

8. tétel: Adatsokaságok jellemzői, a valószínűségszámítás elemei

ADATSOKASÁGOK JELLEMZŐI – STATISZTIKA:

Statisztika: Tömegesen előforduló jelenségek és folyamatok számbavétele, az adatok elemzése, vizsgálati módszerek.

Leíró statisztika: A leíró statisztika azzal foglalkozik, hogy egy adott, meghatározott elemekből álló információhalmazt kiértékeljen. Ezek legtöbbször számok. Egy teljes, ismert csoportból gyűjtünk adatokat.

Matematikai statisztika: Ismert minták alapján következtetünk a teljes csoportra.

Vizsgált csoport: maga a statisztikai sokaság. Egyedeinek száma a statisztikai sokaság mérete.

Adatgyűjtés: - elsődleges adat – önálló gyűjtéssel:

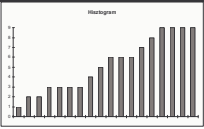
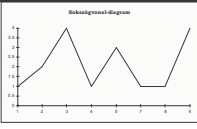
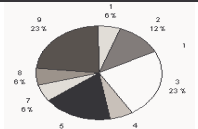
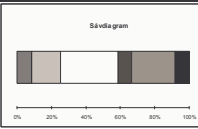
- teljes mintavétel
- reprezentatív minta (amivel foglalkozhat a matematikai statisztika)
- másodlagos adat – meglévő adatbázisból

Adatfajták: névleges adat (pl. nem, hajszín); rendezhető adat (pl. versenyeredmény, osztályzat); mérhető (mennyiségi) adat (pl. tömeg, db, hossz, jövedelem)

Gyakoriság: Az adat előfordulásainak száma a sokaságban. Az adatokat és gyakoriságukat táblázatba foglalva megkapjuk a gyakorisági eloszlást. Ha a közli értékeket egy csoportba soroljuk, osztályokat kapunk, melyeket táblázatba rendezve osztályközös gyakorisági eloszlást kapunk.

Relatív gyakoriság: A gyakoriság és a méret hányadosa. $\frac{k}{n}$, ahol $0 \leq \frac{k}{n} \leq 1$. A relatív gyakoriságot néha százalékban adják meg.

Adatok ábrázolása: Az adatok ábrázolásánál legtöbb esetben az adathalmazban való előfordulási arányt szokták ábrázolni (relatív gyakoriság).

Oszlopdiaqram	Vonaldiagram	Kördiagram	Sávdiaqram	Egyéb
				- kúp - henger - piramis - pont - terület - felület - buborék - árfolyam - sugár - pókháló - piktogram - kartogram
Az oszlopok magassága arányos az adat nagyságával (így lehet negatív is). Jó összehasonlításhoz, időbeli változáshoz.	Az adatok össze vannak kötve egy törtvonallal.	A relatív gyakoriság a körívek középponti szögével arányos. 100%-360°.	A részsávok hossza arányos a megjelenített adat relatív gyakoriságának nagyságával.	
Nem jó, ha az adatok között szélsőségek vannak, vagy ha nagyon egyformák.	Kiemeli a változások mértékét, mert meredekség dominál	Jól látható a részek egymáshoz- az egészhez viszonyított értéke.	A rész és egész viszonya jól látszik, de a részek egymáshoz képesti viszonya nem..	
Hisztogram: adatok helyett a hozzájuk tartozó gyakoriságot ábrázolják	Megtévesztő is lehet, mert folytonosságot sugall.	Trébeli változata a torta- diagram mely a perspektíva miatt manipulációra ad lehetőséget.	Térbeli változatánál a perspektívaváltozás manipulációra ad lehetőséget.	

Fontos a *grafikonkészítés*nél az értéktengelyek helyes beállítása, a megfelelő típusválasztás, a színezés ill. minta jó megválasztása, grafikonnak legyen címe (rövid, hatásos). A tényezők helytelen megválasztása ill. különböző optikai csalások manipulációra adnak lehetőséget.

Statisztikai mutatók, középértékek:

(adattömörítés reprezentatív számértékekkel)

Statisztikai mutatók: Az adatábrázolásból kapott, az adathalmazra jellemző számok.

Középértékek: Összefoglaló neve a módusznak, mediánnak és az átlagnak

Módusz:

A számsokaságban leggyakrabban előforduló számot a számsokaság módusának nevezzük. Ha több ilyen van, akkor azok mindegyike módusz.

Tehát a leggyakoribb adat. Jele: M_0

A módusz akkor használható, ha az adatok közül kiemel egyet, nem szerencsés a használata, ha több adat is közel azonos gyakorisággal emelkedik ki a sokaságból.

Medián:

A nagyságrendi sorba rendezett adatok közül a középső (ha két középső van, akkor ezek átlaga adja a mediánt). Jele: M_e .

Látható, hogy ugyanannyi adat nem nagyobb a mediánnál, mint amennyi nem kisebb.

Csak rendezhető adatoknál használható.

Kvartilisek:

alsó kvartilis: a legkisebb és a medián között középen elhelyezkedő adat számértéke a rendezett mintában. Tehát a medián alatti adatok mediánja. Jele: Q_1

felső kvartilis: a medián és a legnagyobb adat között elhelyezkedő adat számértéke a rendezett mintában. Tehát a medián feletti adatok mediánja. Jele: Q_2

A Q_1 és Q_3 közötti tartomány az *interkvartilis tartomány*.

A kvartilisek a szóráshoz hasonlóan az adatok szóródásáról tájékoztatnak, elsősorban ferde eloszlás esetén érdemes őket használni. (A kvartilisek mutatják a ferdeséget, az szórás nem).

Számtani közép – átlag ($\bar{A}$):

A számsokaság összegét elosztjuk a számsokaság darabszámával, ekkor a számsokaság *átlagát vagy számtani közepét* kapjuk.

$$\text{jele: } \bar{X} = \frac{\text{adatok összege}}{\text{darabszám}} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Csak mérhető adatoknak lehet átlaga.

A kiugró, szélsőséges adatok nagyon befolyásolják.

Egyéb jellemzők – szóródási mutatók:

A középértékek megadása önmagában nem elég. A szórás, szóródás azt adja meg, hogy az adatok mennyire tömörülnek a középértékek körül – minél kisebb a szóródás, annál jobban jellemzőek a középértékek.

Terjedelem:

A mintában előforduló legnagyobb és legkisebb adat különbsége (mérhető adatoknál). Tehát: $t = x_{\max} - x_{\min}$.

Ha ez kicsi, akkor gyakorlatilag bármelyik középérték jól jellemzi az adathalmazt, ha pedig nagy, akkor nem lehet eldönteni, hogy mi mennyi információt szolgáltat. A másik hibája, hogy nagyon érzékeny a kiugró adatokra.

Az interkvartilis terjedelem ($it = Q_3 - Q_1$) viszont már nem érzékeny a szélsőségekre.

Átlagos abszolút eltérés:

Def.: Az egyik középértéktől (Z) vett eltérések abszolútértékeinek átlaga.

$$\text{Képlet: } \overline{\Delta X_Z} = \frac{|x_1 - Z| + |x_2 - Z| + \dots + |x_n - Z|}{n}$$

$\overline{\Delta X_Z}$ minimális, ha Z=medián.

Ha nem mondjuk meg mi a Z, akkor azt a mediánnak kell tekinteni.

Átlagos négyzetes eltérés - szórásnégyzet:

Def.: Az egyik középértéktől vett eltérések négyzeteinek átlaga.

$$\text{Képlet: } \sigma_Z^2 = \frac{(x_1 - Z)^2 + (x_2 - Z)^2 + \dots + (x_n - Z)^2}{n}$$

σ_Z^2 minimális, ha a választott középérték, Z a számtani közép. Ezért ez az általános középérték, ha szórásnégyzetet számolunk.

Szórás:

$$\text{a szórásnégyzet gyöke: } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{n}}$$

Érzékeny a kiugró értékekre

Csebisev – törvény:

Az átlagtól az adatok legfeljebb 25%-a térhet el a szórás kétszeresénél, legfeljebb 10-11%-a térhet el a szórás háromszorosánál, és 5-6%-a a szórás négyszeresénél jobban (ún. normális eloszlásnál). Természetesen maga a Csebisev-egyenlőtlenség durva becslésen alapszik, ezért gyakorlatilag a szórás négyszeresénél jobban nem térnek el az adatok az átlagtól, de az 5-6%-ot biztosan mondhatjuk bármilyen adathalmaz esetén. Ez konkrét adathalmazokra vonatkoztatva az úgynevezett empirikus Csebisev-törvény, vagy az átlag körüli szórás empirikus törvénye.

Alkalmazások (teljes tételhez)

- adatsokaságok jellemzése
- mérési eredmények kiértékelése (pl. sport: NBA, vagy fizikai, kémiai méréseknél)
- iskolai statisztikák készítése: jegyek átlaga, osztályok átlaga, hiányzások száma
- statisztikák, grafikonok, folyamatok elemzése
- közvéleménykutatások, kutatások kiértékelése
- reklámok hatékonyságát növeli egy belerajzolt grafikon ☺
- szavazások
- szerencsejátékok (lottó, totó, tippmix)
- biztosítások
- időjáráselőrejelzés
- betegségekhez való tesztek
- alkatrészek élettartamának becslése (tartalékolási problémák)

A VALÓSZÍNŰSÉGSZÁMÍTÁS ELEMEI:**Definíciók:**

- Véletlen jelenség: Olyan folyamat, melynek kimenetelét az adott körülmények között nem tudjuk meghatározni.
 - Valószínűségi számítás: Véletlen tömegjelenségek matematikai modellezése.
 - *Kísérlet*: Olyan folyamat melynek kimenetele véletlenszerű.
 - *Eseménytér*: Egy kísérlet összes lehetséges kimenetele. Jele: Ω .
Példa: $\Omega_{\text{érmédobás}} = \{\text{fej}; \text{írás}\}$; $\Omega_{\text{kockadobás}} = \{1; 2; 3; 4; 5; 6\}$
 - *Elemi esemény*: Az eseménytér egy eleme, a kísérlet egy kimenetele.
Jelük általában nagy betűk: A; B stb.
Példa: $A = \{1\}$ 1-est dobtunk a dobókockával.
 - Eseménytér számossága: hány darab elemi eseményből áll. Példa: $|\Omega_{\text{kockadobás}}| = 6$.
 - *Esemény*: Az eseménytér egy részhalmaza.
Példa: $A = \text{párosat dobunk egy kockával} = \{2; 4; 6\}$. Tehát $A \in \Omega_{\text{kockadobás}}$. Az esemény akkor következik be, ha a kísérlet végeredménye benne van ebben a részhalmazban ($\Omega \subset \Omega$ és $\emptyset \subset \Omega$ - minden halmaznak részhalmaza önmaga és az üreshalmaz).
Biztos esemény: ha az egész eseményteret felsoroljuk.
Példa: $\Omega_{\text{kockadobás}} = \{1; 2; 3; 4; 5; 6\}$
Lehetetlen esemény: $\emptyset$
 - *Események összessége*: Az összes részhalmaz száma. Jele: $\mathcal{A}$ (írott nagy A betű).
Példa: - $\Omega_{\text{érmédobás}} = \{I; F\}$ $\mathcal{A}_{\text{érmédobás}} = \{\emptyset; \{I\}; \{F\}; \{I; F\}\}$
- $\Omega_{\text{kockadobás}} = \{1; 2; 3; 4; 5; 6\}$ $|\mathcal{A}_{\text{kockadobás}}| = 2^6 = 64$
Általánosan: $|\Omega| = n$ $n \in \mathbb{N}^+$
 $|\mathcal{A}| = 2^n$
- Tétel: n elemű halmaznak 2^n darab részhalmaza van.

Műveletek eseményekkel – esemény algebra:

1. *Komplementer*: $\bar{A}$ bekövetkezik, ha A nem következik be. Jele: $\bar{A} = \Omega \setminus A$.
Tulajdonságok: $\overline{\overline{A}} = A$; $\overline{\Omega} = \emptyset$; $\overline{\emptyset} = \Omega$
2. *Unió*: Két esemény uniója (összege) bekövetkezik, ha legalább az egyik bekövetkezik.
Jele: $A \cup B$; $A+B$
3. *Metszet*: Két esemény metszete bekövetkezik, ha mindkét esemény bekövetkezik.
Jele: $A \cap B$; $A \cdot B$
4. *Különbség*: Két esemény különbsége bekövetkezik, ha az egyik bekövetkezik, de a másik nem (pl. ha A bekövetkezik, akkor B nem következik be).
Jele: $A \setminus B$; $A - B$ (ebben az esetben A bekövetkezik, de B nem).

Műveletei tulajdonságok (tök ugyanaz, mint a halmazoknál):

1. Az unió- és metszetképzés kommutatív és asszociatív műveletek.
2. Események különbsége nem kommutatív művelet: $A \setminus B \neq B \setminus A$.
3. Műveletek lehetetlen eseménnyel: $A \cup \emptyset = A$; $A \cap \emptyset = \emptyset$; $A \setminus \emptyset = A$; $\emptyset \setminus A = \emptyset$
4. Esemény önmagával: $A \cup A = A$; $A \cap A = A$; $A \setminus A = \emptyset$
5. Műveletek biztos eseménnyel: $A \cup \Omega = \Omega$; $A \cap \Omega = A$; $A \setminus \Omega = \emptyset$; $\Omega \setminus A = \bar{A}$
6. De Morgan azonosságok:

$$\overline{A \cup B} = \bar{A} \cap \bar{B}$$

$$\overline{A \cap B} = \bar{A} \cup \bar{B}$$

Kolmogorov-féle valószínűségi modell:

Adott Ω (eseménytér) és $\mathcal{A}$ (események részhalmazainak összessége). Minden $\mathcal{A}$ –beli eseményhez rendeljünk hozzá egy valós számot – ez az esemény valószínűsége $P(A)$ – melyre az alábbiak teljesülnek:

Axiómák:

1. $P(A) \geq 0 \quad \forall A \in \mathcal{A}$ (bármely esemény valószínűsége nemnegatív)
2. $P(\Omega) = 1$ (a biztos esemény valószínűsége 1)
3. $P(A \cup B) = P(A) + P(B)$ ha $A; B$ események diszjunkt, azaz $A \cap B = \emptyset$ (tehát két esemény diszjunkt, ha metszetük üreshalmaz, nincs közös elemük).

Tételek:

1. $P(\emptyset) = 0$
Bizonyítás: $P(A \cup \emptyset) \stackrel{(3.ax)}{=} P(A) + P(\emptyset) \rightarrow P(A) = P(A) + P(\emptyset) \rightarrow 0 = P(\emptyset)$
2. $P(\bar{A}) = 1 - P(A)$
3. Szita-formula:
 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Klasszikus kombinatorikus valószínűségi mező /modell:

Az Ω (eseménytér) véges sok eseményből áll és ezek valószínűség megegyezik.

Klasszikus modelleknél: $P(A) = \frac{\text{kedvező esetek száma}}{\text{összes eset száma}} = \frac{|A|}{|\Omega|}$.

Csak akkor alkalmazható, ha az elemi események valószínűségei egyenlők (különben ugye jön Brad Pitt)

Geometriaival valószínűség:

A klasszikus kombinatorikus valószínűségi modell nem alkalmazható, ha például $|\Omega| = \infty$ (az eseménytér számossága végtelen).

Geometriaival valószínűség:

Tetszőleges A síkidomba esés valószínűsége arányos az A síkidom területével (egy dimenzióban a hosszal).

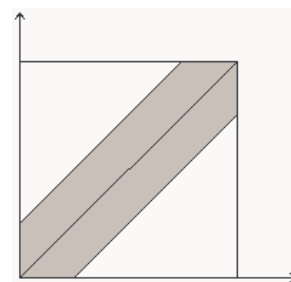
- Példák: 1. céllövölde  $P(A - \text{ba lő}) = \frac{T(A)}{T(\Omega)}$

2.  x: véletlenül választott szám $[a; b]$ -ből

$$P(c < x < d) = \frac{d - c}{b - a}$$

3. Bertrand paradoxon (körbe választott húr

4. Példa: mindenki fél órát a vár a másakra

**Teljes esemény rendszer:**

$A_1; A_2; \dots; A_n$ telje esemény rendszer, ha $A_1 \cup A_2 \cup \dots \cup A_n = \Omega$ és $A_i \cap A_j = \emptyset \quad \forall i; j$.

Példa: kockadobás: $A_1 = \{1; 2\}$, $A_2 = \{3\}$, $A_3 = \{4; 5; 6\}$.

Események függetlensége:

- Def.: A és B esemény független, ha metszetük valószínűsége az események valószínűségének szorzata: $P(A \cap B) = P(A) \cdot P(B)$.
- Példa: A: kockával 6-ost dobunk ; B: érmével fejet dobunk . Ez a két esemény független.

Feltételes valószínűség:

- Def.: Mekkora A esemény valószínűsége, ha B biztosan bekövetkezik.

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

- Ha A és B független események, akkor:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

- Példák: Szűrővizsgálatos valószínűségek; mintavételi feladatok.

Teljes valószínűség tétel:

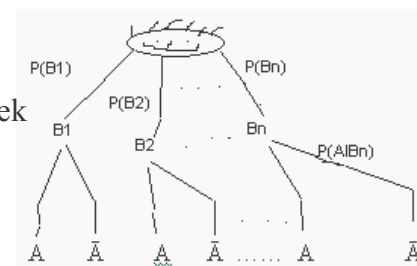
$B_1; B_2; \dots; B_n$ az összes esetet lefedő, egymást kizáró események

$$B_i \cap B_j = \emptyset$$

$$\text{Azaz: } \bigcup_{i=1}^n B_i = B_1 \cup B_2 \cup \dots \cup B_n = \Omega \quad \left. \vphantom{\bigcup_{i=1}^n B_i} \right\} \text{TER}$$

Adott $B_1; B_2; \dots; B_n$ teljes esemény rendszer és $A \subset \Omega$ esemény. Ekkor:

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n) = \\ &= P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + \dots + P(B_n) \cdot P(A|B_n) \end{aligned}$$

**Bayes-tétel:**

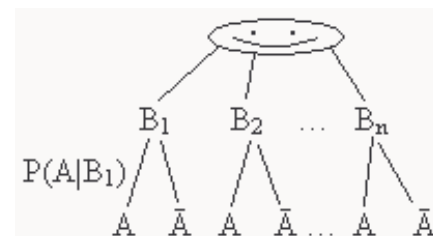
Thomas Bayes: 17. századi angol pap.

Tétel:

$$P(B_1|A) = \frac{P(B_1) \cdot P(A|B_1)}{P(B_1) \cdot P(A|B_1) + \dots + P(B_n) \cdot P(A|B_n)}$$

számláló: „A” „B₁”-el együtt teljesül

nevező: „A” összes valószínűsége

**Eloszlások:**

- Az eseménytér összes elemét felsoroljuk a hozzájuk tartozó valószínűséggel
- Jele: ξ (kszi) vagy η (éta).
- Példa: ξ : érmedobás (0-fej; 1-írás)

x_i	p_i
0	$\frac{1}{2}$
1	$\frac{1}{2}$

- ξ (eloszlás) várható értéke (átlaga): $E(\xi) = M(\xi) = x_1 \cdot p_1 + x_2 \cdot p_2 + \dots + x_n \cdot p_n = \sum x_i \cdot p_i$

Szórás:

- Szórásnégyzet: $D^2(\xi) \stackrel{\text{Def.}}{=} E[(\xi - E(\xi))^2] \stackrel{\text{Tétel}}{=} E(\xi^2) - [E(\xi)]^2$
- Szórás: $\sqrt{D^2(\xi)}$ a szórásnégyzet gyöke

Nevezetes eloszlások:**1. Indikátorváltozó:**

ξ : egy p valószínűségű esemény bekövetkezik (1-bekövetkezik; 0-nem következik be)

x_i	p_i
0	$1-p$
1	p

$$E(\xi) = 0 \cdot (1-p) + 1 \cdot p = p$$

$$D(\xi) = \sqrt{p \cdot (1-p)}$$

2. Diszkrét egyenletes eloszlás:

x_i	1	2	3	...	n
p_i	$1/n$	$1/n$	$1/n$	...	$1/n$

$$E(\xi) = \frac{n+1}{2}$$

3. Binomiális eloszlás:

ξ : „ n ” független kísérletből hányszor következik be egy „ p ” val-ű esemény.

x_i	p_i
0	$(1-p)^n$
1	$\binom{n}{1} \cdot p \cdot (1-p)^{n-1}$
2	$\binom{n}{2} \cdot p^2 \cdot (1-p)^{n-2}$
$\vdots$	$\vdots$
n	p^n

- számít a sorrend

- általánosan:

$$P(\xi = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}, \text{ ahol } 0 \leq k \leq n \text{ és } k \in \mathbb{Z}$$

$$E(\xi) = n \cdot p \quad ; \quad D(\xi) = \sqrt{n \cdot p \cdot (1-p)}$$

példa: Visszatevésees mintavétel: N dologból K jó $\rightarrow$ visszatevéssel húzunk n darabot. Mi a valószínűsége, hogy k darab jó köztük?

$$\begin{aligned} P(k \text{ db jó}) &= \frac{\binom{n}{k} \cdot K^k \cdot (N-K)^{n-k}}{N^n} = \binom{n}{k} \cdot \frac{K^k}{N^k} \cdot \frac{(N-K)^{n-k}}{N^{n-k}} = \\ &= \binom{n}{k} \cdot \left(\frac{K}{N}\right)^k \cdot \left(\frac{N-K}{N}\right)^{n-k} = \binom{n}{k} \cdot \left(\frac{K}{N}\right)^k \cdot \left(1 - \frac{K}{N}\right)^{n-k} \end{aligned}$$

Ha $\frac{K}{N} = p$, akkor megkapjuk a binomiális eloszlás képletét

Visszatevésees mintavételnél, K és N konkrét értékét nem szükséges ismernünk, csak arányuk kell.

4. Poisson-eloszlás:

Adott idő alatt bekövetkezett események számát közelíthetjük ezzel.

Képlet: $P(\xi = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$, ahol k : ahányszor bekövetkezik az esemény és

$\lambda \in \mathbb{R}^+$ a paraméter.

Ekkor: $E(\xi) = \lambda$

$$D(\xi) = \sqrt{\lambda}$$

Példa: 1 óra alatt átlagosan 23 hívás érkezik. Mennyi annak a valószínűsége, hogy 2 hívás fog érkezni 1 óra alatt?

Ekkor: $\lambda = 23$ és $k = 2$, ezeket kell a képletbe behelyettesíteni.