29.B Pravděpodobnost a statistika

Náhodný pokus – pokus, jehož výsledek záleží i při dodržení předem stanovených podmínek na náhodě.

$\Omega=\{\omega_1; \omega_2;\ldots; \omega_n\}$ … množina všech možných výsledků náhodného pokusu

Náhodný jev A je podmnožinou množiny Ω ..... $A \subseteq \Omega$

Klasická (Laplaceova) definice pravděpodobnosti:

Nechť náhodný pokus splňuje předpoklady:

1) Všech možných výsledků je konečný počet
2) Všechny výsledky mají stejnou šanci na realizaci
3) Všechny výsledky se navzájem vylučují (tj. žádné dva nemohou nastat současně)
4) Jeden z výsledků jistě nastane.

Pak pravděpodobností jevu A se nazývá číslo $P(A)=\frac{m}{n}$, kde *n* je počet všech možných výsledků (počet prvků Ω) a *m* je počet výsledků příznivých jevu A (počet prvků A).

Některé vlastnosti pravděpodobnosti:

1. a) *Jistý jev* … A= Ω => $P(A)=\frac{n}{n}=1$

b) *Nemožný jev* … A= Ø => $P(A)=\frac{0}{n}=0$

c) *Náhodný jev* … $A \subseteq \Omega$ => $0 \le P(A) \le 1$

2. **Opačný jev** k jevu A je takový jev A‘, který nastává právě tehdy, když nenastal jev A.

Tedy: A‘ = Ω – A (přesněji $A \cap A' = \emptyset \wedge A \cup A' = \Omega$)

Pak $P(A')=\frac{n-m}{n}=1-\frac{m}{n}=1-P(A)$

3. **Sjednocení jevů** A, B je jev $A \cup B$, který nastane právě tehdy, když nastane aspoň jeden z jevů A nebo B.

a) Platí-li, že se jevy A, B navzájem vylučují (tj. $A \cap B = \emptyset$), pak $P(A \cup B) = P(A) + P(B)$

b) Pokud se jevy A, B navzájem nevylučují (tj. $A \cap B \neq \emptyset$), pak $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4. **Průnik jevů** A, B je jev $A \cap B$, který nastane právě tehdy, když nastane jev A a zároveň jev B. Jestliže A, B jsou **nezávislé jevy**, pak $P(A \cap B) = P(A).P(B)$

Statistická (zobecněná) definice pravděpodobnosti:

Pravděpodobnost P(A) jevu A je určena přibližně jeho relativní četností při dostatečně velkém počtu opakování náhodného pokusu.

Nechť $\Omega = \{\omega_1; \omega_2;\ldots; \omega_n\}$ je množina všech možných výsledků náhodného pokusu a $p_1, p_2, \ldots, p_n$ jsou jejich relativní četnosti (tzn. $\sum_{i=1}^{n} p_i = 1$).

Pak $P(A) = \sum_{i=1}^{k} p_{j_i}$, kde p_{j_i} značí relativní četnost výsledku $\omega_{j_i} \in$ A (k je počet prvků A).

Podmíněná pravděpodobnost – pravděpodobnost jevu A podmíněnou jevem B určíme takto:

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

Bernoulliho schéma:

Nechť při n-násobném opakování náhodného pokusu je stále stejná pravděpodobnost zdaru *p* a pravděpodobnost nezdaru *q (*tedy *q = 1 – p)*. Pak pravděpodobnost jevu A_k, že zdar nastane v těchto n pokusech právě k-krát je dána vztahem: $P(A_k)=\binom{n}{k}.p^k.q^{n-k}$

Základy statistiky

Statistika se zabývá zkoumáním a zpracováním velkého množství dat souvisejících s hromadnými jevy.

Základní pojmy

- **statistický soubor** – konečná neprázdná množina objektů, které zkoumáme (např. obyvatelé Brna, obyvatelé ČR, rodinní příslušníci zaměstnanců určité firmy, dopravní nehody v určité oblasti za určité období, výrobky vyrobené v určité firmě za určité období ,…)
- **statistická jednotka** – prvek statistického souboru (např. jeden určitý obyvatel, jeden daný výrobek)
- **rozsah statistického souboru** – počet prvků statistického souboru
- **statistický znak** – společná vlastnost prvků statistického souboru, kterou zjišťujeme (např. věk, národnost, výše měsíčního příjmu, výška postavy, kvalita výrobku (vadný nebo bez vady), …)

  znak může být
  - ***kvantitativní*** (číselný) – např. počet obyvatel daného věku, výše škody při nehodě, …
  - ***kvalitativní*** (popsán slovy) – např. povolání, druh nemoci, příčina dopravní nehody, …

  Pozn. 1) *kvalitativní* znak může mít někdy více možností (např. příčin nehody může být víc) – pak se musí vybrat jedna, která je hlavní (ostatní mohou tvořit kategorii „jiné“)

  Pozn. 2) *nejjednodušší kvalitativní* znak je znak *alternativní* – dán jevem a jeho opakem – např. voják-nevoják, muž-žena, plavec-neplavec, prospěl-neprospěl, …

- **absolutní četnost** hodnoty znaku x_i – číslo n_i udávající počet prvků daného statistického souboru, které vykazují sledovanou hodnotu x_i, neboli udávající, pro kolik prvků souboru nabývá statistický znak určité hodnoty nebo rozmezí hodnot (např. kolik nezaměstnaných osob je evidováno v dané oblasti, kolik osob má měsíční příjem ve vybraném rozmezí, …)
- **relativní četnost** znaku – poměr $\frac{n_i}{n}$ absolutní četnosti dané hodnoty a rozsahu souboru

  Pozn. Relativní četnost se nejčastěji uvádí v procentech $\frac{n_i}{n}.100\ \%$

Statistické soubory rozdělujeme na
- **základní** (mohou mít pro zkoumání příliš velký rozsah)
- **výběrové** (část základního souboru, na němž se provádí zkoumání)

Charakteristiky statistického souboru

1. **Charakteristiky polohy** hodnot znaku jsou číselné hodnoty, které určitým způsobem charakterizují typickou hodnotu sledovaného znaku

   - **Aritmetický průměr** – součet všech hodnot zjištěných znaků dělených jejich počtem.
     $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ nebo tzv. vážený průměr $\bar{x} = \frac{1}{n}\sum_{j=1}^{r} x_j n_j$
   - **Modus Mod(x) – hodnota** znaku s největší četností.
   - **Medián Med(x) – je** prostřední hodnota znaku, jsou-li hodnoty uspořádány podle velikosti.

     $Med(x) = x_{\frac{n+1}{2}}$, je-li n liché, $Med(x) = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)$, je-li n sudé.

Geometrický průměr $\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

Př. 1: V souboru A byl sledován údaj o počtu dětí v 13 rodinách. Rozsah souboru je n = 13 (liché):

počet dětí	0	1	2	3	4	5	6	7
četnost	2	4	3	2	1	1	0	0

Aritmetický průměr $\bar{x} = \frac{1}{13}\sum_{i=1}^{13} x_i = \frac{0+0+1+1+1+1+2+2+2+3+3+4+5}{13} = \frac{25}{13} = 1{,}923$

(Vážený průměr $\bar{x} = \frac{1}{13}\sum_{j=1}^{8} x_j n_j = \frac{0.2+1.4+2.3+3.2+4.1+5.1+6.0+7.0}{13} = \frac{25}{13} = 1{,}923$)

Modus Mod(x) = 1 (má četnost 4, což je nejvíce)
Medián Med(x) = 2 (7. rodina).

Př. 2: V souboru B byl sledován údaj o počtu dětí ve 14 rodinách. Rozsah souboru je n = 14 (sudé):

počet dětí	0	1	2	3	4	5	6	7	8
četnost	2	5	3	1	1	0	0	1	1

Aritmetický průměr $\bar{x} = \frac{1}{14}\sum_{i=1}^{14} x_i = \frac{0+0+1+1+1+1+1+2+2+2+3+4+7+8}{14} = \frac{33}{14} = 2{,}357$

(Vážený průměr $\bar{x} = \frac{1}{14}\sum_{j=1}^{9} x_j n_j = \frac{0.2+1.5+2.3+3.1+4.1+5.0+6.0+7.1+8.1}{14} = \frac{33}{14} = 2{,}357$)

Modus Mod(x) = 1 (má četnost 5, což je nejvíce)
Medián Med(x) = 1,5 (aritmetický průměr ze 7. a 8.rodiny).

2. Charakteristiky variability

Každá charakteristika polohy je číslo, kolem něhož jednotlivé hodnoty znaku kolísají. Charakteristiky variability vyjadřují „velikost“ onoho kolísání.

- **Rozptyl** s_x^2 se definuje jako průměr druhých mocnin odchylek od aritmetického průměru

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{j=1}^{r}\left(x_j^* - \bar{x}\right)^2 n_j$$

- **Směrodatná odchylka** $s_x = \sqrt{s_x^2}$

- **Variační koeficient** v_x – podíl směrodatné odchylky a aritmetického průměru – udává se v procentech $v_x = \frac{s_x}{\bar{x}}.100\,\%$

- **Koeficient korelace** r - souvisí s tím, že se velmi často zkoumá, zda a jak jsou na sobě závislé dva znaky *x* a *y*. Koeficient korelace vyjadřuje míru vzájemné závislosti těchto znaků *x* a *y*.

$$r = \frac{k}{s_x.s_y},$$ kde $k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$,

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad , \quad s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Pozn.: Vždy platí: $|r| \le 1$. Čím víc se hodnota *r* blíží k **1**, tím považujeme závislost *x* a *y* za silnější.