Statisztika gyakorlat

Geológus szakirány

Játékszabályok

- Az órákon részt kell venni, maximum 3-szor lehet hiányozni.
- Az aláírás megszerzésének lehetséges módjai:
  - vagy ZH írásával
  - vagy egy előre kihirdetett házi feladat beadásával
- A ZH-ról bővebben:
  - utolsó gyakorlaton lesz
  - legalább 50%-ot kell elérni
  - A4-es lapra KÉZZEL írott "puska" és számológép használható
  - pótlási lehetőség: vizsgaidőszak 1. hetén
- A házi feladatról bővebben:
  - A statisztikai számításokat R-ben (esetleg Excelben) kell elkészíteni; az elemzéseket, értelmezéseket pedig Word-ben (vagy Latexben) kell értelmes, kerek magyar mondatokban leírni.
  - Védés: előre egyeztetett időpontban kötetlen, körülbelül 15 perces beszélgetés. Az értintett hallgatókkal az időpontot az utolsó gyakorlat előtt/után egyeztetjük.
  - A házi feladat **önálló** munka legyen! Amennyiben bebizonyosodik, hogy a házi feladatot nem Te írtad, vagy nagyon hasonlít (mondatok, bekezdések azonosak, ugyanazokat a rossz következtetéseket vonod le, ugyanazokat számolod ki rosszul) egy másik hallgatótársadéra, akkor pótZH-t kell írnod.
- R-hez ajánlott szoftver: RStudio

Infók a gyakvezetőről

Név	Varga László
Tanszék	Valószínűségelméleti és Statisztika Tanszék (ELTE TTK)
Szoba	D 3-309
E-mail	`vargal4@cs.elte.hu`
Honlap	`www.cs.elte.hu/~vargal4`

Ajánlott irodalom

- Solymosi Norbert: Bevezetés az R-nyelv és környezet használatába; elérési hely: `http://cran.r-project.org/doc/contrib/Solymosi-Rjegyzet.pdf`
- Móri-Szeidl-Zempléni: Matematikai statisztikai feladatok
- Pröhle-Zempléni: Többdimenziós statisztika számítógépes módszerei; elérési hely: `http://www.cs.elte.hu/~zempleni/tobbdim_stat.pdf`

1.) Legyen $X \sim N(1, 2^2)$. Számítsuk ki a $P(-X + 1 < 3)$ mennyiséget! Becsüljük szimulációval is!

2.) Legyenek $X_1, \dots, X_n$ független

a.) $N(10, 3^2)$ eloszlásúak,

b.) $\mathrm{Geo}\left(\frac{1}{10}\right)$ eloszlásúak.

Vizsgáljuk meg számítógépes szimulációval, hová tart a $\frac{X_1+\dots+X_n}{n}$ mennyiség, ha minél nagyobb n-re számítjuk ki ezt az átlagot!

3.) Legyen $X \sim N(-3, 4^2)$. Számítsuk ki a $P(2X + 2 > -6)$ mennyiséget! Becsüljük szimulációval is!

4.) Legyenek $X_i \sim N(10, 5^2)$ $(i = 1, \dots, 9)$ függetlenek. Számítsuk ki a $P(\overline{X} < 9)$ mennyiséget! Becsüljük szimulációval is!

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

Fontos speciális kvantilisek: kvartilisek:

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

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

Értelmezése: ha

- $\mathrm{skew}(X)=0$, akkor az eloszlás szimmetrikus;
- $\mathrm{skew}(X)>0$, akkor az eloszlás balra ferdült;
- $\mathrm{skew}(X)<0$, akkor az eloszlás jobbra ferdült.

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

Értelmezése: ha

- $\mathrm{kurt}(X)=0$, akkor az eloszlás csúcsossága a standard normáliséval megegyező;
- $\mathrm{kurt}(X)<0$, akkor az eloszlás laposabb a standard normálisnál;
- $\mathrm{kurt}(X)>0$, akkor az eloszlás csúcsosabb a standard normálisnál.

Minta: $X_1, ..., X_n$ valószínűségi változó sorozat. (Jel. $\mathbf{X} = X_1, ..., X_n$)
A továbbiakban feltesszük, hogy függetlenek és azonos eloszlásúak. Magyarosan rövidítve FAE minta, de gyakrabban használják az angol *i.i.d. minta* rövidítést (independent, identically distributed).
Az elméleti értékeket nagy, a konkrét, realizált mintából számolt értékeket mindig kis betű fogja jelölni, azaz minta esetén $x_1, ..., x_n$.

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

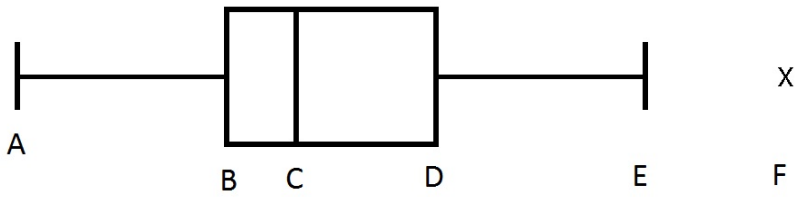
Néhány lényeges statisztika:

- *Rendezett minta*: $X_1^* \leq ... \leq X_n^*$ nem csökkenő sorrendbe tesszük a mintaelemeket
- *Terjedelem*: $R = X_n^* - X_1^*$ (R=range)
- *Mintaátlag*: $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
- *Tapasztalati szórás*: $S_n = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}}$
  Értelmezése: az átlagtól való átlagos eltérés abszolút mértékegységben
- *Korrigált tapasztalati szórás*: $S_n^* = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$
- *Szórási együttható*: $V = \frac{S_n}{\overline{X}}$
  Értelmezése: az átlagtól való átlagos eltérés százalékban
  Megj.: relatív szórásnak is hívják
- *Tapasztalati eloszlásfüggvény*: $F_n(x) = \frac{\sum_{i=1}^{n} I(X_i < x)}{n}$
  ahol $I(X_i < x) = \begin{cases} 1 & \text{ha } X_i < x \\ 0 & \text{ha } X_i \geq x \end{cases}$ $\rightsquigarrow$ karakterisztikus függvény
- *Tapasztalati z-kvantilis*: Realizált mintából sokféleképpen számolható, interpolációs módszer:
  1.) Sorszám megállapítása: $(n+1)z = e + t$ (e:egészrész, t:törtrész)
  2.) $q_z = x_e^* + t(x_{e+1}^* - x_e^*)$
  Értelmezése: a mintaelemek z-ed része q_z-nél kisebb, $(1-z)$-ed része q_z-nél nagyobb
- *Interkvantilis terjedelem*: $IQR = Q_3 - Q_1$
- *Tapasztalati módusz*: a legtöbbször előforduló értékek közül a legkisebb.
  Értelmezése: a minta tipikus, leggyakrabban előforduló értéke.
- *Tapasztalati ferdeség*: $\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{S_n^3}$
- *Tapasztalati csúcsosság*: $\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{S_n^4} - 3$

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

Osztályközös gyakorisági sor készítése: jelölje n a minta elemszámát. Az osztályközök meghatározása nem egyértelmű, általános hüvelykujjszabályként az osztályok k száma legyen $k = \min\{k : 2^k > n\}$. Ha azonos hosszúságú (h) osztályközöket akarunk létrehozni, akkor $h = \frac{x_{\max} - x_{\min}}{k}$.

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



- $A = \max\{x_1^*, Q_1 - 1,5 \cdot IQR\}$;
- $B = Q_1$;
- $C = Me$;
- $D = Q_3$;
- $E = \min\{x_n^*, Q_3 + 1,5 \cdot IQR\}$;
- F: kieső értékek, azokat tüntetjük fel pontokként, amik A-n vagy E-n kívülre esnek.

Feladatok

5.) Egy osztályban a diákok magassága (cm):

180	163	1500
157	165	165
174	191	172
165	1-68	186

a.) Nézzük át nagy vonalakban az adatokat, reálisak-e! Próbáljuk javítani az esetleges adathibákat!

b.) Rajzold fel a tapasztalati eloszlásfüggvényt! Mennyi a tapasztalati eloszlásfüggvény értéke a 180 helyen?

c.) Elemezd a diákok testmagasságát
- átlag;
- korrigált tapasztalati szórás;
- szórási együttható;
- kvartilisek;
- terjedelem;
- interkvartilis terjedelem;
- tapasztalati ferdeség;
- tapasztalati csúcsosság segítségével!

Értelmezd is az eredményeket!

d.) Készíts boxplot ábrát!

e.) Készíts alkalmas osztályközös gyakorisági sort, majd abból hisztogramot!

6.) 2013 nyarán az alábbi maximum hőmérsékleteket mérték egy településen (°C):

június	25	25	28	29	26	23	21	22	25	25
	25	25	29	31	31	32	33	30	28	25
	26	24	22	21	25	29	33	31	32	33
július	34	32	32	35	36	32	31	32	35	35
	32	29	28	25	27	27	28	30	28	27
	29	32	32	34	35	31	33	31	30	30
	30									
augusztus	30	31	32	33	35	32	31	32	28	27
	25	27	28	31	30	30	32	30	28	28
	27	25	28	29	26	22	23	21	24	23
	25									

a.) Készíts számítógép segítségével tapasztalati eloszlásfüggvényt!

b.) Elemezd együtt a nyári maximális hőmérséklet értékeket
- átlag;
- korrigált tapasztalati szórás;
- szórási együttható;
- kvartilisek;
- terjedelem;
- interkvartilis terjedelem;
- tapasztalati ferdeség;
- tapasztalati csúcsosság segítségével!

Értelmezd is az eredményeket!

c.) Készíts boxplot ábrát!

d.) Készíts alkalmas osztályközös gyakorisági sort, majd abból hisztogramot!

Definíció. **Torzítatlan becslés:** $\mathrm{T}(\mathbf{X})$ statisztika torzítatlan becslése θ-nak, ha $E_\theta T(\mathbf{X}) = \theta \quad \forall\theta$-ra.

Definíció. Legyenek $T_1(X)$ és $T_2(X)$ torzítatlan becslései θ-nak. Ekkor azt mondjuk, hogy $T_1(X)$ **hatásosabb** $T_2(X)$-nél, ha $D^2_\theta(T_1(X)) \leq D^2_\theta(T_2(X))$ minden $\theta \in \Theta$ esetén.

Definíció. **Hatásos becslés.** A $T(X)$ tozítatlan becslést hatásosnak nevezzük, ha minden torzítatlan becslésnél hatásosabb.

Definíció. **Konzisztencia:** A $T_n(X)$ becléssorozat $(n = 1, 2, ...)$ konzisztens becslése a θ paraméternek, ha $T_n(X)$ sztochasztikusan a θ paraméterhez tart $\forall\theta$ esetén.

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

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

Definíció. **Log-likelihood függvény:** $l(\theta, \mathbf{x}) = \log(L(\theta, \mathbf{x}))$.

Paraméterbecslési módszerek

- **Maximum likelihood módszer (ML-módszer):** Azt a paraméterértéket keressük, ahol a likelihood függvény a legnagyobb értéket veszi fel: $\max_\theta L(\theta, \mathbf{x})$

  Amennyiben a függvény deriválható θ szerint, akkor a maximumot kereshetjük a szokásos módon, az első és második deriváltak segítségével, azonban a feladatunkat jelentősen megnehezíti, hogy olyan n-szeres

szorzatot kellene deriválni, amelyiknek minden tagjában ott van az a változó, ami szerint deriválnunk kellene. Ezért likelihood függvény helyett a log-likelihood függvény maximumhelyét keressük.

Ha θ 1 dimenziós, akkor az

· elsőrendű feltétel: $\partial_\theta l(\theta, \mathbf{x}) = 0 \quad \rightsquigarrow \quad \hat{\theta}$

· másodrendű feltétel: $\partial^2_\theta l(\theta, \mathbf{x}) < 0$

Ha θ p dimenziós, akkor $\theta = (\theta_1, ..., \theta_p)$, az

· elsőrendű feltétel:
$\partial_{\theta_i} l(\theta, \mathbf{x}) = 0 \quad \rightsquigarrow \quad \hat{\theta}_i \ \ (i = 1, ..., p) \quad \rightsquigarrow \quad \hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)$

· másodrendű feltétel: $H(\theta_1, ..., \theta_p) = \left(\partial_{\theta_i}\partial_{\theta_j} l(\theta, \mathbf{x})\right)_{i,j=1,...,p}$ Hesse-mátrix negatív definit a $\theta = \hat{\theta}$ helyen

- **Momentum módszer:** A mintából számítható tapasztalati momentumokat ($m_i := \frac{\sum_j x_j^i}{n}$) egyenlővé tesszük az elméleti momentumokkal ($M_i := E_\theta X^i$), az elsőtől kezdve, mégpedig annyit, amennyi paraméter van. Tehát p darab ismeretlen paraméter esetén a következő p ismeretlenes egyenletrendszert oldjuk meg:

$$M_1 = m_1$$
$$\vdots$$
$$M_p = m_p$$

Megjegyzés: $m_1 = \overline{x}$

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

Feladatok

7.) Legyen $X_1, ..., X_n$ független, azonos abszolút folytonos eloszlású valószínűségi változók sorozata.
Adjuk meg $\min(X_1, ..., X_n)$, illetve $\max(X_1, ..., X_n)$ eloszlás- és sűrűségfüggvényét! A minimumnál külön is vizsgáljuk meg azt az esetet, ha az X_i változók exponenciális eloszlásúak!

8.) Adjunk torzítatlan becslést a val.szám. vizsga bukási arányára, ha 300-ból 100-an buktak meg. Mekkora a becslésünk szórása? (Adjunk rá felső becslést.)

9.) Legyen $X_1, ..., X_n$ i.i.d. minta ismeretlen eloszlásból.

a.) Torzítatlan becslés-e a várható értékre nézve az átlag?

b.) Torzítatlan becslés-e a szórásnégyzetre nézve a tapasztalati szórásnégyzet?

Amennyiben nem az, hogyan tudnánk torzítatlanná tenni?

10.) n-elemű λ-paraméterű exponenciális minta esetén adjunk torzítatlan becslést $e^{-3\lambda}$-ra és $\frac{1}{\lambda}$-ra! Vizsgáljuk meg szimulációval is!

11.) n-elemű λ-paraméterű Poisson minta esetén adjunk torzítatlan becslést $e^{-\lambda}$-ra és λ^2-re! Vizsgáljuk meg szimulációval is!

12.) Adjunk meg torzítatlan becslést a [0,θ] intervallumon egyenletes eloszlás paraméterére

a.) a mintaátlag

b.) a maximum

segítségével. Melyik a hatásosabb a kettő közül? Konzisztens-e a két becslés? Vizsgáljuk meg szimulációval is!

13.) Mutassuk meg, hogy exponenciális eloszlású minta esetén $T(\mathbf{X}) = n \cdot \min(X_1, ..., X_n)$ statisztika torzítatlan a várható értékre. Mekkora a szórása? Konzisztens a becslés?

14.) Legyen $X_1, ..., X_n$ i.i.d. minta valamely véges szórású eloszlásból, és tekintsük a $\mathrm{T}(\mathbf{X}) = a_1X_1 + ... + a_nX_n$ alakú lineáris becsléseket, ahol $a_1, ..., a_n \in \mathbb{R}$. Feltéve, hogy $\mathrm{T}(\mathbf{X})$ a várható érték torzítatlan becslése, mely $a_1, ..., a_n$ számokra lesz minimális a $D^2(T(\mathbf{X}))$?

15.) Határozzuk meg az ismeretlen paraméter(ek) ML becslését, ha a minta

a.) Pascal (=Geom(p));

b.) Bin(m, p), ahol m ismert, p paraméter;

c.) E(a, b) eloszlású, ahol $a < b$, mindkettő paraméter;

d.) Exp(λ);

e.) Poi(λ).

16.) Tegyük fel, hogy a minta kétparaméteres eloszláscsaládból származik, a paraméterek a és b.

Ekkor mutassuk meg, hogy az $\begin{cases} E_{a,b}X = m_1 \\ E_{a,b}X^2 = m_2 \end{cases}$ egyenletrendszer megoldása megegyezik az $\begin{cases} E_{a,b}X = m_1 \\ D^2_{a,b}X = s_n^2 \end{cases}$ egyenletrendszer megoldásával.

17.) Becsüld a paramétert momentum-módszerrel az alábbi esetekben:

a.) Exp(λ);

b.) Poi(λ);

c.) E(a, b);

d.) E($-a, a$).

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

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

Mostantól α egy 0-hoz közeli pozitív szám lesz (például $0.05 = 5\%$), és vezessük be a következő jelöléseket:

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

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

Paramétertér: $\Theta = \Theta_0 \cup^* \Theta_1 \longrightarrow$ "valóság"
Mintatér: $\mathcal{X} = \mathcal{X}_e \cup^* \mathcal{X}_k \longrightarrow$ "látszat" - MINTÁBÓL
$\mathcal{X}_k$: kritikus tartomány - azon $\mathbf{X}$ megfigyelések halmaza, amikre *elutasítjuk* a nullhipotézist
$\mathcal{X}_e$: elfogadási tartomány - azon $\mathbf{X}$ megfigyelések halmaza, amikre *elfogadjuk* a nullhipotézist

Hipotézisvizsgálati feladat:
$H_0: \ \vartheta \in \Theta_0 \quad \rightsquigarrow$ nullhipotézis
$\underline{H_1: \ \vartheta \in \Theta_1} \quad \rightsquigarrow$ ellenhipotézis

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

Az $\mathcal{X}$ mintatér felosztását általában egy statisztika (neve: próbastatisztika) segítségével végezzük el:
legyen T: $\mathcal{X} \to \mathbb{R}$, $\mathcal{X}_k = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) > c\}$ c neve: kritikus érték
$\mathcal{X}_e = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) \le c\}$

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

P(elsőfajú hiba)=$\alpha(\vartheta)$=$\mathrm{P}_\vartheta(\mathcal{X}_k)$, ahol $\vartheta \in \Theta_0$
P(másodfajú hiba)=$\beta(\vartheta)$=$\mathrm{P}_\vartheta(\mathcal{X}_e)$, ahol $\vartheta \in \Theta_1$
Erőfüggvény: ψ: $\Theta_1 \to \mathbb{R}$, $\psi(\vartheta) = P_\vartheta(\mathcal{X}_k)$
Terjedelem: $\alpha = \sup\{\alpha(\vartheta): \vartheta \in \Theta_0\}$
p-érték: az az α terjedelem, ami esetén a próbastatisztika értéke egyenlő a kritikus értékkel : $\mathrm{T}(\mathbf{x}) = c_\alpha$.
A p-érték a legkisebb terjedelem, amire még elutasítjuk a H_0-t. Ha egy próbát számítógép segítségével végzünk el, rendszerint a p-érték révén tudunk dönteni: ha (p-érték)$< \alpha$, akkor elvetjük H_0-t.

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

1.) Egymintás próbák

a.) Egymintás u-próba

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

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

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

b.) Egymintás t-próba

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

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

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

2.) Kétmintás próbák

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

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

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

a.) F-próba

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

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

A próbastatisztika: $F = \begin{cases} \frac{(s_1^*)^2}{(s_2^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{n-1,m-1} & \text{ha } s_1^* > s_2^* \\ \frac{(s_2^*)^2}{(s_1^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{m-1,n-1} & \text{ha } s_2^* > s_1^* \end{cases}$

b.) kétmintás u-próba

m_1, m_2 paraméterek, σ_1, σ_2 ismert

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

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

c.) kétmintás t-próba

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

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

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

d.) Welch-próba

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

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

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

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

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

Feladatok

18.) Valaki azt állítja, hogy a klíma változik, és ezt azzal véli bizonyítottnak, hogy az elmúlt 10 évben 2-szer is volt jégeső, pedig korábban az egyes évekre a jégeső valószínűsége a hivatalos adatok alapján csupán p=0.1 volt. Írjuk fel a hipotéziseket, a próbát és állapítsuk meg az elsőfajú hiba valószínűségét, valamint az erőfüggvényt a p=0.2 pontban!

19.) Az alábbi minta 4 év október 18-án Budapesten mért napi középhőmérséklet adatait tartalmazza. Ellenőrizzük a H_0: m =15 hipotézist α =0.05 elsőfajú hibavalószínűség mellett *értelmes* alternatív hipotézissel szemben.

Középhőm. (C fok) adatok:

14,8	12,2	16,8	11,1

a.) A korábbi tapasztalatok alapján tekintsük az értékek szórását 2-nek. Adjuk meg a p-értéket is.

b.) Ne használjunk a szórásra vonatkozóan előzetes információt.

20.) Tegyük fel, hogy az emberi magasság normális eloszlású.

a.) Végezzünk statisztikai próbát arra vonatkozóan, hogy a gyakorlaton lévő lányok átlagmagassága 170 cm!

b.) Végezzünk statisztikai próbát arra vonatkozóan, hogy a gyakorlaton lévő fiúk átlagmagassága 180 cm!

21.) A Dezinformatikai Kar III. évfolyamán 10-en írtak statisztika zárthelyit. 2 feladatsor volt, mindkettőben 30 pontot lehetett elérni. Tegyük fel, hogy az elért pontszámok normális eloszlásúak. A pontszámokat tartalmazza az alábbi táblázat:

1. feladatsor	12	11	8	14	10
2. feladatsor	15	14	9	16	11

a.) Vajon az első feladatsor nehezebb volt?

b.) Mennyiben változik a helyzet, ha nem 10 diákról, hanem csak 5-ről van szó, és a 2. feladatsor a pótZH eredménye?

22.) Tegyük fel, hogy az emberi magasság normális eloszlású. Végezzünk statisztikai próbát arra vonatkozóan, hogy a gyakorlaton lévő fiúk magasabbake a lányoknál!

23.) Az alábbi két minta 10 egyforma képességűnek feltételezett sportoló súlylökésben elért eredményeit tartalmazza. A sportolók két ötfős csoportban készültek az edzőtáborban. Edzéstervük ugyanaz volt, de az első csoportban készülők minden reggel fejenként 10 tojást és 25 túró rudit ettek meg. A második csoportban készülőknek reggel és este 1-1 kg szalonnát és 1-1 kg madártejet kellett megenni. 2 hét felkészülés után értékelték az eredményeket. Tételezzük fel, hogy normális eloszlásból származnak a minták és a terjedelem 5%.

1. csoport	15,8	15,2	16,3	17,1	16,1
2. csoport	19,0	12,1	17,2	14,7	21,0

a.) Melyik diéta volt jobb, ha a dobások szórását 2-nek tekintjük?
b.) Állíthatjuk-e, hogy a második csoportban nagyobb változékonyságot mutat a sportolók teljesítménye?
c.) Ha nem ismerjük a szórást, akkor tekinthetjük-e valamelyik diétát jobbnak?

χ^2-próbák

a.) Diszkrét illeszkedésvizsgálat

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

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

H_0 : a valószínűségek: $\mathbf{p}$=$(p_1, \ldots, p_r)$
H_1: nem ezek a valószínűségek

A próbastatisztika: $T_n = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ eloszlásban, ha $n \to \infty$
A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

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

Illeszkedésvizsgálat "szemmel": **Q-Q plot** és **P-P plot**
Jelölje F az illesztett eloszlás eloszlásfüggvényét, x_k^* pedig a k. rendezett mintaelemet.
Q-Q plot: az illesztett eloszlás kvantiliseit vetjük össze a tapasztalati kvantilisekkel, azaz a következő pontokat ábrázoljuk: $\left(F^{-1}\left(\frac{k}{n+1}\right), x_k^*\right)$, ahol $k = 1, ..., n$.
P-P plot: az illesztett eloszlás valószínűségeit vetjük össze a tapasztalati valószínűségekkel, azaz a következő pontokat ábrázoljuk: $\left(\frac{k}{n+1}, F(x_k^*)\right)$, ahol $k = 1, ..., n$.
Mindkét ábránál be szokták húzni a 45 fokos egyenest és minél jobban rásimulnak a pontok az egyenesre, annál jobbnak tekinthető az illeszkedés.

24.) Rendelkezésünkre áll a következő minta: 0,55; 0,59; 0,34; 0,69; 0,95; 0,34; 0,53; 0,54; 0,03; 0,11; 0,15; 0,67; 0,48; 0,09; 0,55; 0,02; 0,37; 0,76; 0,83; 0,92. A megoldás során alkalmazzunk diszkretizálást, azaz képezzünk alkalmas gyakorisági sort az adatokból.
a.) Elfogadhatjuk-e azt a hipotézist, hogy a minta (0,2) intervallumon egyenletes eloszlású? Vizsgáljuk meg Q-Q plot-tal is!
b.) Elfogadhatjuk-e azt a hipotézist, hogy a minta egyenletes eloszlású? Vizsgáljuk meg Q-Q plot-tal is!
c.) Elfogadhatjuk-e azt a hipotézist, hogy a minta exponenciális eloszlású? Vizsgáljuk meg Q-Q plot-tal is!

25.) Az Informatikai Kar III. évfolyamán 300-an tanulnak. Megszámolták, hogy a legutóbbi vizsgaidőszakban hányszor buktak az egyes hallgatók. Az eredményeket tartalmazza az alábbi táblázat.

Bukások száma	0	1	2	3	4
Hallgatók száma	80	113	77	27	3

a.) Elfogadhatjuk-e azt a hipotézist, hogy egy hallgató bukásszáma Bin(4; 0,25) eloszlású?
b.) és azt, hogy Bin(4;p) eloszlású?

26.) A "Reggeli ital" tejgyárban minden szállítás előtt megvizsgálják a 25 dkg-os túrókban található hajszálak számát. Több éves tapasztalat szerint egy csomagban nincs 2 hajszálnál több. A H_0 hipotézis (a minőség elfogadható) szerint egy csomagban 1/2 valószínűséggel nincs hajszál, 1/3 valószínűséggel 1 hajszál van és 1/6 valószínűséggel 2 hajszál esett bele. A túró minőségét 2011. április 7-én 100 csomag túró tételes ellenőrzésével tesztelték. 40 csomagban nem volt hajszál, 40-ben egy hajszál volt és 20-ban 2 hajszál. Elfogadjuk-e a megfelelőség hipotézisét?

27.) CASCO biztosítással rendelkezők éves kárszámát vizsgáltuk. 4000 vezető adatait az alábbi táblázat tartalmazza. Vajon elfogadható-e 1%-os terjedelem mellett, hogy a kárszám Poisson eloszlású?

Kárszám	0	1	2	3	4	5	>5
Vezetők száma	3691	232	68	5	3	1	0

b.) Függetlenségvizsgálat

Feladat: van egy minta, két szempont szerint csoportosítva. Azt kell eldön-

teni, hogy a két szempont független-e egymástól.
$p_{i,j}$=P(egy megfigyelés az (i,j) osztályba kerül)
$N_{i,j}$=ennyi megfigyelés kerül az (i,j) osztályba

A mintavétel eredménye:

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

$N_{i\bullet} = \sum_{j=1}^{s} N_{i,j}$ $\qquad$ $N_{\bullet j} = \sum_{i=1}^{r} N_{i,j}$

H_0 : a szempontok függetlenek, azaz $p_{i,j} = p_{i\bullet} \cdot p_{\bullet j}$ $\forall i,j$-re
H_1: nem azok

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

28.) Az alábbi kontingencia-táblázat mutatja, hogy 100 évben a csapadék mennyisége és az átlaghőmérséklet hogyan alakult.

Csapadék / Hőmérséklet	Kevés	Átlagos	Sok
Hűvös	15	10	5
Átlagos	10	10	20
Meleg	5	20	5

(A cellákban az egyes esetek gyakoriságai találhatóak.) Tekinthető-e a csapadékmennyiség és a hőmérséklet függetlennek?

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

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

Feladat (**lineáris regresszió**): Adottak $(x_1,y_1),\ldots,(x_n,y_n)$ pontok, ezekre szeretnénk egyenest illeszteni (neve: *regressziós egyenes*) legkisebb négyzetek módszerével.
A modell: $Y_i = aX_i + b + \varepsilon_i$, ahol $\mathrm{E}\varepsilon_i = 0$ és $\mathrm{D}^2\varepsilon_i = \sigma^2 < \infty$ $(i = 1,\ldots,n)$
Megoldás: $\hat{a} = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$, $\hat{b} = \overline{y} - \hat{a}\overline{x}$
Reziduumok: $\hat{\varepsilon}_i = y_i - \hat{a}x_i - \hat{b}$ (i=1,..., n)
Reziduális négyzetösszeg: RNÖ=$\sum \hat{\varepsilon}_i^2 = \sum(y_i-\overline{y})^2 - \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$
$\hat{\sigma}^2 = \frac{\text{RNÖ}}{n-2}$
Tapasztalati korrelációs együttható: $R = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2 \cdot \sum(y_i-\overline{y})^2}}$. Ennek négyzetét, R^2-et *determinációs együtthatónak* hívjuk, és ezzel mérjük a modell jóságát. Az R^2 mutatja meg, hogy százalékban a modell az Y változékonyságából mennyit magyaráz meg. Értéke 0 és 1 között lehet, ha 0-hoz közeli, akkor a modell gyengén teljesít, ha 1-hez, akkor jól.

29.) Legyenek adottak a következő (x,y) párok:

x_i	0	1	6	5	3
y_i	4	3	0	1	2

a.) Határozzuk meg és ábrázoljuk is az $aX+b$ alakú regessziós egyenest!
b.) Számoljuk ki a reziduálisokat és becsüljük meg a hiba-szórásnégyzetet!
c.) Mennyire jó a modell?
d.) Adjunk előrejelzést x=10-re a regressziós egyenes alapján!
e.) Oldjuk meg a feladatot R segítségével!

30.) A Statisztika II. vizsga után kiválasztottunk 8 hallgatót, akiktől megkérdeztük, mennyi órát készültek a vizsgára és hány pontot szereztek a tantárgy előfeltételének számító Statisztika I. tantárgyból a vizsgán:

Statisztika II. pontszám	49	55	56	62	65	70	78	92
Hány órát készült a vizsgára (ó)	15	16	14	13	12	19	21	24
Statisztika I. pontszám	60	50	66	53	67	76	88	87

a.) Vizsgáljuk meg lineáris regresszióval a tanulási idő hatását a Statisztika II. pontszámra! Ábrázoljuk a regressziós egyenest!
b.) Illesszünk négyzetes regressziós függvényt a Statisztika II. pontszámra, ha a magyarázó változó a tanulási idő! Ábrázoljuk a regressziós egyenest!
c.) Illesszünk lineáris regressziót a Statisztika II. pontszámára, ha a magyarázó változók a tanulási idő és a Statisztika I. pontszám!
d.) Vessük össze a modelleket!