nemlineáris trendszámítás, logisztikus modell

A rövid távon exponenciális jelleget mutató folyamatok hosszabb távon gyakran ún. S görbe alakú korlátos növekedési folyamattá válnak.

S görbe: kezdetben gyorsuló módon, exponenciális függvény szerint nő. Közeledve a felső korláthoz (telítettségi szint) a növekedés lassul.

A logisztikus trend modellje a megfigyelt idősorra:

$$y_t = \frac{k}{1+\beta_0 \mathrm{e}^{-\beta_1 t}} + \varepsilon_t, \qquad t = 1, 2, \ldots, n.$$

Paraméterezés:

- k: telítődési paraméter;
- β_0: helyzetparaméter;
- β_1: alakparaméter.

Nemlineáris trendszámítás, polinomiális modell

A polinomiális trend modellje a megfigyelt idősorra:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p + \varepsilon_t, \qquad t = 1, 2, \ldots, n.$$

Paraméterbecslések: legkisebb négyzetek módszere. Az előrejelzés pontossága nagy mértékben függ a p fokszám választásától.

Javaslatok az alkalmazáshoz

- Elemezzük a megfigyeléseken kívüli információkat, amelyek alapján valószínűsíthetjük a trend alakját, jellemzőit.
- Ne használjunk túl magas fokszámú polinomot, legfeljebb harmadfokú polinom javasolt. SPSS: $p = 1, 2, 3$.
- Az előrejelzéseknél gondoljuk végig, reális lehet-e azok iránya és nagysága.
- Ha össze akarjuk hasonlítani a különböző fokszámú polinomok illeszkedését, az $s_e^2 = \frac{1}{n-p-1} \sum_{t=1}^{n} e_t^2$ szabadságfokkal korrigált reziduális varianciát és az R_a^2 szabadságfokkal korrigált R^2 mutatót használjuk.

Mozgóátlagolású trendszámítás

Probléma. Nincs elég ismeretünk a vizsgált folyamatról ahhoz, hogy megadjuk a trend analitikus alakját.

Mozgóátlagolású trendszámítás: az idősor t-edik eleméhez úgy rendelünk trendértéket, hogy átlagoljuk az idősor t-edik elemének bizonyos környezetében lévő elemeket.

Példa. Három tagú szimmetrikus mozgóátlag, MA(3).

$y_1, y_2, \ldots, y_n$: a megfigyelt idősor.

Mozgóátlagolású trend:

$$\widehat{y}_t = \frac{y_{t-1} + y_t + y_{t+1}}{3}, \qquad t = 2, 3, \ldots, n-1.$$

A mozgóátlagolású trend mindig **rövidebb** az eredeti idősornál. Szimmetrikus esetben az idősor mindkét végén hiányoznak a becsült trendértékek.

Általános szimmetrikus mozgóátlagok

$y_1, y_2, \dots, y_n$: a megfigyelt idősor.

Az m tagú mozgóátlagolású (MA(m)) trend alakja:

- m páratlan, azaz $m = 2k+1$, $k = 0, 1, 2, \dots$:
$$\widehat{y}_t = \frac{y_{t-k} + y_{t-k-1} + \dots + y_t + \dots + y_{t+k}}{2k+1};$$
- m páros, azaz $m = 2k$, $k = 0, 1, 2, \dots$:
$$\widehat{y}_t = \frac{\frac{1}{2}y_{t-k} + y_{t-k-1} + \dots + y_t + \dots + y_{t+k-1} + \frac{1}{2}y_{t+k}}{2k},$$

  ahol $t = k+1, k+2, \dots, n-k$.

A trend első és utolsó k értékét nem tudjuk kiszámolni.

SPSS: a fentiek mellett páros m esetén nem azonos súlyozású mozgóátlagot is számol:
$$\widehat{y}_t = \frac{y_{t-k} + y_{t-k-1} + \dots + y_t + \dots + y_{t+k-2} + y_{t+k-1}}{2k},$$
ahol $t = k+1, k+2, \dots, n-k+1$.

Példa

Lineáris trend a megfigyelt idősorra:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t, \qquad \mathsf{E}(\varepsilon_t) = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2, \qquad t = 1, 2, \ldots, n.$$

Három tagú mozgóátlagolású trend:

$$\widehat{y}_t = \frac{y_{t-1} + y_t + y_{t+1}}{3} = \frac{\beta_0 + \beta_1(t-1) + \varepsilon_{t-1} + \beta_0 + \beta_1 t + \varepsilon_t + \beta_0 + \beta_1(t+1) + \varepsilon_{t+1}}{3}$$
$$= \beta_0 + \beta_1 t + \frac{\varepsilon_{t-1} + \varepsilon_t + \varepsilon_{t+1}}{3}.$$

A trend várható értéke és varianciája:

$$\mathsf{E}\big(\widehat{y}_t\big) = \beta_0 + \beta_1 t + \frac{1}{3}\mathsf{E}\big(\varepsilon_{t-1} + \varepsilon_t + \varepsilon_{t+1}\big) = \beta_0 + \beta_1 t,$$

$$\mathrm{Var}\,\big(\widehat{y}_t\big) = \frac{1}{9}\,\mathrm{Var}\,\big(\varepsilon_{t-1} + \varepsilon_t + \varepsilon_{t+1}\big) = \frac{1}{9}(3\sigma^2) = \frac{\sigma^2}{3}.$$

$\widehat{y}_t$ az y_t idősor trendjének **torzítatlan** becslése, szórásnégyzete az eredeti szórásnégyzet **harmada**.

Példa

y_t: a vodka éves fogyasztói átlagára (Ft) a 2011-2022 időszakban.

Időparaméter: $t = 1, 2, \ldots, 12$ ($t = 1$: a 2011. év).

A becsült lineáris trendfüggvény: $\widehat{y}_t = 1501.8182 + 132.0280 \cdot t$.

Három tagú mozgóátlagolású trend:

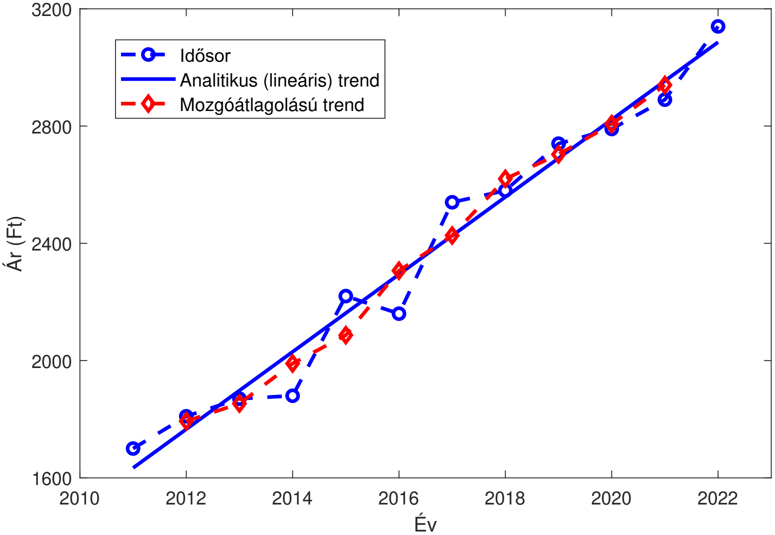
$$\widehat{y}_t^* = \frac{y_{t-1} + y_t + y_{t+1}}{3}, \quad t = 2, 3, \ldots, 9.$$

Év	2011	2012	2013	2014	2015	2016
y_t :	1700	1810	1870	1880	2220	2160
$\widehat{y}_t$:	1633.8	1765.9	1897.9	2029.9	2162.0	2294.0
$\widehat{y}_t^*$:	–	1793.3	1853.3	1990.0	2086.7	2306.7

Év	2017	2018	2019	2020	2021	2022
y_t :	2540	2580	2740	2790	2890	3140
$\widehat{y}_t$:	2426.0	2558.0	2690.1	2822.1	2954.1	3086.2
$\widehat{y}_t^*$:	2426.7	2620.0	2703.3	2806.7	2940.0	–

Példa

y_t: a vodka éves fogyasztói átlagára (Ft) a 2011-2022 időszakban.



Kapcsolat a szezonalitással

Szezonalitás: rövid távú ingadozás, ami a determinisztikus alapmodell része. Hullámhosszát és amplitúdóját állandónak feltételezzük.

Általában éven belüli (havi, negyedéves) adatok idősoraiban jelenik meg.

p: az szezonális ingadozás hullámhossza, az egy peridóduson belüli időszakok száma.

A mozgóátlagolás akkor simítja ki megfelelően az idősort, ha annak m tagszáma a p hullámhossz egész számú többszöröse. Egyébként vagy nem simít eléggé, vagy nem létező ciklikus hatásokat visz az idősorba.

Gyakorlati alkalmazások esetén az $m = p$ választás javasolt.

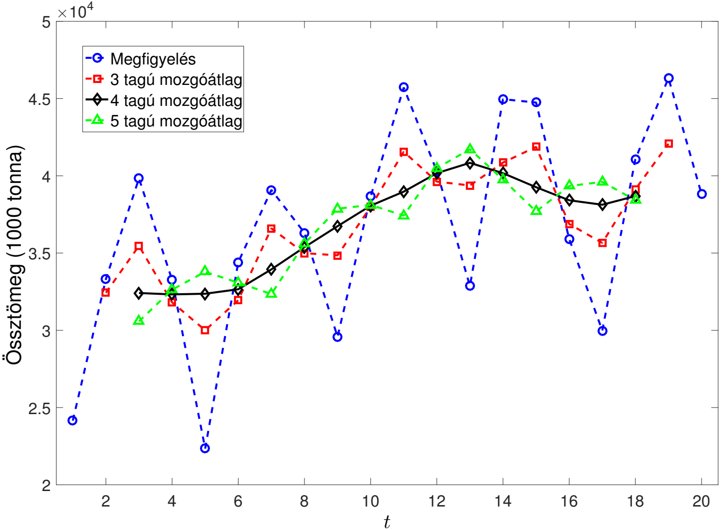
Példa

A belföldön közúton szállított áruk össztömege (1000 tonna) a 2012-2016 időszakban (forrás: KSH).

Év	negyedév	Megfigyelés	MA(3)	MA(4)	MA(5)
2012	I.	24152	-	-	-
	II.	33314	32431.7	-	-
	III.	39829	35467.7	32413.3	30580.6
	IV.	33260	31812.3	32324.0	32631.0
2013	I.	22348	30004.0	32367.6	33785.8
	II.	34404	31946.7	32656.1	33081.8
	III.	39088	36600.3	33941.8	32346.6
	IV.	36309	34993.7	35378.4	35609.2
2014	I.	29584	34851.3	36738.9	37871.4
	II.	38661	37986.7	38065.5	38112.8
	III.	45715	41557.0	38974.4	37424.8
	IV	40295	39626.3	40172.4	40500.0
2015	I.	32869	39374.7	40838.3	41716.4
	II.	44960	40857.3	40169.6	39757.0
	III.	44743	41873.7	39261.6	37694.4
	IV.	35918	36881.0	38415.4	39336.0
2016	I.	29982	35659.0	38128.0	39609.4
	II.	41077	39128.7	38687.4	38422.6
	III.	46327	42071.0	-	-
	IV.	38809	-	-	-

Példa

A belföldön közúton szállított áruk össztömege (1000 tonna) a 2012-2016 időszakban ($t = 1$: 2012 I. negyedév).



A trendszámítási módszerek összehasonlítása

Az **analitikus trend** számítása:

- feltételez egy ismert analitikus függvényt, amely jól leírja a hosszú távú tendenciát;
- magába foglalja a függvény ismeretlen paramétereinek becslését, amely paraméterek fontos jellemzői lehetnek az időszaknak;
- lehetőséget ad a trendfüggvény értékeinek meghatározására mind a megfigyelési időszakon belül, mind azon kívül, azaz lehetőséget ad trendelőrejelzések készítésére.

A **mozgóátlagolású trend** számítása során:

- nem szükséges előre rögzíteni, valószínűsíteni a növekedési pálya jellegét;
- az eljárás nem eredményez az idősor tömör jellemzésére használható paramétereket;
- elkészíthetők a megfigyelési időszak egy részére a becsült trendértékek.

A konjuktúraciklus kimutatása

Dekompozíciós modellek:

$$Y = \widehat{Y} + S + C + \varepsilon, \qquad \text{vagy} \qquad Y = \widehat{Y} \cdot S^* \cdot C^* \cdot \nu.$$

Cél: a C vagy C^* hosszú távú szabálytalan ciklus meghatározása.

Első megoldás:

1. elkészítjük a megfigyelt idősor mozgóátlagolású trendjét;
2. a mozgóátlagolású trendből analitikus trendet számolunk;
3. kiszámoljuk a mozgóátlagolású és az analitikus trend különbségét (hányadosát), ami megadja a ciklus empirikus értékeit.

Második megoldás:

1. az idősorra analitikus trendet illesztünk;
2. additív modellnél az idősor elemeiből levonjuk az analitikus trendet, multiplikatív esetben osztunk a trendértékekkel;
3. a kapott maradékokra mozgóátlagolású trendet illesztünk, ami megadja a ciklus empirikus értékeit.

Szezonális ingadozások

Dekompozíciós modellek:

$$Y = \widehat{Y} + S + C + \varepsilon, \qquad \text{vagy} \qquad Y = \widehat{Y} \cdot S^* \cdot C^* \cdot \nu.$$

Szezonális ingadozásnak $(S,\ S^*)$ a rendszeresen ismétlődő, azonos hullámhosszú (periodicitású) és szabályos amplitúdójú, többnyire rövid távú ingadozásokat tekintjük.

Jelölések:

- n: az idősor megfigyeléseinek száma;
- p: egy perióduson belüli időszakok száma;
- y_{ij}: az i-edik periódus j-edik megfigyelt eleme, ahol $i = 1, 2, \ldots, n/p$, $j = 1, 2, \ldots, p$.

Technikai feltétel: n/p egész szám, azaz a megfigyelt idősor teljes periódusokból áll.

Példa. Öt évnyi, negyedéves adatokat tartalmazó idősor esetén $p = 4$, $n = 20$; ugyanilyen hosszú időszakra havi adatokkal $p = 12$, $n = 60$.

Additív szezonalitás

Additív modell a megfigyelt idősorra:

$$y_{ij} = \widehat{y}_{ij} + S_j + e_{ij}, \qquad i = 1, 2, \ldots, n/p, \quad j = 1, 2, \ldots, p.$$

$\widehat{y}_{ij}$: a trendfüggvény becsült értéke.

e_{ij}: a véletlen komponensnek a trendszámítás után maradt értéke.

S_j: a szezonális ingadozás a j-edik szezonban, nem függ az i periódusindextől.

$$y_{ij} - \widehat{y}_{ij} = S_j + e_{ij}, \qquad \frac{\sum_{i=1}^{n/p}(y_{ij} - \widehat{y}_{ij})}{n/p} = S_j + \frac{\sum_{i=1}^{n/p} e_{ij}}{n/p}.$$

Az S_j, $j = 1, 2, \ldots, p$, nyers szezonális eltérések becslése:

$$s_j := \frac{\sum_{i=1}^{n/p}(y_{ij} - \widehat{y}_{ij})}{n/p}, \qquad j = 1, 2, \ldots, p.$$

Az s_j becsült nyers szezonális eltérés azt mutatja, hogy a megfigyelt idősor a j-edik szezonban átlagosan mennyivel tér el a trendértéktől a szabályosan ismétlődő szezonhatás következtében.

Korrigált szezonális eltérések

Additív modell a megfigyelt idősorra:

$$y_{ij} = \widehat{y}_{ij} + S_j + e_{ij}, \qquad i = 1, 2, \ldots, n/p, \quad j = 1, 2, \ldots, p.$$

Becsült nyers szezonális eltérés:

$$s_j := \frac{\sum_{i=1}^{n/p} (y_{ij} - \widehat{y}_{ij})}{n/p}, \qquad j = 1, 2, \ldots, p.$$

S_j és s_j mértékegysége megegyezik a megfigyelt idősor mértékegységével.

Természetes követelmény: $\sum_{j=1}^{p} S_j = 0$, a szezonális hatások egy perióduson belül kiegyenlítik egymást. Ez indokolja a becslések korrekcióját.

Korrigált szezonális eltérések:

$$\widetilde{s}_j := s_j - \overline{s}, \quad j = 1, 2, \ldots, p, \qquad \text{ahol} \qquad \overline{s} := \frac{1}{p} \sum_{j=1}^{p} s_j.$$

Példa

A belföldön közúton szállított áruk össztömege (1000 tonna) a 2012-2016 időszakban (forrás: KSH).

Év	negyedév	y_{ij}	$\widehat{y}_{ij}$, MA(4)	$y_{ij} - \widehat{y}_{ij}$
2012	I.	24152	-	-
	II.	33314	-	-
	III.	39829	32413.3	7415.7
	IV.	33260	32324.0	936.0
2013	I.	22348	32367.6	-10019.6
	II.	34404	32656.1	1747.9
	III.	39088	33941.8	5146.2
	IV.	36309	35378.4	930.6
2014	I.	29584	36738.9	-7154.9
	II.	38661	38065.5	595.5
	III.	45715	38974.4	6740.6
	IV	40295	40172.4	122.6
2015	I.	32869	40838.3	-7969.3
	II.	44960	40169.6	4790.4
	III.	44743	39261.6	5481.4
	IV.	35918	38415.4	-2497.4
2016	I.	29982	38128.0	-8146.0
	II.	41077	38687.4	2389.6
	III.	46327	-	-
	IV.	38809	-	-

Példa

y_{ij}: a belföldön közúton szállított áruk össztömege (1000 tonna), negyedéves adatok.

Év	$y_{ij} - \widehat{y}_{ij}$			
	I.	II.	III.	IV.
	negyedév			
2012	-	-	7415.7	936.0
2013	-10019.6	1747.9	5146.2	930.6
2014	-7154.9	595.5	6740.6	122.6
2015	-7969.3	4790.4	5481.4	-2497.4
2016	-8146.0	2389.6	-	-
Összesen	-33289.8	9523.4	24783.9	-508.2
Átlag	-8322.45	2380.85	6195.975	-127.05

Becsült nyers szezonális eltérések: $s_1 = -8322.45$, $s_2 = 2380.85$, $s_3 = 6195.975$, $s_4 = -127.05$.

Az eltérések átlaga:

$$\bar{s} = \frac{s_1 + s_2 + s_3 + s_4}{4} = \frac{-8322.45 + 2380.85 + 6195.975 - 127.05}{4} = 32.83125.$$

Korrigált szezonális eltérések $(\widetilde{s}_i = s_i - \bar{s})$:

$$\widetilde{s}_1 = -8354.28125, \quad \widetilde{s}_2 = 2349.01875, \quad \widetilde{s}_3 = 6164.14375, \quad \widetilde{s}_4 = -158.88125.$$

Multiplikatív szezonalitás

Multiplikatív modell a megfigyelt idősorra:

$$y_{ij} = \widehat{y}_{ij} \cdot S_j^* \cdot u_{ij}, \qquad i = 1, 2, \ldots, n/p, \quad j = 1, 2, \ldots, p.$$

Az S_j^*, $j = 1, 2, \ldots, p$, nyers szezonindex becslése:

$$s_j^* := \frac{\sum_{i=1}^{n/p} (y_{ij}/\widehat{y}_{ij})}{n/p}, \qquad j = 1, 2, \ldots, p.$$

Az s_j^* becsült nyers szezonindex azt fejezi ki mértékegység nélkül, illetve százalékos formában, hogy a j-edik szezonban a megfigyelt idősor átlagosan hányszorosa, illetve hány százaléka a trendértéknek a szezonhatás következtében.

Természetes követelmény: $\frac{1}{p} \sum_{j=1}^{p} S_j^* = 1$, a szezonális hatások egy perióduson belül kiegyenlítik egymást.

Korrigált szezonindex:

$$\widetilde{s}_j^* := \frac{s_j^*}{\overline{s}^*}, \quad j = 1, 2, \ldots, p, \qquad \text{ahol} \qquad \overline{s}^* := \frac{1}{p} \sum_{j=1}^{p} s_j.$$

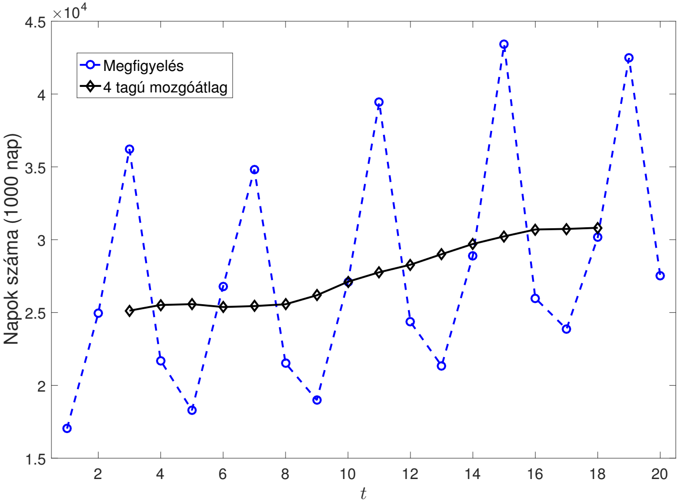
Példa

A Magyarországra tett külföldi látogatások során (a tehergépkocsi-vezetők nélkül) az országban eltöltött napok száma (1000 nap) a 2012-2016 időszakban (forrás: KSH).

Év	negyedév	y_{ij}	$\widehat{y}_{ij}$, MA(4)	$y_{ij}/\widehat{y}_{ij}$
2012	I.	17032	-	-
	II.	24938	-	-
	III.	36215	25127.9	1.4412
	IV.	21696	25517.9	0.8502
2013	I.	18293	25575.0	0.7153
	II.	26797	25377.4	1.0559
	III.	34813	25442.0	1.3683
	IV.	21517	25567.1	0.8416
2014	I.	18989	26183.0	0.7252
	II.	27102	27115.3	0.9995
	III.	39435	27763.6	1.4204
	IV	24353	28282.5	0.8611
2015	I.	21340	29005.6	0.7357
	II.	28902	29707.5	0.9729
	III.	43420	30226.6	1.4365
	IV.	25983	30699.9	0.8464
2016	I.	23863	30739.3	0.7763
	II.	30165	30815.8	0.9789
	III.	42472	-	-
	IV.	27543	-	-

Példa

A Magyarországra tett külföldi látogatások során (a tehergépkocsi-vezetők nélkül) az országban eltöltött napok száma (1000 nap) a 2012-2016 időszakban ($t = 1$: 2012 I. negyedév).



Példa

y_{ij}: a Magyarországra tett külföldi látogatások során (a tehergépkocsi-vezetők nélkül) az országban eltöltött napok száma (1000 nap), negyedéves adatok.

Év	$y_{ij}/\widehat{y}_{ij}$			
	I.	II.	III.	IV.
	negyedév			
2012	-	-	1.4412	0.8502
2013	0.7153	1.0559	1.3683	0.8416
2014	0.7252	0.9995	1.4204	0.8611
2015	0.7357	0.9729	1.4365	0.8464
2016	0.7763	0.9789	-	-
Összesen	2.9525	4.0072	5.6664	3.3992
Átlag	0.7381	1.0018	1.4166	0.8498

Becsült nyers szezonindexek: $s_1^* = 0.7381, \quad s_2^* = 1.0018, \quad s_3^* = 1.4166, \quad s_4^* = 0.8498.$

Az eltérések átlaga:

$$\overline{s}^* = \frac{s_1^* + s_2^* + s_3^* + s_4^*}{4} = \frac{0.7381 + 1.0018 + 1.4166 + 0.8498}{4} = 1.0016.$$

Korrigált szezonindexek $(\widetilde{s}_i^* = s_i^*/\overline{s}^*)$:

$$\widetilde{s}_1^* = 0.7370, \quad \widetilde{s}_2^* = 1.0002, \quad \widetilde{s}_3^* = 1.4144, \quad \widetilde{s}_4^* = 0.8485.$$

Additív vs. multiplikatív szezonalitás

Szezonális modellek a megfigyelt idősorra:

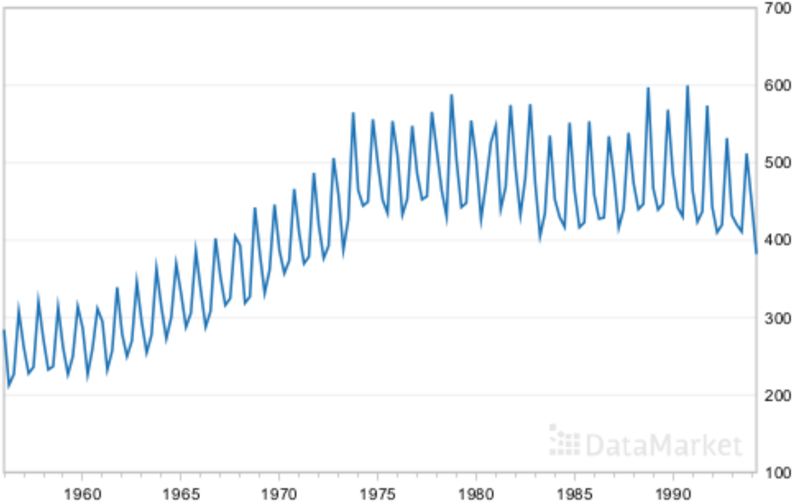
$$
\begin{array}{lll}
\text{additív:} & y_{ij} = \widehat{y}_{ij} + S_j + e_{ij}, & i = 1, 2, \ldots, n/p, \quad j = 1, 2, \ldots, p; \\
\text{multiplikatív:} & y_{ij} = \widehat{y}_{ij} \cdot S_j^* \cdot u_{ij}, & i = 1, 2, \ldots, n/p, \quad j = 1, 2, \ldots, p.
\end{array}
$$

Az additív és multiplikatív modell közti választás a szezonalitás jellege alapján történik.

- **Additív** modellt használunk, ha minden egyes periódusban azonos mértékű a kilengések nagysága, függetlenül a trend értékétől.
- **Multiplikatív** modellt használunk, ha a szezonális kilengések nagysága a trendértékkel arányosan változik, magasabb szinten nagyobbak, alacsonyabb szinten kisebbek az ingadozások.

Példa additív szezonalitásra

Ausztrália negyedéves sörtermelése (megaliter) 1956. március és 1995. június között (forrás: Time Series Data Library; Australian Bureau of Statistics).



Példa multiplikatív szezonalitásra

Ausztrália havi áramtermelése (millió kWh) 1956. január és 1995. augusztus között (forrás: Time Series Data Library; Australian Bureau of Statistics).