7.3 Statistika

Předmětem matematické statistiky jsou hromadné jevů.
Cílem je vytvoření spolehlivých závěrů o povaze sledovaných jevů na základě informací, které o nich máme.

Definice 7.3.1
Statistický soubor je skupina objektů, jejichž vlastnosti sledujeme a studujeme.
Prvek statistického souboru je každý jednotlivý objekt.
Rozsah statistického souboru je počet jeho prvků. Značíme obvykle n.

Prvky statistického souboru musí mít některé vlastnosti společné, což je základní podmínkou při tvoření statistických souborů. Čím více je společných vlastností, tím více je soubor homogenní.

Definice 7.3.2
Znakem nazýváme vlastnosti vyšetřované společné všem prvkům.
Nabývanou jakost znaku pro konkrétní prvek nazveme **hodnotou znaku**.

Znaky dělíme na:

a) Kvalitativní – udávané kvalitou

Speciálně alternativní znak nabývá pouze dvou kvalitativních hodnot znaku.

b) Kvantitativní – zjišťované měřením a udávané číselnou hodnotou

Hodnota sledovaného znaku se u jednotlivých prvků obecně mění. Znak je tedy proměnnou veličinou. Nelze-li na základě určité zákonitosti předem stanovit hodnotu znaku určitého prvku mluvíme o **náhodné veličině**. Náhodnou veličinu (proměnnou) značíme velkými písmeny $X, Y, \ldots$ a konkrétní nabývané hodnoty malými písmeny $x, y, \ldots$.

Úplnou informaci, většinou značně nepřehledně, určuje a podává statistická tabulka základních dat.

Proto data třídíme a určujeme charakteristiky statistického souboru.

Třídění dat

Definice 7.3.3
Četnost hodnoty znaku ν_i je počet opakování i – té hodnoty znaku v statistickém souboru.

Relativní četnost hodnoty znaku je podíl četnosti znaku a rozsahu souboru $f_i = \frac{\nu_i}{n}$

respektive. $f_i = \frac{\nu_i}{n} \cdot 100\%$.

Věta 7.3.1
Pro daný statistický soubor a jeho četnosti platí:

a) $\sum_{i=1}^{k} \nu_i = n$ b) $\sum_{i=1}^{k} f_i = 1$, kde k je počet různých hodnot znaku.

Tabulka četností a relativních četností pro soubory s malým počtem různých hodnot znaku

x_i	9,36	10,50	10,94	10,24	10,90	11,68	10,21	11,12	10,80	9,93
ν_i										
f_i										

Často je používáno rozdělení hodnot znaku na třídy, kterých je k a které jsou určeny hranicí a délkou, přičemž každou třídu pak reprezentujeme třídním znakem.
Hranice udává, kterou nejmenší a největší hodnotu do třídy zahrnujeme.
Délka h intervalů bývá konstantní, kvůli dalšímu zpracovávání a obvykle platí

$$h = 0{,}08 \cdot (x_{\max} - x_{\min}) \text{ nebo } h < \frac{x_{\max} - x_{\min}}{12} < 2h \ .$$

Počet tříd určíme vhodným způsobem z hodnoty zlomku $k = \frac{x_{\max} - x_{\min}}{h}$

Třídní znak pak je $x_j = \frac{x_{j,\max} + x_{j,\min}}{2}$, kde j značí pořadové číslo třídy.

Poznámka
Jako triviální rozdělení na třídy můžeme chápat situaci, kdy základní data nabývají tak mála různých hodnot, že každá z nich tvoří samostatnou skupinu.

Někdy uvažujeme též kumulativní absolutní četnost N_j a kumulativní relativní četnost F_j.
Kumulativní četnosti vyjadřují součet všech četností od prvé do j – té včetně.

Tabulka četností a relativních četností pro netriviální rozdělení na třídy

x_j	9,36	10,50	10,94	10,24	10,90	11,68	10,21	11,12	10,80	9,93
ν_j										
f_j										
N_j										
F_j										

Grafické znázornění rozdělení četností

- Kruhový diagram

– četnostem odpovídají kruhové výseče s poměrným středovým úhlem
Kartézské grafy
- na vodorovnou osu vynášíme hodnoty sledovaného znaku, na svislou osu četnost

- Spojnicový diagram – jednotlivým četnostem odpovídají body pospojované lomenou čarou
- Sloupcový diagram – osa sloupců prochází třídní hodnotou znaku a šířka sloupců se volí libovolně se zřetelem na celkový vzhled diagramu

- Histogram - osa sloupců prochází třídní hodnotou znaku a šířka sloupců délce třídy

Často se u kartézských grafů na svislou osu vynáší relativní četnost.

Věta 7.3.2
Celková plocha vymezená obsahem sloupců histogramu relativních četností je
$P = \sum_{j=1}^{k} h \cdot f_j = h \cdot \sum_{j=1}^{k} f_j = h$ a pokud $h = 1$ činí velikost plochy $P = 1$.

7.3.1 Charakteristiky statistického souboru

Veličiny udávající stručnou informaci o hodnotách kvantitativního znaku statistického souboru. Rozdělujeme je na charakteristiky polohy a variability.

Charakteristiky polohy

Tyto veličiny určují střední hodnotu zkoumaného znaku.

Aritmetický průměr

$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ respektive ve váženém tvaru $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{l} \nu_i \cdot x_i$

a) matematické vyjádření a stanovení aritmetického průměru je jednoduché a snadno použitelné
b) výpočet aritmetického průměru je založen na všech zjištěných hodnotách
c) součet odchylek jednotlivých hodnot od aritmetického průměru je roven nule
$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$
d) součet čtverců odchylek jednotlivých hodnot od aritmetického průměru je menší než součet čtverců odchylek jednotlivých hodnot od jakékoli jiné hodnoty a
$\sum_{i=1}^{n} (x_i - \bar{x})^2 < \sum_{i=1}^{n} (x_i - a)^2$
e) aritmetický průměr je ovlivňován krajními naměřenými hodnotami

Geometrický průměr

$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$ respektive ve váženém tvaru $\bar{x}_G = \sqrt[n]{\prod_{i=1}^{l} x_i^{\nu_i}}$, pro nezáporná čísla x_i

Harmonický průměr

$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$ respektive ve váženém tvaru $\bar{x}_H = \frac{n}{\sum_{i=1}^{l} \nu_i \frac{1}{x_i}}$

Věta 7.3.3

Pro libovolná nezáporná reálná čísla $x_i, i = 1,2,3,...,n$ a jejich průměry platí

$x_{\max} \geq \bar{x} \geq \bar{x}_G \geq \bar{x}_H \geq x_{\min}$

Modus
Modus je dán hodnotou znaku s největší četností

Medián
Medián je dán hodnotou středního prvku statistického souboru, uspořádaného podle velikosti. Pokud je rozsah souboru lichý jedná se o prostřední prvek, pokud je sudý, určíme aritmetický průměr dvou středních prvků.

Poznámka
Nejběžnější je určeni aritmetického průměru. Geometrický a harmonický průměr se určuje, pokud má specifický význam. Modus a medián lépe charakterizují soubory s několika výrazně odlišnými hodnotami znaku (hrubé chyby).

Charakteristiky variability

Tyto veličiny podávají informaci, jak jsou jednotlivé pozorované hodnoty ve sledovaném souboru rozptýleny.

Variační rozpětí
$R = x_{\max} - x_{\min}$

Střední odchylka
$$\Delta\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Rozptyl
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Z čehož lze odvodit $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$

Směrodatná odchylka
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$
Směrodatná odchylka má stejný rozměr jako měřený znak a jeho střední hodnota.

Variační koeficient
$$v = \frac{s}{\bar{x}}$$

Poznámka
Nejběžnější je určení směrodatné odchylky.

Při ručním zpracování můžeme použít pro kontrolu správnosti numerických výpočtů tzv. Charlierův test: $\sum_{i=1}^{n}(x_i+1)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i + n$

Poznámka
Pro vyhodnocení například přijímacích testů různých obtížností se často používají kvantily.

Definice
Nechť $F(x)$ je distribuční funkce (viz Statistika 2) spojité náhodné proměnné X a nechť dané číslo $p \in (0,1)$. Pokud je Q_p řešením rovnice $F(Q_p) = p$ tj. $P(X < Q_p) \leq p$, potom Q_p nazýváme p - krát procentním kvantilem.

Speciální označení
Medián je kvantil rozdělující statistický soubor na dvě stejně početné množiny, tj. $med(X) = Q_{0,5}$.
Kvartil
Dolní kvartil (1.) kvartil je určen $Q_{0,25}$ a horní (3.) kvartil $Q_{0,75}$.

Decil
Decil dělí statistický soubor na desetiny. Jako k - tý decil označujem $Q_{0,k}$, $k \in \{1,2,3,...,9\}$.

Percentil
Percentil dělí statistický soubor na setiny. Jako k - tý percentil označujem $Q_{0,k}$, $k \in \{1,2,3,...,99\}$.

Hodnoty kvantilů představují charakteristiky polohy, přičemž při známém rozsahu souboru lze určit i absolutní pořadí.

Pokud užijeme kvantilů k určení charakteristik variability, stanovujeme:

Mezikvartilové rozpětí	$Q_{0,75} - Q_{0,25}$
Mezidecilové rozpětí	$Q_{0,9} - Q_{0,1}$
Mezipercentilové rozpětí	$Q_{0,99} - Q_{0,01}$