Valószínűségszámítás és statisztika

oktatási segédanyag

Kupán Pál

Tartalomjegyzék

1. FEJEZET

Valószínűségszámítási alapfogalmak

1.1. Események

Modell: determinisztikus vagy sztochasztikus.

Kísérlet eredménye=elemi esemény, jel. $\omega_1, \omega_2, \ldots$ Az elemi események halmazát jel. $\Omega = \{\omega_1, \omega_2, \ldots\} = \{\omega_i\}, i \in I$.

Események $\mathcal{P}(\Omega)$ -ban:

- biztos esemény=mindig bekövetkezik - azonosítjuk Ω-val.
- lehetetlen esemény=soha nem következik be - jel. $\emptyset$
- $A \in \mathcal{P}(\Omega)$ ellentét eseménye=akkor következik be mikor A nem - jel. $\overline{A}$.

1. Példa.

(1) Kocka dobás: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_6\}$ ahol ω_i=" a kocka i pontot mutat" elemi esemény. A ="páros szám"$=\{\omega_2, \omega_4, \omega_6\} \Rightarrow \overline{A} = \textit{páratlanszám} = \{\omega_1, \omega_3, \omega_5\}$. Az egyszerűség kedvéért az elemi eseményeket $\{1, 2, \ldots, 6\} = \Omega$ alakban is felírhatjuk.

(2) Két kocka dobása: a kockákat különbözőnek tekintjük (például piros, illetve kék színűnek) tehát az elemi események rendezett párost alkotnak:

$$\Omega: \begin{matrix} (1,1) & (1,2) & \ldots & (1,6) \\ (2,1) & (2,2) & \ldots & (2,6) \end{matrix}$$

$$\begin{matrix} (6,1) & (6,2) & \ldots & (6,6). \end{matrix}$$

A ="a két kocka azonos pontszámot mutat"$=\{(1,1), (2,2), \ldots, (6,6)\}$.

B ="a kockák „összege" ≥ 10"$=\{(4,6), (6,4), (5,5), (5,6), (6,5), (6,6)\}$.

(3) Pénzérme dobása: $\Omega = \{"fej", "\textit{írás}"\}$.

(4) Két érme dobása: $\Omega = \{"fej - fej", "fej - \textit{írás}", "\textit{írás} - fej", "\textit{írás} - \textit{írás}"\}$.
A ="az érmék azonos jelt mutatnak"$=\{"fej - fej", "\textit{írás} - \textit{írás}"\}$.

1.1.1. Műveletek eseményekkel $A, B \in \mathcal{P}(\Omega)$.

- események összeadása: $A + B = A \cup B$ - az az esemény amely akkor következik be ha az A vagy a B bekövetkezik.
- események szorzása: $A \cdot B = A \cap B$ - az az esemény amely akkor következik be ha az A és a B bekövetkezik.
- ellentétes esemény: $\overline{A}$ akkor következik be ha A nem következik be.
- események különbsége $A - B = A \cap \overline{B}$.

2. Tétel. *A $P(\Omega)$ halmaz a $\cup, \cap, \bar{}$, műveletre nézve boole algebrát képez, vagyis az alábbi tulajdonságok érvényesülnek:*

- *asszociativitás:* $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$
- *kommutativitás:* $A \cup B = B \cup A$, $A \cap B = B \cap A$
- *elnyelési tulajdonság:* $A \cup (A \cap B) = A$, $A \cap (A \cup B) = A$
- *disztributivitás:* $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$,
- *komplementer képzés:* $A \cup \overline{A} = \Omega$, $A \cap \overline{A} = \emptyset$.

3. Definíció. A $(P(\Omega), \cup, \cap, \bar{}, \Omega, \emptyset)$ struktúrát eseménytérnek nevezzük.

Az említett műveletekre az alábbi tulajdonságok is érvényesülnek:

- de Morgan azonosságok: $\overline{A \cup B} = \overline{A} \cap \overline{B}$, $\overline{A} \cap \overline{B} = \overline{A} \cup \overline{B}$,
- idempotencia: $A \cup A = A$, $A \cap A = A$,
- $\overline{\Omega} = \emptyset$, $\overline{\emptyset} = \Omega$,
- $\overline{\overline{A}} = A$.

4. Definíció. Az $A, B \in \mathcal{P}(\Omega)$ eseményeket egymást kizárónak nevezzük ha egy-időben nem következhetnek be: $A \cap B = \emptyset$.

5. Definíció. Az $\{A_1, A_2, \ldots\}$ rendszer az $A \in \mathcal{P}(\Omega)$ esemény egy felbontását (particióját) képezi ha

$$A_i \cap A_j = \emptyset,\ i \neq j,$$
$$\bigcup_{i \in I} A_i = A.$$

Ha a Ω tér felbontását végeztük akkor azt mondjuk, hogy $\{A_1, A_2, \ldots\}$ egy teljes esemény-rendszer.

6. Definíció. Az $\mathcal{A} \in \mathcal{P}(\Omega)$ algebrát alkot ha

$$A \in \mathcal{A} \Rightarrow \overline{A} \in \mathcal{A}$$
$$A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}.$$

Az $(\Omega, \mathcal{A})$ párost esemény-mezőnek nevezzük (véges vagy végtelen).

1.2. A valószínűség fogalma

Feltételezzük, hogy Ω véges és az ω_i elemi események előfordulásának esélye azonos (!).

7. Definíció. Az $A \in \mathcal{P}(\Omega)$ esemény klasszikus értelemben vett valószínűségén az alábbi törtet értjük:

$$P(A) = \frac{A - nak\ kedvező\ esetek\ száma}{összes\ eset}.$$

8. Példa.

(1) Dobókocka. A=”a kocka 2 pontot mutat (2 -es)” ,$\Rightarrow P(A) = \frac{1}{6}$. B=”páros szám” ,$\Rightarrow P(B) = \frac{3}{6}$.

(2) Két dobókocka. A=”a két kocka $(2,6)$ -ost mutat”, $\Rightarrow P(A) = \frac{1}{36}$. B=”mindkét kocka azonos számot mutat”, $\Rightarrow P(B) = \frac{6}{36}$. C=”a kockák „összege” ≥ 10”, $\Rightarrow P(C) = \frac{6}{36}$.

$$\begin{array}{ccccc} (6,6) & (6,5) & \dots & (6,1) & (6,0) \\ & (5,5) & & & (5,0) \\ & & \ddots & & \\ & & & (1,1) & (1,0) \\ & & & & (0,0) \end{array}$$

(3) Dominó játék Annak a valószínűsége, hogy egyformát emeljünk ki $\frac{7}{28}$.

9. Példa. Galton paradoxon. Ha feldobunk 3 érmét ebből (legalább) 2 egyforma. Mivel a harmadik érme vagy fej, vagy irást mutat annak a val., hogy mind a három érme egyformát mutat egyenlő $\frac{1}{2}$, igaz? A válasz hamis: a 8 lehetőségből $FFF, FFI, FIF, IFF, FII, IFI, IIF, III$ csak két esetben (FFF, III) teljesül a kérés, tehát a val: $\frac{2}{8}$. Nem egy TETSZŐLEGES érméről van szó, hanem A HARMADIK érméről!

10. Példa. (Osztozkodási feladat) Feltételezve, hogy egy játék megnyeréséhez 10 csatát kell megnyernie az A, illetve B játékosoknak, hogyan kell elosztani a nyereményt ha $8-7$ állásnál (az A játékos javára) a játék félbemarad?

Bizonyítás. Figyelembe véve, hogy maximum 4 további csata eldönti a játék menetét a nyereményt a hátramaradt csaták nyerési esély arányában osztjuk el. A lehetséges kimenetelek a következők:

A nyer	*aaaa*	*aaab*	*aaba*	*abaa*	*baaa*	*aabb*	*abab*	*baab*	*abba*	*baba*	*bbaa*
B nyer	*bbbb*	*bbba*	*bbab*	*babb*	*abbb*						

ahol például *babb* azt jelenti, hogy az első, a harmadik és a negyedik csatát a B játékos nyerte, a másodikat pedig az A játékos. Természetesen a fölösleges csatákat nem szokták lejátszani. Tehát annak a valószínűsége, hogy A nyer $=\frac{11}{16}$, míg B nyerési esélye $=\frac{5}{16}$. A nyereményt ugyanebben az arányban kell elosztani a két játékos között. □

Axiomatikus értelmezés.

11. Definíció. $P:\mathcal{A}\to[0,1]$

- $P(A)\geq 0$;
- $P(\Omega)=1$ (biztos esemény valószínűsége $=1$);
- $P(A\cup B)=P(A)+P(B)$, ha $A\cap B=\emptyset$.

A $(\Omega,\mathcal{A},P)$ hármast valószínűségi-mezőnek nevezzük.

12. Tétel. *Tulajdonságok*

(1) $P(\emptyset)=0$.
(2) $P\left(\overline{A}\right)=1-P(A)$.
(3) $P(B-A)=P(B)-P(A\cap B)$.
(4) $A\subseteq B\Rightarrow P(A)\leq P(B)$.
(5) $P(A\cup B)=P(A)+P(B)-P(A\cap B)$.
(6) $P(A\cup B)\leq P(A)+P(B)$.

13. Példa. Dominó játék. Mi annak a valószínűsége, hogy a következő konfigurációt folytassuk?

- $(3,4)-(4,0)-(0,0)-(0,3)-(3,2)$
- $(3,4)-(4,0)-(0,0)$

- $(3,4)-(4,0)-(0,0)-(0,3)$

14. DEFINÍCIÓ. Feltételes valószínűség ($P(B)\neq 0$):

$$P_A(B)=\frac{P(A\cap B)}{P(A)}.$$

Teljes valószínűség és Bayes képlet: Ha A_i az Ω egy teljes felbontását jelenti és $B\in\Omega$

$$P(B)=P(A_1)P_{A_1}(B)+P(A_2)P_{A_2}(B)+\dots$$

Szorzási képlet:

$$P(A\cap B)=P(A)\cdot P_A(B).$$

15. DEFINÍCIÓ. Az A és B események függetlenek ha $P_A(B)=P(B)$.

16. TÉTEL. *Ha A és B függetlenek akkor*

$$P(A\cap B)=P(A)\cdot P(B).$$

1.3. Valószínűségi változók

Ω,ω_i

$$X:\omega_i\to x$$

Diszkrét vv. $X(\Omega)=\{x_1,x_2,\dots,x_n\}$ lehetséges értékek $P(X=x_i)=p_i$
X vv. eloszlása:

$$X:\begin{pmatrix}x_1 & x_2 & \dots & x_n\\ p_1 & p_2 & \dots & p_n\end{pmatrix}.$$

17. PÉLDA. Dobókocka. $X:\begin{pmatrix}1 & 2 & \dots & 6\\ \frac{1}{6} & \frac{1}{6} & \dots & \frac{1}{6}\end{pmatrix}$

18. DEFINÍCIÓ. Eloszlásfüggvény (kumulált): $F:\mathbb{R}\to\mathbb{R}$

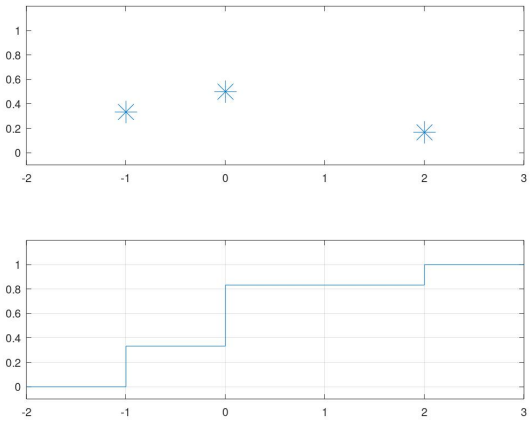
$$F(x)=P(X<x).$$

Tulajdonságok:

- $F(-\infty)=0$;
- $F(\infty)=1$
- F szakadásos (lépcsős)

- F monoton növekvő.

19. Példa. $X : \begin{pmatrix} -1 & 0 & 2 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{pmatrix}$ akkor $F(x) = \begin{cases} 0, & x \in (-\infty, -1] \\ \frac{1}{3}, & x \in (-1, 0] \\ \frac{5}{6}, & x \in (0, 2] \\ 1, & x \in (2, \infty) \end{cases}$.



1.3.1. ábra. Az X eloszlása és eloszlásfüggvénye

20. Tétel.

$$P(a \leq X < b) = F(b) - F(a).$$

Bizonyítás. $A =$ "$X < a$", $B =$ "$X < b$" $\Rightarrow P(a \leq X < b) = P(\overline{A} \cap B) = P(B - A) = P(B) - P(A \cap B) = P(B) - P(A) = F(b) - F(a)$. □

21. Definíció. Az $X : \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$ véges változó

(1) várható értéke vagy átlaga

$$M(X) = \sum_{i=1}^{n} x_i p_i \tag{1.3.1}$$

(2) szórásnégyzete

$$D^2(X) = \sum_{i=1}^{n} (x_i - m)^2 p_i, \tag{1.3.2}$$

ahol $m = M(X)$ az X várható értéke

(3) szórása

$$D(X) = \sqrt{D^2(X)}. \tag{1.3.3}$$

22. Tétel.

$$D^2(X) = M\left(X^2\right) - (M(X))^2$$

23. Példa. Monte Carlo roulette 0-36. 0=zöld, páratlan (1,3,...,35)= piros, páros (2,4,...,36)= fekete. Ha 1eurót teszünk fel pirosra és piros jön ki nyerünk egy eurót, különben elveszítjük az 1 eurót. Az általunk nyert összeg egy X vv. Adjuk meg a vv. eloszlását, az eloszláfggv., ábrázolás. Várható érték, szórás. Mi történik akkor ha a 0-ra egy fél eurót visszakapunk (a másik fél a banké)?

24. Példa. 2 kocka összege egy S vv. Adjuk meg az S eloszlását, várható értékét, szórását! Ha X, Y vv. a két kocka által mutattott pontok, milyen összefüggés van az S és az X, Y, illetve a várható értékek és szórások között?

25. Példa. Játék 2 kocka. Ha az összeg 12 vagy 2 kapok 8 lejt, ha az összeg 7 (a leggyakoribb) kapok 2 lejt. Ezenkívül fizetek 1 lejt. A nyert összeg egy X vv. Adjuk meg a vv. eloszlását, az eloszlásfüggvényt, ábrázolás. Várható érték, szórás. A játék korrekt (várható érték = 0)?

1.4. Nevezetes eloszlások

1.4.1. Diszkrét eloszlások

1.4.1.1. Binomiális eloszlás Ha egy kisérlet folyamán egy A esemény valószínűsége $p = P(A)$ nem módosul, akkor annak a valószínűsége, hogy n kísérletből az A esemény k-szor előforduljon:

$$b(n, k; p) = C_n^k p^k q^{(n-k)}, \tag{1.4.1}$$

ahol $q = 1 - p$.

Az eloszlás jól modellezhető egy urnával amelyben N golyó van (ezek közül N_1 egy színű (piros) $(N - N_1)$ más színű (fehér)) és amelyből - visszatevéssel - kiemelünk n -szor egy-egy golyót. Ha A-val jelöljük a „kiemelt golyó piros" eseményt aminek a valószínűsége $\frac{N_1}{N} := p$, akkor annak valószínűsége, hogy az n kiemelt

golyóból k-szor piros forduljon (és $(n-k)$-szor fehér) a (1.4.1) képlettel kapjuk meg.

26. PÉLDA. Egy dobozban 7 piros és 3 fehér golyó van. 12 alkalommal (visszatevéssel) kiemelünk egy-egy golyót. Mi a valószínűsége annak, hogy a piros golyó 8 -szor szerepeljen?

A="a golyó piros" $p=\frac{7}{10}, \Rightarrow b\left(12,8;p\right)=C_{12}^{8}\left(\frac{7}{10}\right)^{8}\left(\frac{3}{10}\right)^{4}=0.2311.$

27. PÉLDA. Egy dobókockát 10 -szer dobunk fel. Mi a valószínűsége, hogy legalább 8-szor lesz páros szám?

A="páros szám" $p=\frac{1}{2}, \Rightarrow b\left(10,8;p\right)+b\left(10,9;p\right)+b\left(10,10;p\right)=C_{10}^{8}\left(\frac{1}{2}\right)^{8}\left(\frac{1}{2}\right)^{2}+C_{10}^{9}\left(\frac{1}{2}\right)^{9}\left(\frac{1}{2}\right)^{1}+C_{10}^{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^{0}=0.0547.$

28. DEFINÍCIÓ. Egy X vv. binomiális eloszlást követ ha eloszlása

$$X:\left(\begin{array}{c} k \\ C_n^k p^k q^{(n-k)} \end{array}\right)_{k=0,\ldots,n}. \tag{1.4.2}$$

29. TÉTEL. *A (1.4.2) binomiális eloszlást követő X vv. várható értéke, szórásnégyzete:*

$$M(X)=np,\ D^2\left(X\right)=npq.$$

30. PÉLDA. Egy dobókockát 4-szer dobunk fel. Jel. X-el a páros számok megjelenésének a vv. Adjuk meg az X eloszlását. Mi a valószínűsége annak, hogy legfeljebb 3-ast dobunk?

$A=$ "páros szám" $\Rightarrow P\left(A\right)=\frac{1}{2}\Rightarrow X:\left(\begin{array}{cccccc} 0 & & 1 & & 2 & 3 \quad 4 \\ C_4^0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4 & & C_4^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 & & C_4^2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 & \frac{1}{4} \quad \frac{1}{16} \end{array}\right).$

$P\left(\text{"legfeljebb 3"}\right)=P\left(\left(X=0\right)\cup\left(X=1\right)\cup\left(X=2\right)\cup\left(X=3\right)\right)=P\left(X=0\right)+P\left(X=1\right)+P\left(X=2\right)+P\left(X=3\right)=1-P\left(X=4\right)=\frac{5}{16}.$

1.4.1.2. Hipergeometrikus eloszlás Az eloszlás jól szemléltethető az urnás modellel: Egy urnában van N golyó amiből N_1 piros. Az urnából kiemelünk n golyót visszatevés nélkül. Annak a valószínűsége, hogy a kiemelt golyókból pontosan k golyó lesz piros:

$$P\left(n,k\right)=\frac{C_{N_1}^{k}C_{N-N_1}^{n-k}}{C_N^n}. \tag{1.4.3}$$

A fenti képletben feltételezzük, hogy a létezési feltételek teljesülnek: $n\leq N$, stb.

31. Példa. Egy urnában 5 piros és 3 fehér golyó van. Visszatevés nélkül kiemelünk 6 golyót. Mi a valószínűsége, hogy a kiemelt golyókból pontosan 4 piros?

$P(6,4) = \frac{C_5^4 C_3^2}{C_8^6} = 0.5357.$

32. Definíció. Egy X vv. hipergeometrikus eloszlást követ ha eloszlása

$$X : \left(\begin{array}{c} k \\ \frac{C_{N_1}^k C_{N-N_1}^{n-k}}{C_N^n} \end{array} \right)_{k=0,\ldots,n}. \tag{1.4.4}$$

33. Tétel. *A (1.4.4) hipergeometrikus eloszlást követő X vv. várható értéke, szórásnegyzete:*

$$M(X) = np,\ D^2(X) = npq\left(1 - \frac{n-1}{N-1}\right),$$

ahol $p = \frac{N_1}{N},\ q = 1 - p.$

34. Példa. Adjuk meg az X vv. ha X a 6/49 Lottó nyerő számainak a valószínűségi változója. Mi a valószínűsége, hogy legalább 5-ös találatunk legyen? Számítsuk ki az átlagot, illetve a szórást!

$$X : \left(\begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{C_6^0 C_{43}^6}{C_{49}^6} & \frac{C_6^1 C_{43}^5}{C_{49}^6} & \frac{C_6^2 C_{43}^4}{C_{49}^6} & \frac{C_6^3 C_{43}^3}{C_{49}^6} & \frac{1}{10324} & \frac{1}{54201} & \frac{1}{13983816} \end{array} \right).$$

P("legalább 5 találat") = P("pontosan 5 találat")+P("pontosan 6 találat") = $\frac{1}{54201} + \frac{1}{13983816} = 1.852 \cdot 10^{-5}$.

$n = 6,\ p = \frac{6}{49} \Rightarrow M(X) = 6 \cdot \frac{6}{49} = 0.73$, átlagban 0.73 találatunk lesz! $D^2(X) = 6\frac{6}{49}\frac{43}{49}\left(1 - \frac{5}{48}\right) = 0.57 \Rightarrow D(X) = 0.76.$

1.4.1.3. Poisson eloszlás

35. Definíció. Egy X vv. $\lambda > 0$ paraméterű Poisson eloszlást követ ha eloszlása:

$$X : \left(\begin{array}{c} k \\ \frac{\lambda^k}{k!} e^{-\lambda} \end{array} \right)_{k=0,1,2,\ldots}. \tag{1.4.5}$$

36. Tétel. *A (1.4.5) Poisson eloszlást követő X vv. várható értéke, szórásnégyzete:*

$$M(X) = \lambda,\ D^2(X) = \lambda.$$

A Poisson eloszlás a binomiális eloszlás határeseteként is értelmezhető: $n \to \infty$ és $\lambda = pn$ állandó, ugyanis

$$\lim b(n,k,p) = \lim_{n\to\infty} C_n^k p^k (1-p)^{n-k} = \lim_{n\to\infty} \frac{1}{k!}\frac{n!}{(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} =$$

$$= \frac{\lambda^k}{k!}\lim_n \left(\frac{n!}{(n-k)!n^k}\left(\left(1-\frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right)^{\frac{-\lambda}{n}(n-k)}\right) = \frac{\lambda^k}{k!}e^{-\lambda}.$$

37. Példa. A $\lambda = \frac{1}{3}$ paraméterű Poisson eloszlása $k = 2$-re $P(X=2) = \frac{\left(\frac{1}{3}\right)^2}{2!}e^{-\frac{1}{3}} = 0.0398$ és az X eloszlása:

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & \dots \\ 0.7165 & 0.2388 & 0.0398 & \frac{\lambda^3}{3!}e^{-\lambda} & \end{pmatrix}.$$

1.4.1.4. Geometriai eloszlás Annak a valószínűsége, hogy egy A esemény (csak) a k-ik kísérletnél következzen be

$$P = pq^{k-1},$$

ahol $p = P(A)$, $q = 1-p$.

38. Példa. Mi a valószínűsége, hogy egy pénzérme feldobásánál csak az 5 -ik dobásnál legyen fej?

$P = \frac{1}{2}\left(\frac{1}{2}\right)^4 = \frac{1}{32}$.

39. Definíció. Egy X vv. geometriai eloszlást követ ha eloszlása:

$$X : \begin{pmatrix} k \\ pq^{k-1} \end{pmatrix}_{k=1,2,\dots}. \tag{1.4.6}$$

40. Példa. Mi a valószínűsége, hogy egy pénzérme feldobásánál csak az 5 -ik dobásnál legyen fej?

$P()$

41. Tétel. *A (1.4.6) geometriai eloszlást követő X vv. várható értéke, szórásnégyzete:*

$$M(X) = \frac{1}{p},\ D^2(X) = \frac{1-p}{p^2}.$$

1.4.2. Folytonos eloszlások Egy X valószínűségi változó folytonos ha az eloszlásfüggvénye folytonos.

1.4.2.1. Egyenletes eloszlás

42. DEFINÍCIÓ. Egy X vv. egyenletes eloszlást követ ha sűrűségfüggvénye:

$$f(x) = \begin{cases} \frac{1}{(b-a)} & x \in [a,b] \\ 0 & \text{különben} \end{cases}.$$

1.4.2.2. Normális eloszlás Az Euler-féle gamma fggv.

43. DEFINÍCIÓ. Euler-féle gamma fggv. nevezzük az alábbi fggv-t:

$$\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx,\ p > 0.$$

$p = 1, 2$ a gamma függvény értéke

$$\Gamma(1) = \int_0^\infty e^{-x}dx = -e^{-x}\,|_0^\infty = 1$$

$$\Gamma(2) = \int_0^\infty xe^{-x}dx = -\int_0^\infty x\left(e^{-x}\right)'dx = -x\left(e^{-x}\right)|_0^\infty + \int_0^\infty e^{-x}dx = 1,$$

illetve további p pozitív egészekre:

$$\Gamma(p+1) = \int_0^\infty x^p e^{-x}dx = -\int_0^\infty x^p\left(e^{-x}\right)'dx = -x^p\left(e^{-x}\right)|_0^\infty + p\int_0^\infty x^{p-1}e^{-x}dx = p\Gamma(p)$$

vagyis

$$\Gamma(p+1) = p\Gamma(p). \tag{1.4.7}$$

A (1.4.7)-ből következik, hogy

$$\Gamma(p+1) = p(p-1)\ldots\Gamma(1),$$

és

$$\Gamma(p+1) = p!, \tag{1.4.8}$$

vagyis a Γ függvény általánosítása a faktoriális függvénynek.

44. TÉTEL. *Euler-féle tükrözési képlet.*

$$\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin \pi x}. \tag{1.4.9}$$

Az előbbi tételből $x = \frac{1}{2}$-re következik, hogy

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{1.4.10}$$

45. Definíció. Gauss-féle integrál

$$\int_0^\infty e^{-x^2} dx.$$

A Gauss integrál kifejezhető a gamma fggv. segítségével:

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx = \sqrt{\pi},$$

ahonnan $x = t^2$ változócserével következik $dx = 2t \cdot dt$ és

$$\int_0^\infty \left(t^2\right)^{-\frac{1}{2}} e^{-t^2} 2t \cdot dt = \sqrt{\pi}$$

vagyis

$$\int_0^\infty e^{-t^2} dt = \frac{\sqrt{\pi}}{2}.$$

46. Definíció. Egy X vv. m, σ $\left(m \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\right)$ paraméterű normális eloszlást követ ha sűrűségfüggvénye:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2},\ x \in \mathbb{R}.$$

1.4.2.3. t-eloszlás

47. Definíció. Egy X vv. t (Student) ν-szabadsáfokú eloszlást követ ha sűrűségfüggvénye:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}},\ x \in \mathbb{R}.$$

Ha $\nu \to \infty$ az eloszlás közelíti a standard normális eloszlást.

1.4.2.4. χ^2 (khi-négyzet) eloszlás

48. Definíció. Egy X vv. ν-szabadsáfokú χ^2 eloszlást követ ha sűrűségfüggvénye:

$$f(x) = \frac{x^{\frac{\nu-2}{2}} e^{-\frac{x}{2}}}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)},\ x \in \mathbb{R}.$$

2. FEJEZET

Bevezetés a statisztikába

2.1. Statisztikai minta, gyakoriság

Alapsokaság=a statisztikai megfigyelés tárgyát képező egyedek összessége, halmaza.
Statisztikai minta=az alapsokaságból kiválasztott egyedekhez tartozó adatok. A minta kiválasztásánál figyelni kell, hogy a minta *reprezentatív* legyen, vagyis az adott sokaságot jellemezze. Egy $X_1, ..., X_n$ minta elemei is valószínűségi változók mert értékük (jel. $x_1, ..., x_n$) a mintázási technikától vagyis a véletlentől függ. Ezeket a változókat függetlennek és azonos eloszlásúnak tekintjük. A belőlük képzett valószínűségi változók *statisztikai függvényeknek* vagy egyszerűen *statisztikák*nak nevezzük. Mivel ezeket tapasztalati úton szerezzük empirikus vagy tapasztatlati statisztikáknak nevezzük.

2.2. A minta számszerű jellemzői

Jelöljük $X_1, X_2, ...X_n$ egy minta elemeit.

Mintaátlagnak (mintaközép, empirikus várható érték) nevezzük az alábbi mérőszámot

$$\overline{X} = \frac{X_1 + ... + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{2.2.1}$$

Ha $\overline{X}$ -el toljuk el az X mintát akkor ennek a várható értéke zéró

$$\sum_{i=1}^{n} \left(X_i - \overline{X}\right) = 0.$$

Az átláthatóság érdekében gyakran csoportosítjuk az adatokat (például növekvő sorrendbe), vagy osztályokba rendezzük.

49. Példa. Az

$$\begin{aligned} X: \quad & 6,2,7,2,5,3,7,3,4,6,4,5,4,8,4,5,6,7,5,9, \\ & 5,3,5,6,6,7,6,7,5,7,3,8,2,8,5,6,8,3,6,5 \end{aligned}$$

minta átlaga $\overline{X} = (6 + 2 + ... + 2)/40 = 5.325$. A minta átláthatóbb ha növekvő sorrendbe rendezzük

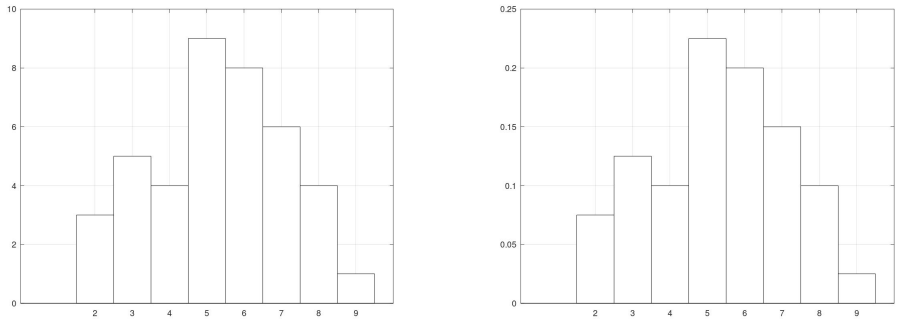
$$\begin{aligned} X^*: \quad & 2,2,2,3,3,3,3,3,3,4,4,4,4,5,5,5,5,5,5,5,5, \\ & 5,6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,8,8,8,8,9 \end{aligned}$$

illetve csoportosítjuk

X_i	2	3	4	5	6	7	8	9
f_i	3	5	4	9	8	6	4	1

1. táblázat. Csoportosított adatok

ahol f_i az X_i gyakorisága. Az adatok eloszlását téglalapokkal szemléltetjük amelyeknek a magasságát abszolút vagy relatív értékben adjuk meg. Ez utóbbi esetében a terület 1.



2.2.1. ábra. Hisztogram (abszolút/relatív skála)

Az adatokat - az átláthatóság érdekében- gyakran osztályokra bontjuk:

osztály	oszt. közép X_i	gyakoriság f_i	rel. gyak. g_i	kumulált rel. gyak. $\sum_i g_i$
$2-4$	3	12	12/40	12/40
$5-6$	5.5	17	17/40	29/40
$7-9$	8	11	11/40	1

2. táblázat. Adatok osztályozása

Az osztályok száma általában=$\sqrt{n}$-hez közeli érték, de szükség szerint változtatható az osztályok száma/hossza.

Csoportosított adatok esetén az átlag:

$$\overline{X} = \frac{f_1X_1 + ... + f_kX_k}{f_1 + ... + f_k} = \frac{\sum_{i=1}^{k} f_iX_i}{n}. \tag{2.2.2}$$

50. Példa. Az előbbi mintát véve alapul a mintaátlag

$$\overline{X} = \frac{3\cdot 2 + 5\cdot 3 + 4\cdot 4 + 9\cdot 5 + 8\cdot 6 + 6\cdot 7 + 4\cdot 8 + 1\cdot 9}{3+5+4+9+8+6+4+1} = 5.325,$$

vagy az osztály csoportosítást véve alapul

$$\overline{X} = \frac{3\cdot 12 + 5.5\cdot 17 + 8\cdot 11}{40} = 5.4375.$$

Habár az utolsó eredmény nem egyezik meg pontosan az előbbivel, az osztályokra való csoportosítás hasznos mert megkönnyíti az adatok feldolgozását főleg nagy számú minta esetén.

A mintára egy másik jellemző adat a *módusz,* vagyis a leggyakrabban előforduló adat. Az előbbi példában $Mod = 5$ mivel ennek a legnagyobb a gyakorisága= 9.

Egy másik szám amivel jellemezhető a minta a *medián*, illetve általánosítása a *kvartilisek, kvantilisek.* A medián a minta közepét jelöli, vagyis ha $X_1^* \leq ... \leq X_n^*$ a rendezett minta, akkor

$$Me = \begin{cases} X^*_{m+1} & \text{ha } n = 2m+1 \\ \frac{X^*_m + X^*_{m+1}}{2} & \text{ha } n = 2m \end{cases}. \tag{2.2.3}$$

Az első, második és harmadik kvartilis a minta negyedét, felét, háromnegyedét jelöli. A második kvartilis megegyezik a mediánnal. A p-kvantilisek a kvartilisek általánosításai.

51. Példa. Az adott minta mediánja $Me = 5$, és az első, második, harmadik kvartilise egyenlő rendre $4, 5, 7$-el.

Bizonyos statisztikák esetén a medián hitelesebben tükrözi a valóságot, mivel „kiszűri" a nagy kilengéseket. Például az átlagbér esetében a medián „kiszűri" a nagyon magas fizetések és a rengeteg minimálbéres által okozott torzítást. (Romániában 2018-ban bruttó átlagbér =4512 lej (980 euró), viszont 92% a lakosságnak ez alatt van a bére.) https://www.maszol.ro/index.php/hatter/121087-unios-szint-minimalber-fulel-uzemmodban-van-az-europai-bizottsag

A *mintaterjedelem* a minta legnagyobb és legkisebb érték közötti különbség

$$R = X_n^* - X_1^*. \tag{2.2.4}$$

Az említett mérőszámokon kívül szóródási mutatókat is számítunk.
A *tapasztalati szórásnégyzet* s^2, a mintaelemek mintaközéptől való eltérései négyzetének átlaga, azaz

$$s^2 = \frac{\left(X_1 - \overline{X}\right)^2 + ... + \left(X_n - \overline{X}\right)^2}{n} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{n}. \tag{2.2.5}$$

Kis minták esetében a *korrigált tapasztalati szórásnégyzetet* s^{*2} használjuk:

$$s^{*2} = \frac{\left(X_1 - \overline{X}\right)^2 + ... + \left(X_n - \overline{X}\right)^2}{n-1} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{n-1}. \tag{2.2.6}$$

A *tapasztalati* s, illetve *korrigált szórás* s^*

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{n}}, \tag{2.2.7}$$

$$s^* = \sqrt{s^{*2}} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}. \tag{2.2.8}$$

A (2.2.7),(2.2.8)-ből következik, hogy

$$\frac{s^*}{s} = \sqrt{\frac{n}{n-1}}$$

vagyis nagy n értékre a két szórás megegyezik.

Csoportosított adatok esetén a szórásokat a

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i \left(X_i - \overline{X}\right)^2}{n}} \tag{2.2.9}$$

$$s^* = \sqrt{\frac{\sum_{i=1}^{k} f_i \left(X_i - \overline{X}\right)^2}{n-1}} \tag{2.2.10}$$

képletekkel számítjuk ki ahol k a csoportosítás után létrejött osztályok számát jelöli.

Az Excelben a következő függvényekkel számítjuk ki az ismertetett paramétereket: ÁTLAG(), MÓDUSZ(), MEDIÁN(), SZÓRÁSP(), SZÓRÁS().

2.3. Statisztikai becslések

A mintavételezés elsődleges célja, hogy információkat szolgáltasson az alapsokaság jellemzőire, például a várható értékre, szórásra.

2.3.1. Pontbecslés Egy becslés akkor tekinthető jónak ha eleget tesz legalább az egyiknek az alábbi követelményeknek: torzítatlan, hatásos, konzisztens, elégséges.

52. Tétel. *A mintaátlag torzítatlan becslést ad az X alapsokaság várható értékére, vagyis*

$$M\left(\overline{X}\right) = M\left(X\right). \tag{2.3.1}$$

Bizonyítás. A várható érték additivitását és homogenitását felhasználva következik, hogy

$$M\left(\overline{X}\right) = M\left(\frac{X_1 + ... + X_n}{n}\right) = \frac{1}{n} M\left(X_1 + ... + X_n\right) = \frac{1}{n}\left(M\left(X_1\right) + ... + M\left(X_n\right)\right) =$$
$$\frac{1}{n}\left(M\left(X\right) + ... + M\left(X\right)\right) = \frac{1}{n} n M\left(X\right) = M\left(X\right).$$

□

53. Tétel. *Az empirikus korrigált szórásnégyzet torzítatlan becslést ad az X alapsokaság $D^2\left(X\right)$ szórásnégyzetére, vagyis*

$$M\left(s^{*2}\right) = D^2\left(X\right).$$

Bizonyítás.

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 + \overline{X}^2 - 2X_i\overline{X}\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{1}{n}\sum_{i=1}^{n}\overline{X}^2 - \frac{2\overline{X}}{n}\sum_{i=1}^{n} X_i =$$
$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{1}{n}\sum_{i=1}^{n}\overline{X}^2 - \frac{2\overline{X}}{n} n\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \overline{X}^2 - 2\overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

Figyelembe véve, hogy $M\left(\overline{X}\right) = M\left(X\right)$, $M\left(X_i\right) = M\left(X\right)$ következik

$$M\left(s^2\right) = M\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left(M\left(X_i^2\right) - M\left(X_i\right)^2\right) - \left(M\left(\overline{X}^2\right) - M\left(\overline{X}\right)^2\right) =$$
$$= \frac{1}{n}\sum_{i=1}^{n} D^2\left(X_i\right) - D^2\left(\overline{X}\right) = \frac{1}{n} n D^2\left(X\right) - \frac{1}{n^2} n D^2\left(X\right) = \frac{n-1}{n} D^2\left(X\right),$$

tehát a s^2 torzított becslést ad, viszont a

$$s^{*2} = \frac{n}{n-1} s^2 \tag{2.3.2}$$

értelmezett korrigált empirikus szórás becslése torzítatlan

$$M\left(s^{*2}\right) = D^2\left(X\right).$$

□

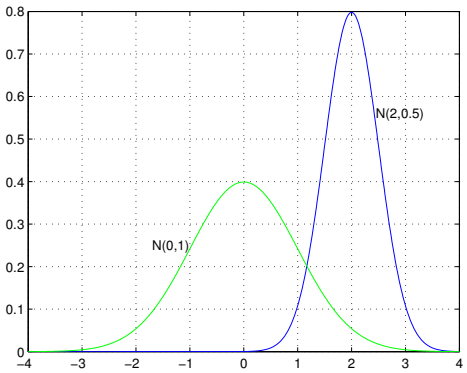
54. TÉTEL. *Az empirikus korrigált szórásnégyzet konzisztens becslést ad az X alapsokaság $D^2(X)$ szórásnégyzetére, vagyis*

$$\lim_{n\to\infty} D^2\left(s^{*2}\right) = 0.$$

2.3.2. Intervallumbecslés Az előbbi fejezettől eltérően ebben a fejezetben egy (szimmetrikus) *konfidencia (megbízhatósági) intervallumot* állapítunk meg amely nagy valószínűséggel tartalmazza az alapsokaság jellemzőjét. Az említett nagy valószínűséget $(1-p)$ -vel jelöljük, *konfidencia (megbízhatósági) szint*nek nevezzük és a leggyakrabban 0.99, 0.95 vagy 0.90 értékeket veszi fel. A megfelelő p értéket $(0.01,\ 0.05,\ 0.1)$ *tévedési* vagy *szignifikanciaszint*nek nevezzük.

Egy X valószínűségi változó m,σ paraméterű, *normális eloszlás*únak nevezzük (jele $X \in N(m,\sigma)$) ha sűrűségfüggvénye:

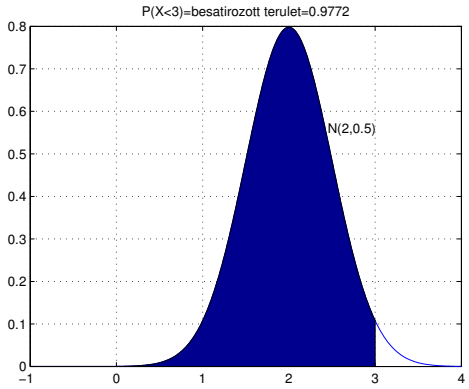
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},\ x \in \mathbb{R}.$$



2.3.1. ábra. Normális eloszlás

Az X változót *standard normális eloszlású*nak nevezzük ha $X \in N(0,1)$. Annak a valószínűségét, hogy az X kisebb legyen egy adott x értéknél az F eloszlásfüggvénnyel fejezzük ki:

$$P(X < x) = F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$



2.3.2. ábra. $F(x)$

Geometriailag $F(x)$ egyenlő a besatírozott területtel $-\infty$-től x-ig.

2.3.2.1. A várható érték becslése

55. Tétel. *Ha $X_1, ..., X_n \in N(m, \sigma)$ normális eloszlású változók akkor*

$$\overline{X} = \frac{X_1 + ... + X_n}{n} \in N\left(m, \frac{\sigma}{\sqrt{n}}\right).$$

Bizonyítás.

$$M\left(\overline{X}\right) = M\left(\frac{X_1 + ... + X_n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{1}{n} nm = m$$

$$D^2\left(\overline{X}\right) = D^2\left(\frac{X_1 + ... + X_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} D^2(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n},$$

ahonnan

$$D\left(\overline{X}\right) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

□

56. KÖVETKEZMÉNY. *Az*

$$u = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \tag{2.3.3}$$

változó standard normális eloszlású $u \in N(0,1)$.

Ajánlott a standard eloszlás előnyeit (szimmetria, egyszerűbb képlet, táblázatban megadott értékek) kihasználni ezért a (2.3.3) standardizálással bármely $N(m,\sigma)$ normális eloszlás visszavezethetünk a $N(0,1)$ standard normális eloszlásra.

Mivel $u \in N(0,1)$ a konfidencia-intervallumot a 0-ra szimmetrikusan $[-u_p, u_p]$ fogjuk meghatározni

$$P\left(-u_p \leq \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} < u_p\right) = 1 - p. \tag{2.3.4}$$

Következik, hogy

$$\Phi(u_p) - \Phi(-u_p) = 1 - p$$

tehát

$$2\Phi(u_p) - 1 = 1 - p$$

ahonnan

$$\Phi(u_p) = 1 - \frac{p}{2} \tag{2.3.5}$$

ahol Φ a standard eloszlás eloszlásfüggvénye. Az (2.3.5) képletből kiszámítható u_p értéke

$$u_p = \Phi^{-1}\left(1 - \frac{p}{2}\right) \tag{2.3.6}$$

aminek segítségével megszerkesztjük a konfidencia-intervallumot

$$-u_p \leq \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} < u_p.$$

A műveletek elvégzése után kapjuk az m várható értékre igaz a alábbi egyenlőtlenség

$$\overline{X} - u_p\frac{\sigma}{\sqrt{n}} < m \leq \overline{X} + u_p\frac{\sigma}{\sqrt{n}}$$

Tehát az m konfidencia-intervallumaként használjuk a

$$\left(\overline{X} - u_p\frac{\sigma}{\sqrt{n}}, \overline{X} + u_p\frac{\sigma}{\sqrt{n}}\right] \tag{2.3.7}$$

ami azt jelenti hogy a alapsokaság m várható értéke $1-p$ valószínűséggel esik ebbe az intervallumba.

Az Excelben a φ normális sűrűség eloszlást, illetve Φ eloszlásfüggvényt a NORM.ELOSZL() függvénnyel számítjuk ki. Az eloszlásfüggvény inverzét a INVERZ.NORM() függvénnyel számítjuk ki.

57. Példa. Az előbbi példa adataira határozzuk meg a konfidencia-intervallumot 95%-os megbízhatósági szinten ha az alapsokaság szórása ismert $\sigma = 1.83$. Az

$$u_p = \Phi^{-1}\left(1 - \frac{0.05}{2}\right) = 1.96,$$

tehát a konfidencia-intervallum

$$\left(5.325 - 1.96\frac{1.83}{\sqrt{40}}, 5.325 + 1.96\frac{1.83}{\sqrt{40}}\right] = (4.7579, 5.8921]\,,$$

vagyis 95%-os valószínűséggel az alapsokaság várható értéke ebben az intervallumban található.

Nagyobb valószínűség esetén az intervallum bővül. Például 99%-os megbízhatósági szinten $u_p = 2.5758$ és az intervallum $(4.5797, 6.0703]$.

Ha az alapsokaság szórása ismeretlen akkor σ helyett az s^* minta korrigált szórást helyettesítve az u (2.3.3) kifejezésbe a

$$t = \frac{\overline{X} - m}{\frac{s^*}{\sqrt{n}}} \tag{2.3.8}$$

valószínűségi változót kapjuk, amely egy $n-1$ szabadságfokú t (Student) eloszlású valószínűségi változó. A (2.3.4) hasonlóan, a t-eloszlás táblázatából kikeressük azt a t_p értéket amelyre

$$P\left(-t_p \le \frac{\overline{X} - m}{\frac{s^*}{\sqrt{n}}} \le t_p\right) = P\left(|t| \le t_p\right) = 1 - p,$$

vagyis annak a valószínűsége, hogy t értéke a $[-t_p, t_p]$ intervallumba essen egyenlő $1-p$-vel, ami ekvivalens azzal, hogy annak a valószínűsége, hogy az alapsokaság m várható értéke a

$$\left[\overline{X} - t_p \frac{s^*}{\sqrt{n}}, \overline{X} + t_p \frac{s^*}{\sqrt{n}}\right] \tag{2.3.9}$$

konfidencia intervallumba essen egyenlő $1-p$-vel.

Ha az előbbi példában a szórás ismeretlen akkor helyette használjuk a korrigált szórást $s^* = 1.8312$, a t-eloszlás táblázatából $1-p = 0.95$ és $df = 39$ szabadságfoknak megfelel a $t_p = 2.0227$ érték, tehát a konfidencia-intervallum

$$\left[5.325 - 2.0227\frac{1.8312}{\sqrt{40}}, 5.325 + 2.0227\frac{1.8312}{\sqrt{40}}\right] = [4.\,739\,4, 5.\,910\,6]\,.$$

A t-eloszlásfüggvényt az Excelben a T.ELOSZLÁS() függvénnyel, az inverzét INVERZ.T()-vel számítjuk ki.

2.4. Statisztikai hipotézisek vizsgálata

A statisztikai vizsgálatok során hipotéziseket (feltételezéseket) teszünk az alapsokaság bizonyos jellemzőire (várható érték, szórás, eloszlás, függetlenség, stb.).

A vizsgálat egy alap- vagy nullhipotézis (jel. H_0) és egy ellenhipotézis (jel. H) felállításából áll majd ezek helyességének ellenőrzéséből. Például, ha egy X alapsokaságra szeretnénk leellenőrizni, hogy a várható értéke $M(X) = m$ egyenlő-e egy m_0 értékkel akkor a nullhipotézis

$$H_0 : \ M(X) = m_0. \tag{2.4.1}$$

A vele szembe felállítandó ellenhipotézis kétféle lehet: kétoldali vagy egyoldali. A kétoldali ellenhipotézis

$$H : \ M(X) \neq m_0 \tag{2.4.2}$$

míg az egyoldali

$$H : \ M(X) < m_0 \text{ vagy } H : \ M(X) > m_0. \tag{2.4.3}$$

Alapszabályként, ha elfogadjuk a nullhipotézist elvetjük az ellenhipotézist és fordítva, ha elvetjük a nullhipotézist elfogadjuk az ellenhipotézist.

Az eljárást amellyel a nullhipotézist elfogadjuk vagy elvetjük *statisztikai próbának* nevezzük. A döntést statisztikai függvények segítségével hozzuk meg amit a mintából számítunk ki bizonyos $1 - p$ megbízhatóság szint mellett. Ennek ellenére, a döntés eredménye, mivel valószínűségi értékek alapján hozzuk, lehet hibás is. *Elsőfajú hibát* követünk el ha elvetjük az igaz H_0 hipotézist. *Másodfajú hibát* követünk el ha elfogadjuk a hibás H_0 hipotézist. A hibás döntés valószínűsége az első- és másodfajú hiba valószínűségeinek összegéből adódik

$$P(\text{hiba}) = P(\text{elsőfajú hiba}) + P(\text{másodfajú hiba}).$$

Az alábbi táblázat a helyes és hibás döntéseket szemlélteti

	H_0 igaz	H_0 hamis
H_0 elfogadása	helyes döntés	másodfajú hiba
H_0 elvetése	elsőfajú hiba	helyes döntés

2.4.1. Egymintás u-próba Feltételezzük, hogy adott egy $X \in N(m, \sigma)$ normális eloszlású valószínűségi változó amelynek m várható értékét nem, de σ szórását ismerjük. Vizsgálni szeretnénk, hogy a várható érték $M(X) = m$ megegyezike egy m_0 értékkel. Ennek érdekében egy $X_1, X_2, ..., X_n$ n elemű mintát veszünk amelynek $\overline{X}$ átlaga csak közelíteni fogja m értékét. Ha a nullhipotézis

$$H_0 : \ M(X) = m_0$$

igaz a kétoldali ellenhipotézissel szemben

$$H : \ M(X) \neq m_0$$

akkor a

$$u^* = \frac{\overline{X} - m_0}{\frac{\sigma}{\sqrt{n}}} \tag{2.4.4}$$

próbafüggvény standard normális eloszlású $u^* \in N(0,1)$, és $1-p$ valószínűséggel esik a $[-u_p, u_p]$ konfidencia-intervallumba, ahol u_p a táblázatból kiolvasott, $1-p$ szintnek megfelelő érték. Tehát a H_0 nullhipotézist elfogadjuk ha

$$|u^*| \leq u_p, \tag{2.4.5}$$

vagyis az u^* benne van az *elfogadási tartomány*ban. Ebben az esetben azt mondjuk, hogy az alapsokaság várható értéke és az m_0 feltételezett érték *különbsége nem szignifikáns.*

A nullhipotézist elvetjük (az ellenhipotézist elfogadjuk) ha

$$|u^*| > u_p, \tag{2.4.6}$$

vagyis ha az u^* a *kritikus tartomány*ban van. Ebben az esetben az alapsokaság várható értéke és az m_0 feltételezett érték között *szignifikáns eltérés* van.

58. Példa. Egy gépet 5.5 cm pálcikák vágására állították be. A beállítás ellenőrzésére egy 40 elemű mintát vesznek:
$6, 2, 7, 2, 5, 3, 7, 3, 4, 6, 4, 5, 4, 8, 4, 5, 6, 7, 5, 9, 5, 3, 5, 6, 6, 7, 6, 7, 5, 7, 3, 8, 2, 8, 5, 6, 8, 3, 6, 5.$
$p = 5\%$-os tévedési szinten, kell-e állítani a gépen ha a pálcikák hosszának szórása ismert $\sigma = 1$.

A pálcikák hossza normális eloszlást követ aminek szórása ismert. A várható értékre a hipotézisek

$$\begin{aligned} H_0 &: \quad M(X) = 5.5 \\ H &: \quad M(X) \neq 5.5. \end{aligned}$$

A minta átlaga $\overline{X} = 5.325$, tehát a (2.4.4) képletből $u^* = \frac{5.325-5.5}{\frac{1}{\sqrt{40}}} = -1.1068$, és a táblázatból $u_{p=0.05} = 1.96$. Mivel $|u^*| \leq u_p$ a nullhipotézist elfogadjuk, vagyis 0.95 valószínűséggel az alapsokaságban a pálcikák hossza nem tér el szignifikánsan az 5.5 értéktől, a gép jól működik.

2.4.2. Egymintás t-próba A legtöbbször a normális eloszlású alapsokaság szórását nem ismerjük. Ebben az esetben a Student-féle eloszlást használjuk. A

$$H_0 : \ M(X) = m_0$$

nullhipotézis ellenőrzésére a

$$t^* = \frac{\overline{X} - m_0}{\frac{s^*}{\sqrt{n}}} \tag{2.4.7}$$

próbastatisztikát használjuk, ahol n a mintaszám, $\overline{X}$ a mintaátlag és s^* a minta korrigált szórása. Az u próbához hasonlóan a nullhipotézist p szignifikancia szinten fogadjuk el ha

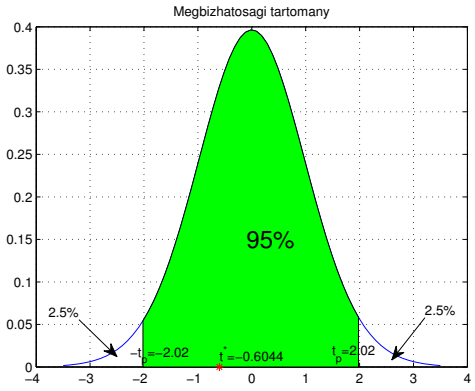
$$|t^*| \leq t_p \tag{2.4.8}$$

ahol t_p értékét a Student-féle táblázatból határozzuk meg $n-1$ szabadságfok és p szignifikanciának megfelelően.

59. Példa. Ha az előbbi példában a pálcikák szórása ismeretlen akkor

$$t^* = \frac{5.325 - 5.5}{\frac{1.8312}{\sqrt{40}}} = -0.604\,41,$$

és a $p = 0.05$ szignifikancia szintnek, illetve $df = 39$ szabadságfoknak megfelel a $t_p = 2.0227$ érték. Mivel $|t^*| \leq t_p$ következik, hogy 0.95 megbízhatósági szinten a nullhipotézist elfogadjuk. Az alábbi ábrán látszik, hogy a t^* számított érték benne van a megbízhatósági tartományban.



2.4.1. ábra. Megbízhatósági tartomány

Az Excelben lásd Z.PRÓBA(), T.PRÓBA().

2.4.3. χ^2-próba szórásvizsgálatra A hipotézis vizsgálatot szórásra is elvégezhetjük. Feltételezzük, hogy $X \in N(m, \sigma)$ egy normális eloszlású valószínűségi változó és vizsgálni szeretnénk, hogy a szórás értéke $D^2(X) = \sigma$ megegyezik-e egy σ_0 értékkel, vagyis származhat-e egy adott minta σ_0 szórású alapsokaságból? A null-illetve ellenhipotézisek

$$\begin{aligned} H_0 &: \quad \sigma = \sigma_0 \\ H &: \quad \sigma \neq \sigma_0 \text{ kétoldali ellenhipotézis} \\ H &: \quad \sigma < \sigma_0 \text{ vagy } \sigma > \sigma_0 \text{ egyoldali ellenhipotézis.} \end{aligned}$$

A kiválasztott n elemű mintában kiszámítjuk az s^{*2} korrigált szórásnégyzetet és a χ^{*2} próbastatisztikát

$$\chi^{*2} = \frac{(n-1)\, s^{*2}}{\sigma_0^2}, \tag{2.4.9}$$

majd összehasonlítjuk a táblázatbeli χ^2 (khi-négyzet) értékkel. Ha χ^{*2} benne van a megbízhatósági tartományban

$$\chi^{*2} < \chi^2 \tag{2.4.10}$$

akkor a nullhipotézist elfogadjuk.

60. Példa. Az adott minta származhat-e egy alapsokaságból amelynek szórása kisebb mint 1.5? A minta nagysága = 40, és a korrigált szórás $s^* = 1.83$, tehát $\chi^{*2} = \frac{39*1.83^2}{1.5^2} = 58.048$, míg a táblázatbeli érték $\chi^2 = 54.57$, tehát a (2.4.10) nem teljesül, a nullhipotézist elvetjük 95%-os megbízhatóság mellett.

2.4.4. Kétmintás u-próba Legyen $X \in N(m_X, \sigma_X)$ és $Y \in N(m_Y, \sigma_Y)$ két független, normális eloszlású valószínűségi változó, amelyeknek ismerjük a σ_X, illetve σ_Y szórásait. Legyen $X_1, X_2, ...$ és $Y_1, Y_2, ...$a változók egy n_X, illetve n_Y elemű minta. Vizsgálni kell, hogy a két sokaság várható értéke megegyezik-e. Ha

$$H_0 : \ M(X) = M(Y) \tag{2.4.11}$$

nullhipotézis teljesül a

$$H \ : \ M(X) \neq M(Y) \tag{2.4.12}$$

$$H \ : \ M(X) < M(Y) \text{ vagy } M(X) > M(Y) \tag{2.4.13}$$

ellenhipotézissel szemben, akkor az

$$u^* = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \tag{2.4.14}$$

próbastatisztika $N(0,1)$ eloszlású, ahol $\overline{X}$, $\overline{Y}$ az adott minták átlagai. Az egymintás próbához hasonlóan, a számított u^* változó $1-p$ valószínűséggel esik a $[-u_p, u_p]$ konfidencia-intervallumba, ahol u_p a táblázatból kiolvasott, $1-p$ szintnek megfelelő érték. Tehát a H_0 nullhipotézist elfogadjuk ha

$$|u^*| \leq u_p \tag{2.4.15}$$

és ebben az esetben azt állítjuk, hogy a két minta várható értékei között nincs szignifikáns eltérés.

2.4.5. Kétmintás t-próba Az előbbi alponthoz hasonlóan az $X \in N(m_X, \sigma_X)$ és $Y \in N(m_Y, \sigma_Y)$ minták várható értékének a vizsgálatát végezzük ha a minták szórásai ismeretlenek de feltehetően **azonosak**. Ebben az esetben a (2.4.11) nullhipotézis helyességét a

$$t^* = \frac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \tag{2.4.16}$$

$n_X + n_Y - 2$ szabadságfokú Student-eloszlású próbastatisztikával végezzük, ahol S a két minta alapján becsült közös szórás

$$S^2 = \frac{(n_X - 1)\, s_X^{*2} + (n_Y - 1)\, s_Y^{*2}}{n_X + n_Y - 2}, \tag{2.4.17}$$

és s_X^*, s_Y^* a két minta korrigált szórása.

A H_0 (2.4.12) nullhipotézist elfogadjuk ha t^* a $[-t_p, t_p]$ megbízhatósági intervallumba esik, vagyis

$$|t^*| \leq t_p.$$

A kétmintás t-próba csak azonos szórású mintára alkalmazzuk, ennek ellenőrzésére viszont az F-próbát (Fisher) használjuk.

2.4.6. F-próba Az F-próbával két $X \in N(m_X, \sigma_X)$ és $Y \in N(m_Y, \sigma_Y)$ független változó szórásainak $D(X) = \sigma_X$, illetve $D(Y) = \sigma_Y$ összehasonlítását végezzük. A nullhipotézis

$$H_0 : \ \sigma_X = \sigma_Y \tag{2.4.18}$$

és a megfelelő egyoldali, illetve kétoldali hipotézisek

$$H \ : \ \sigma_X \neq \sigma_Y \tag{2.4.19}$$

$$H \ : \ \sigma_X < \sigma_Y \text{ vagy } \sigma_X > \sigma_Y. \tag{2.4.20}$$

Jelöljük S^2, illetve s^2-el a két alapsokaságból kiválasztott minta korrigált szórásnégyzeteinek maximumát, illetve minimumát

$$S^2 = \max\left(s_X^{*2}, s_Y^{*2}\right), \ s^2 = \min\left(s_X^{*2}, s_Y^{*2}\right) \tag{2.4.21}$$

és jelöljük DF, illetve df -el a megfelelő minták szabadságfokát. Képezzük a

$$F^* = \frac{S^2}{s^2} \tag{2.4.22}$$

statisztikát amely $(DF - 1, df - 1)$ szabadságfokú F-eloszlást követ. A Fisher táblázatból kikeressük a F_p p-szignifikancia szintnek megfelelő értéket és ha

$$F^* \leq F_p, \tag{2.4.23}$$

akkor a H_0 hipotézist $p-$szignifikancia szinten elfogadjuk, vagyis a szórások közötti eltérés a véletlennek tulajdonítható.

A kétmintás próbának az alkalmazásához szükséges a szórások egyenlősége (amit az F-próbával vizsgálunk). Ha ez nem teljesül akkor a Welch próbát alkalmazzuk.

61. Példa. Két terület pH értékét mérve az alábbi mintákat kapták

X_i	6.3	6.2	6.0	6.3	6.7	6.4	6.7	6.6		
Y_j	6.2	6.1	6.2	6.0	6.4	6.4	6.3	6.7	6.0	6.2

Állítható-e, hogy a két terület talaj pH várható értéke megegyezik 95% megbízhatósági szinten?

Bizonyítás. A két minta átlagai $\overline{X} = 6.4$, $\overline{Y} = 6.25$ és korrigált szórásai $s_X^* = 0.2507$, $s_Y^* = 0.2121$. Az alapsokaság várható értékére a null-hipotézisek

$$H_0 : \ M(X) = M(Y)$$

és kétoldali ellenhipotézis

$$H : \ M(X) \neq M(Y).$$

Mivel az alapsokaság szórásait nem ismerjük a kétmintás t-próbát használjuk, de ehhez a szórások egyenlőségét kell leellenőrizni F-próbával.

$$H_0 \ : \quad D(X) = D(Y)$$
$$H \ : \quad D(X) \neq D(Y)$$

A (2.4.21) képletből $S^2 = 0.2507^2 = 0.062\,85$, $s^2 = 0.2121^2 = 0.0\,498\,6$, és $DF = n_X - 1 = 7$, $df = n_Y - 1 = 9$, majd (2.4.22)-ből $F^* = \frac{0.062\,85}{0.0\,498\,6} = 1.\,260\,5$. A $(7, 9)$ szabadságfok és $p = 0.05$ -nek megfelelő kétoldali Fisher táblázatbeli érték $F_p = 4.197$. Mivel $F^* < F_p$ a szórások egyenlőségére vonatkozó nullhipotézist elfogadjuk ($\sigma_X = \sigma_Y$). A továbbiakban a t-próbát használva vizsgáljuk, hogy a két várható érték egyenlő-e. A (2.4.17) képletből

$$S^2 = \frac{(8-1)\,0.2507^2 + (10-1)\,0.2121^2}{8 + 10 - 2} = 0.0528,$$

majd (2.4.16)-ből

$$t^* = \frac{6.4 - 6.25}{\sqrt{0.0528}\sqrt{\frac{1}{8} + \frac{1}{10}}} = 1.\,376\,2.$$

A t-eloszlás táblázatban $8+10-2$ szabadságfok és $p = 0.05$ szignifikancia szintnek $t_p = 2.12$ érték felel meg ahonnan $t^* < t_p$, tehát a nullhipotézist megtartjuk, vagyis

a két talaj pH várható értéke egyenlőnek tekinthető az adott szignifikancia szinten (a különbség a véletlennek tulajdonítható). □

2.5. Korreláció és regresszióanalízis

A *korreláció-analízis* két változó X és Y kapcsolatának szorosságát vizsgálja. A *regresszióanalízis* a változók közötti összefüggést próbálja függvényszerű leírni. A leíró függvényt *regressziós függvénynek* nevezzük.

A változók közötti szorosságot a korrelációs együtthatóval mérjük

$$\rho = \frac{M\left(\left(X - M\left(X\right)\right)\left(Y - M\left(Y\right)\right)\right)}{\sqrt{D^2\left(X\right) D^2\left(Y\right)}} = \frac{M\left(XY\right) - M\left(X\right) M\left(Y\right)}{D\left(X\right) D\left(Y\right)}$$

amely -1 és 1 közötti értékeket vesz fel

$$-1 \leq \rho \leq 1.$$

Ha X és Y függetlenek akkor $\rho = 0$ fordítva viszont nem igaz. Ha $\rho = 0$ akkor azt mondjuk, hogy X és Y *korrelálatlanok*. Az X és Y között annál szorosabb a kapcsolat minél közelebb van ρ abszolút értéke 1-hez. X és Y között $Y = aX + b$ lineáris kapcsolat van $\Leftrightarrow |\rho| = 1$.

Legyen $X_1, X_2, ..., X_n$ és $Y_1, Y_2, ..., Y_n$ két n elemű minta a két alapsokaságból. Az r *tapasztalati korrelációs együttható*

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 \sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2}} = \\ &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}}. \end{aligned} \tag{2.5.1}$$

Statisztikai próbával kell eldönteni, hogy az alapsokaságok amelyből az X és Y minták származnak korreláltak-e vagy sem. Jelöljük ρ-val az alapsokaságok korrelációs együtthatóját. Ekkor a null, illetve ellenhipotézis:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

a próba statisztika pedig

$$t_\rho = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{2.5.2}$$

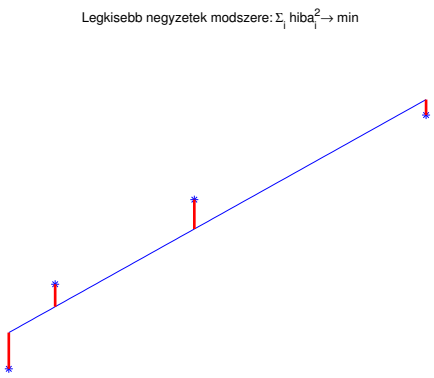
amely $df = n-2$ szabadságfokú t eloszlást követ. Ha egy adott $1-p$ megbízhatósági szintnek a t_ρ benne van az elfogadási tartományban $(-t_p, t_p)$, akkor a nullhipotézist elfogadjuk, különben elvetjük.

Ha a korrelációs együttható 0 és X, Y normális eloszlásúak, akkor függetlenek is, különben csak korrelálatlanok.

Ha X ás Y között lineáris kapcsolat van akkor

$$Y = aX + b$$

és a mintákat felhasználva becslést kell adni az a és b paraméterekre. Az egyik módszer az úgy nevezett legkisebb négyzetek módszere. A módszer abban áll, hogy az $y = ax + b$ egyenes és a mintában szereplő pontok közötti függőleges távolságok négyzeteinek összege minimális legyen.



2.5.1. ábra. Regressziós egyenes

Matematikailag az

$$F(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \to \min \tag{2.5.3}$$

kétváltozós függvény meghatározását jelenti. A stacionárius pontok

$$\begin{cases} \frac{\partial F}{\partial a} &= 0 \\ \frac{\partial F}{\partial b} &= 0 \end{cases}$$

aminek megoldása a következő lineáris egyenletrendszerhez vezet

$$\begin{cases} \sum_{i=1}^{n} 2\left(y_i - ax_i - b\right)\left(-x_i\right) &= 0 \\ \sum_{i=1}^{n} 2\left(y_i - ax_i - b\right)\left(-1\right) &= 0 \end{cases}$$

amely átrendezve az alábbi *normálegyenletrendszert* eredményezi

$$\begin{cases} bn + a\sum_i x_i &= \sum_i y_i \\ b\sum_i x_i + a\sum_i x_i^2 &= \sum_i x_i y_i \end{cases} \tag{2.5.4}$$

aminek megoldása

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}},\ a = \overline{Y} - b\overline{X}. \tag{2.5.5}$$

62. Példa. Hat páciensnél az alábbi vércukorszintet mérték

X_i páciens kora	43	21	25	42	57	59
Y_i vércukorszint	99	65	79	75	87	81

Számítsuk ki a minták korrelációs együtthatóját, majd tanulmányozzuk, hogy létezik-e összefüggés a kor és a vércukorszint között.

$\overline{X} = \frac{\sum X_i}{n} = 41.1667$; $\overline{Y} = 81$; $D^2(X) = 206.8056$; $D^2(Y) = 109.333$;

$r = 0.5298$:

Egy $0.95 = 1-p$ megbízhatósági szinten a kor és a vércukorszint kapcsolatának a feltárásához felállítjuk a null, illetve ellenhipotézist:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

majd a (2.5.2) képlettel kiszámítjuk a t_ρ próbastatisztikát: $t_\rho = 1.2494$. Mivel $1 - p = 0.95$ szinten $t_p = 2.7764$ ($df = 6 - 2 = 4$) a $t_\rho \in (-t_p; t_p)$ (t_ρ benne van az elfogadási tartományban), tehát a nullhipotézist nincs okunk elvetni, vagyis a kor és a vércukorszint korrelálatlan változók.

Irodalomjegyzék

[1] Balogh Gábor, *Visual Basic és Excel programozás,* Computerbooks Kiadó, Budapest, 2001

[2] Baran S., et. al., *Bevezetés a matematikai statisztikába*, Debreceni Egyetem Kiadó, Debrecen, 2005.

[3] Baráth Cs., Ittzés A., Ugrósdy Gy.: *Biometria*, Mezőgazda Kiadó, Budapest, 1996.

[4] Bálint Gy., *Statisztika: elmélet és gyakorlat*, Scientia Kiadó, Kolozsvár, 2009.

[5] Ciucu G., Craiu V., Săcuiu I., *Probleme de statistică matematică*, Editura Tehnică, Bucureşti, 1974.

[6] Cseke V., *A valószínűségszámítás és gyakorlati alkalmazásai*, Dacia Könyvkiadó, Kolozsvár, 1982.

[7] Denkinger G., *Valószínűségszámítás,* Tankönynvkiadó, Budapest, 1978.

[8] Denkinger G., *Valószínűségszámítási gyakorlatok,* Tankönynvkiadó, Budapest, 1977.

[9] Jánosa A., *Adatelemzés számítógéppel*, Perfekt Kiadó, 2005.

[10] Kelly, J., *Microsoft Excel: utilizare*, Teora, Bucuresti, 2000.

[11] Korpás Attiláné (szerk.), *Általános statisztika I-II*, Nemzeti Tankönynvkiadó, Budapest, 1996.

[12] Köves P., Párniczky G., *Általános statisztika*, Közgazdasági és Jogi Könyvkiadó, Budapest, 1975.

[13] Manczel J, (szerk), *Statisztikai módszerek alkalmazása a mezőgazdaságban*, Mezőgazdasági Kiadó, Budapest, 1983.

[14] Nash. J.C., *Teaching statistics with Excel and other spreadsheets*, Computational Statistics and Data Analysis, 2008.

[15] Obádovics Gy., *Valószínűségszámítás és matematikai statisztika*, Scolar Kiadó, Budapest, 2001.

[16] Perczelné Zalai M., *Biometria a kertészetben*, Kertészeti Egyetem, Budapest, 1996.

[17] Podmaniczky L., Illés B.Cs., *A számítógépes tervezés lehetőségei a mezőgazdaságban*, Pannon Agrártudományi Egyetem, Gödöllő, 1997.

[18] Sachs, L., *Angewandte statistik*, 4.kiad, Springer Verlag, Berlin-Heidelberg-New York, 1974.

[19] Váradi Zs., *Az Excel fortélyai nem csak haladóknak*, Műszaki Könyvkiadó, Budapest, 1997.