11. P o p i s n á s t a t i s t i k a

11.1. Poznámka: Při statistickém zkoumání nás zajímají hromadné jevy a procesy, u kterých zkoumáme zákonitosti, které se projevují u velkého počtu prvků. Prvky zkoumání nazýváme *statistické jednotky.* Sledujeme vlastnosti statistických jednotek, které nazýváme *statistické znaky* nebo stručněji *veličiny (variable).* Souhrn znaků a veličin tvoří *data.* Při zkoumání používáme dva základní druhy statistiky, *popisnou statistiku* a *interferenční statistiku.*

Popisná statistika zjišťuje a sumarizuje informace, zpracovává je ve formě grafů a tabulek a vypočítává jejich číselné charakteristiky jako průměr, rozptyl percentily, rozpětí a pod.

Interferenční statistika činí závěry na základě dat získaných z šetření provedených pro vybraný soubor respondentů. Analyzuje tyto závěry a predikuje z nich závěr pro celý soubor. (Volební průzkum, průzkum trhu a pod.)

Při statistickém šetření máme k dispozici:

- **základní soubor** je soubor všech statistických jednotek;
- **výběrový soubor** je vybraná část ze základního souboru.

Rozsah základního (výběrového) souboru je počet jednotek v souboru.

Při vytváření souboru jednotek provádíme výběr ve tvaru *prostého náhodého výběru.*

11.2. Definice: Prostý náhodný výběr (*simple random sample*) je náhodný výběr ze základního souboru vytvořený tak, že každá statistická jednotka ze základního souboru má stejnou pravděpodobnost, že bude vybrána.

Pokud je možné vybrat tutéž jednotku znova, mluvíme o výběru *s vracením,* pokud opakovaný výběr není možný jedná se o výběr *bez vracení.*

Popisná statistika

Vlastnosti, které se pro jednotlivé jednotky mění nazýváme *veličinami,* případně *statistickými znaky* nebo *proměnnými.*

Vyskytují se veličiny

- *kvantitativní,* popsané číselnou hodnotou (výška, váha, cena);
- *kvalitativní,* popsané vlastnostmi (muž, žena, barva očí, dosažené vzdělání).

Kvantitativní veličiny mohou být *diskrétní,* nabývající hodnot ze zadané konečné množiny, nebo *spojité,* které nabývají hodnot ze zadaného intervalu.

Pozorovaním nebo měřením hodnot zkoumané veličiny na několika statistických jednotkách získáme *vstupní data.* Soubor těchto údajů nazýváme *datový soubor.* Tento soubor je *jednorozměrný,* jestliže sledujeme jeden znak, nebo *vícerozměrný* (*multistage random sample*), pokud sledujeme více znaků.

Při zpracování jednorozměrného datového souboru kvantitativních dat $x_1, x_2, \ldots, x_n$ potřebujeme pro některá šetření data uspořádat podle velikosti. Dostaneme pak *uspořádaný* datový soubor tvaru

$$x_{(1)} \leq x_{(2)} \ldots \leq x_{(n)},$$

kde $x_{(1)} = min\{x_i;\ 1 \leq i \leq n\}$ a $x_{(n)} = max\{x_i;\ 1 \leq i \leq n\}$.

Metody zpracovaní dat

11.3. Třídění dat je rozdělení dat do skupin provedené tak, aby vynikly charakteristické vlastnosti sledovaných jevů. Uspořádáme a zhustíme data do přehlednější formy. Rozeznáváme

- **jednostupňové třídění,** jestliže třídíme data podle změn jednoho statistického znaku;
- **vícestupňové třídění,** pokud provádíme třídění podle více znaků najednou.

Nejčastěji při jednostupňovém třídění kvantitativních dat uspořádáme data podle velikosti a stanovíme intervaly, které odpovídají jednotlivým třídám. Mluvíme pak o *intervalovém třídění.*

Máme-li datový soubor $\{x_1, x_2, \ldots, x_n\}$, který obsahuje celkem n prvků, pak interval mezi největší a nejmenší hodnotou rozdělíme na k disjunktních intervalů, tříd (*classes*), tvaru $(a_{i-1}, a_i\rangle,\ 1 \le i \le k$. Potom prvek x_j patří do $i-$té třídy, pokud je $a_{i-1} < x_j \le a_i$. Používáme následujících termínů a označení:

- **třída** (*class*) je část dat zařazená do jedné skupiny, intervalu $(a_{i-1}, a_i\rangle$;
- **dolní hranice třídy** (*lower class limit*) je hodnota a_{i-1};
- **horní hranice třídy** (*upper class limit*) je hodnota a_i;
- **střed třídy** (*class mark*) je průměr horní a dolní hranice třídy, tedy $y_i = \frac{1}{2}(a_{i-1} + a_i)$;
- **šířka třídy** (*class width*) je rozdíl horní a dolní hranice třídy, tedy hodnota $a_i - a_{i-1}$;
- **(absolutní) četnost třídy** (*frequency*) n_i je počet prvků souboru, které patří do $i-$té třídy;
- **relativní četnost** (*relative frequency*) $p_i = \frac{n_i}{n}$ je poměr četnosti třídy ku celkovému počtu dat;
- **kumulativní (absolutní) četnost** (*cumulative frequency*) $N_i = n_1 + n_2 + \ldots + n_i$ je součet četnosti třídy a četností tříd předchozích;
- **kumulativní relativní četnost** (*cumulative relative frequency*) $P_i = p_1 + p_2 + \ldots + p_i$ je součet relativní četnosti třídy a relativních četností tříd předchozích.

Potom platí:

$$\sum_{i=1}^{k} n_i = n, \quad \sum_{i=1}^{k} p_i = 1, \quad \sum_{j=1}^{i} n_j = N_i, \quad \sum_{j=1}^{i} p_j = P_i, \quad N_k = n, \quad P_k = 1.$$

Při stanovení hranic tříd obvykle zachováváme tato dvě pravidla:

- šířku třídy h volíme pro všechny intervaly shodnou, s vyjímkou krajních tříd pokud tvoří neomezené intervaly:
- při stanovení šířky třídy h dodržujeme Sturgesovo pravidlo, kdy pro počet tříd k platí, že $k \doteq 1 + 3,3 \log n$. V tabulce jsou uvedeny počty tříd pro některé hodnoty rozsahů souboru.

n	5	10	20	40	50	100	200	1000
k	3	4	5	6	7	8	9	11

- pokud jsou krajní intervaly dělení neomezené, pak za střed první, resp. poslední třídy volíme bod, který má od konečného krajního bodu třídy stejnou vzdálenost jakou má od středu sousední třídy.

Při třídění kvalitativních dat postupujeme obdobně. Jenom místo intervalu tvoří třídu prvky, které mají stejný znak, nebo skupinu znaků.

11.4. Grafická znázornění

Pro větší názornost požíváme místo tabulek znázornění datového souboru pomocí grafů. Používá se několika typů.

Histogram (*histogram*) je graf kdy na vodorovnou osu znázorníme třídy a na svislou osu četnosti či relativní četnosti. Často se používá ve tvaru, kdy se hodnota odpovídající třídě znázorní jako sloupec s intervalem třídy jako základnou a výška je dána četností.

Polygon četností a relativních četností je graf, kdy úsečkami spojíme body (y_i, n_i), resp. (y_i, p_i).

Bodový graf dostaneme tak, že na vodorovnou osu vyneseme třídy jako body i, $1 \leq i \leq k$, a ve svislém směru vynášíme jednotlivé prvky třídy znázorněné jako jednotlivé body $(i, j),\ j = 1, 2, \ldots n_i$.

Sloupkový graf je podobný histogramu, ale sloupce bývají oddělené, mají stejnou šířku a každý sloupec odpovídá jedné třídě. Používáme je především u kvalitativních dat.

Kruhový (výsečový) diagram (*pie chart*) je znázornění pomocí výsečí kruhu, kde každé třídě odpovídá jedna výseč. Velikosti obsahů výsečí odpovídají četnostem třídy.

Stem-and-Leaf diagram je uspořádání dat do tabulky, kdy první sloupec -stem=stonek odpovídá třídě a do řádku -leaf=list vypisujeme prvky třídy. Pokud tyto prvky uspořádáme podle velikosti mluvíme o uspořádaném diagramu.

Krabicový nebo vrubový krabicový graf (*box or whiskers plot*) znázorňuje význačné a extrémní hodnoty souboru.

Řada vlastností datového souboru se dá vyčíst z tvaru histogranu či polygonu četností. Ty odpovídají grafu hustoty u rozdělení pravděpodobnosti náhodné veličiny. Rozlišuje se několik charakteristických průběhů těchto grafů.

- **souměrný** ve tvaru zvonu, trojúhelníku či rovnoměrný;
- **nesouměrné** ve tvaru J, obráceného J, vpravo či vlevo protažené;
- **podle počtu vrcholů** jedno-, dvou-, či vícevrcholové (*unimodal, bimodal, multimodal*)

.

11.6. Charakteristiky (míry) polohy. Nejznámější a nejčastěji používanou charakteristkou polohy je aritmetický průměr hodnot souboru. **Průměr** (*mean*) datového souboru $\{x_1, x_2, \ldots, x_n\}$ je definován vztahem

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

Pokud jsou $\{z_1, z_k, \ldots, z_m\}$ různé hodnoty souboru s četnostmi n_j, $j = 1, 2, \ldots, m$, a s relativními četnostmi p_j, pak

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{m} z_j n_j = \sum_{j=1}^{m} z_j p_j.$$

Věta 1. Vlastnosti průměru Pro průměr datového souboru platí:

1. Součet odchylek hodnot souboru od průměru je roven nule, t.j. $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$.

2. Přičteme-li k hodnotám souboru konstantu a, pak průměr nového souboru $\{y_i = x_i + a\}$ je $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} (x_i + a) = \overline{x} + a$.

3. Násobíme-li hodnoty souboru číslem b, násobí se průměr také b, neboť pro soubor $\{y_i = bx_i\}$ je $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} bx_i = b\overline{x}$.

Pokud soubor $\{x_0, x_1, \ldots, x_n\}$ tvoří data, která odpovídají časové řadě sledující trend vývoje, pak jako charakteristiku polohy používáme **průměrný přírůstek**. Zavádíme jej jako průměr $\overline{y}$ souboru $\{y_i = x_i - x_0,\ 1 \le i \le n\}$. Je pak

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{x}(x_i - x_0) = \frac{1}{n}(x_n - x_0).$$

Medián Průměr datového souboru je citlivý na hrubé chyby, kdy jedna chybná hodnota může výrazně změnit hodnotu průměru.

Proto někdy používáme tzv. *robustních* charakteristik, které jsou méně citlivé na zadání chybné hodnoty. Mezi ně patří *medián* (*median*) $\tilde{x}$, který je pro datový soubor $x_1, x_2, \ldots x_n$ definován vztahem

$$\tilde{x} = \begin{cases} x_{(m)}, & \text{pro } n = 2m-1, \\ \frac{1}{2}\left(x_{(m)} + x_{(m+1)}\right), & \text{pro } n = 2m. \end{cases}$$

Další z robustních charakteristik je *modus* (*mode*) $\hat{x}$, který je definován jako hodnota souboru s největší četností, tedy

$$\hat{x} = z_j,\ n_j \ge n_i,\ 1 \le i \le m.$$

Používáme jej v případech, kdy nás zajímají „špičkové“ hodnoty souboru, např. při sledování dopravní zátěže v místě, počet cestujících v hromadné dopravě, spotřeba elektrické energie během dne a roku, či průtok řekou.

Kvantily, kvartily, decily, percentily

Definujeme pro p, $0 < p < 1$, $p - kvantil$, resp. $100p\%kvantil$, (*quantile*) jako tu hodnotu $\tilde{x}_{100p}$ ze souboru $\{x_1, x_2, \ldots, x_n\}$, pro kterou je přibližně $100p\%$ hodnot ze souboru menších a $100(1-p)\%$ hodnot je větších než $\tilde{x}_{100p}$.

Nejjemnější používané rozdělení souboru je pomocí *percentilů* (*percentile*) $\tilde{x}_1,\ \tilde{x}_2, \ldots, \tilde{x}_{99}$. Často se využívají *decily* $\tilde{x}_{10},\ \tilde{x}_{20}, \ldots, \tilde{x}_{90}$.

Speciální názvy mají kvantily:

- $\tilde{x}_{50}$ je *medián* (*median*);
- $\tilde{x}_{25}$ *dolní kvartil* (*lower quartile*);
- $\tilde{x}_{75}$ *horní kvartil* (*upper quartile*).

Jako *mezikvartilové rozpětí* IQR se definuje rozdíl $IQR = \tilde{x}_{75} - \tilde{x}_{25}$.

Jsou-li $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ hodnoty souboru uspořádané podle velikosti pak $p-$kvantil, resp. $100p\%$ kvantil určíme podle vzorce

$$\tilde{x}_{100p} = \begin{cases} x_{([np]+1)}, & \text{pokud } np \text{ není celé číslo}, \\ \frac{1}{2}(x_{(np)} + x_{(np)+1}) & \text{pro } np \text{ celé}, \end{cases}$$

kde $[np]$ je celá část čísla, tedy celé číslo, které je nejbližší menší. Při větších rozdílech mezi jednotlivými daty používáme pro přesnější vymezení kvantilů lineární aproximace mezi sousedními hodnotami.

Závěr

modus snadno se najde, má ale minimální vypovídací hodnotu:
medián určuje střed souboru a je méně citlivý na chyby;
průměr zohledňuje všechny hodnoty, ale je citlivý na chyby.

Useknuté průměry

Je-li $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ uspořádaný výběr, pak pro číslo $0 < \alpha < 0,5$ nazýváme hodnotu

$$\overline{x}_\alpha = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}$$

α-useknutým průměrem (*alpha-trimmed mean*).

Hodnotu

$$\overline{x}_{\alpha w} = \frac{1}{n} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} \Big([n\alpha] x_{([n\alpha])} + x_{(i)} + [n\alpha] x_{(n-[n\alpha]+1)} \Big)$$

nazýváme *α-winsorizovaný průměr* (*α-winsored mean*). Symbol $[n\alpha]$ označuje největší celé číslo k, pro které je $k \leq n\alpha$.

Jiné průměrové charakteristiky polohy. Pro soubory kladných dat používáme také jiné průměry. Jsou to:

Geometrický průměr (*geometric mean*) $\overline{x}_G$, který je pro soubor $x_1, x_2, \ldots, x_n$ kladných dat definován vztahem

$$\overline{x}_G = \sqrt[n]{x_1 x_2 \ldots x_n}.$$

Vlastnosti geometrického průměru. Násobíme-li hodnoty původního souboru číslem c, násobí se týmž číslem i geometrický průměr.

Pro logaritmus geometrického průměru platí:

$$\ln \overline{x}_G = \overline{\ln x} = \frac{1}{n} \sum_{i=1}^{n} \ln x_i.$$

Věta 2. Pro soubor s kladnými daty je

$$\overline{x}_G \leq \overline{x}$$

a rovnost nastane jedině pro $x_1 = x_2 = \ldots = x_n$.

Harmonický průměr (*harmonic mean*) $\overline{x}_H$, který je pro soubor kladných dat definován vztahem

$$\overline{x}_H = \frac{n}{x_1^{-1} + x_2^{-1} + \ldots + x_n^{-1}}.$$

Věta 3. Pro soubor s kladnými daty je

$$\overline{x}_H \leq \overline{x}_G \leq \overline{x},$$

přičmž rovnost nastane pouze pro $x_1 = x_2 = \ldots = x_n$.

Kvadratický průměr (*quadratic mean*)$\overline{x}_K$ je definován vztahem

$$\overline{x}_K = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}.$$

Věta 4. Je $\overline{x} \leq \overline{x}_K$ a rovnost platí pouze v případě, že $x_1 = x_2 = \ldots x_n$.

Věta 5. Pro soubory kladných dat je

$$x_{(1)} \leq \overline{x}_H \leq \overline{x}_G \leq \overline{x} \leq \overline{x}_K \leq x_{(n)}$$

a rovnost nastane pouze v případě, že $x_1 = x_2 = \ldots = x_n$.

11.7. Charakteristiky (míry) rozptýlenosti.

Rozpětí datového souboru (*range*) je hodnota $R = x_{max} - x_{min}$.

Hodnota se po uspořádání souboru snadno spočítá, ale její hodnota je citlivá na zavlečené chyby. Vychází pouze ze dvou hodnot a ignoruje informaci z ostatních hodnot souboru. V některých případech proto používáme jako charakteristiku tohoto druhu hodnotu $\tilde{x}_{90} - \tilde{x}_{10}$. Provedeme vlastně „ořezání“ souboru, když vynecháme hodnoty menší než $\tilde{x}_{10}$ a větší než $\tilde{x}_{90}$, tedy 10% nejmenších a 10% největších hodnot.Odstraníme tím vliv případných chybných hodnot, které leží na hranicích souboru.

Podobnou charakteristikou je **mezikvartilové rozpětí** (*interquartile range*) $IQR = \tilde{x}_{75} - \tilde{x}_{25}$.

Střední kvadratická odchylka (MSD) (*mean of squared deviation*) je průměr čtverců odchylek od průměru a je definován vztahem

$$MSD = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Rozptyl (*dispersion, variance*) je definován vzrcem

$$S^2 = \frac{n}{n-1}MSD = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

a **směrodatná odchylka** (*standard deviation*) S je odmocninou z rozptylu.

Věta 6. Vlastnosti rozptylu a MSD a vzorce pro výpočet.
1. Je

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n(\overline{x})^2\right), \qquad s^2 = MSD = \overline{x^2} - (\overline{x})^2.$$

2. Je-li $y_i = bx_i + a,\ 1 \leq i \leq n$, pak $s_y^2 = b^2 s_x^2,\ s_y = |b|s_x$ a $S_y^2 = b^2 S_x^2,\ S_y = |b|S_x$

Věta 7. Funkce $S(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^2$ nabývá svého minima s^2 pro $\alpha = \overline{x}$.

Pro soubory, které obsahují velké množství dat je výhodnější charakteristiky polohy a rozpětí odhadovat. Uvedeme některé jednoduché odhady a o dalších pojednáme později.

Pomocné tvrzení (Cauchyova nerovnost):
Pro $n-$tice čísel $(a_1, a_2, \ldots, a_k)$ a $(b_1, b_2, \ldots, b_k)$ je

$$\left(\sum_{i=1}^{k}a_i b_i\right)^2 \leq \left(\sum_{i=1}^{k}a_i^2\right)\left(\sum_{i=1}^{k}b_i^2\right).$$

Věta 8. Pro soubor $x_i,\ 1 \leq i \leq n$ platí

$$max\{|x_i - \overline{x}|;\ 1 \leq i \leq n\} \leq s\sqrt{n-1}.$$

Věta 9. Pro rozpětí souboru platí

$$s^2 \leq \frac{R^2}{4}, \quad S^2 \leq \frac{nR^2}{4(n-1)} \text{ tedy } s \leq \frac{R}{2}\sqrt{\frac{n}{n-1}}.$$

Průměrná odchylka (*mean of absolute deviation*) d_a od bodu a je pro soubor dat x_i definována vztahem

$$d_a = \frac{1}{n}\sum_{i=1}^{n}|x_i - a|.$$

Nejčastěji se používá průměrná odchylka od aritmetického průměru $\overline{x}$ nebo mediánu $\tilde{x}$. K tomu nás vede následující vlastnost.

Věta 10. Funkce d_a nabývá svého minima pro medián $a = \tilde{x}$.

Pokud používáme jako charakteristiku polohy medián $\tilde{x} = x_{0,5}$, pak místo směrodatné odchylky s používáme jako charakteristiku rozptylu mezikvartilové rozpětí $IQR = \tilde{x}_{0,75} - \tilde{x}_{0,25}$. V tomto intervalu leží 50% hodnot souboru. Omezujeme tím vliv případných extrémních hodnot, které mohou být zatížené chybou.

Pětičíselná charakteristika (*five-number summary*)souboru je pětice čísel

$$x_{min}, \tilde{x}_{25}, \tilde{x}_{50}, \tilde{x}_{75}, x_{max},$$

na které jsou založeny krabicové grafy.

Relativní variabilita

Můžeme také používat charakteristiky relativní variability, které jsou definovány jako poměr směrodatné odchylky a některého průměru. Nejčastěji se používá **variační koeficient,** který je definován vztahem

$$V = \frac{s}{\overline{x}}.$$

Určuje nám jakou částí se podílí směrodatná odchylka na aritmetickém průměru dat. Je-li $V > 0,5$ pak se jedná o nesourodý soubor. Variační koeficient má tyto vlastnosti, které pro jednoduchost budeme uvažovat pro kladná data.

Věta 11. Označme $\boldsymbol{x}$ soubor dat $\{x_i\},\ 1 \leq i \leq n$, $b\boldsymbol{x} = \{bx_i\},\ b > 0$ a $\boldsymbol{x} \pm a = \{x_i \pm a\},\ a > 0$. Potom pro variační koeficient V platí:

a) $V(b\boldsymbol{x}) = V(\boldsymbol{x})$;

b) $V(\boldsymbol{x} + a) < V(\boldsymbol{x})$;

c) $V(\boldsymbol{x} + a) < V(\boldsymbol{x}) < V(\boldsymbol{x} - a),\ 0 < a < \overline{x}$.

Jako aproximace se používá **relativní kvartilová odchylka** Q_r je definována vztahem

$$Q_r = \frac{\tilde{x}_{0,75} - \tilde{x}_{0,25}}{\tilde{x}_{0,75} + \tilde{x}_{0,25}}$$

Jiné charakteristiky

Koeficient šikmosti

$$A_3 = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

a **koeficient špičatosti**

$$A_4 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \overline{x})^4 - 3$$

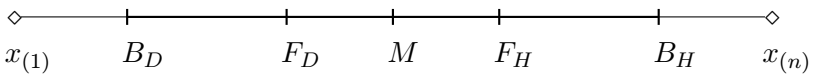
Pro data, která jsou rozložena symetricky kolem hodnoty $\overline{x}$ je $A_3 = 0$. Hodnoty A_3 blízké nule odpovídají rozdělení, které se blíží symetrickému. Je-li $A_3 > 0$, pak je rozložení dat sešikmené vpravo, menší hodnoty než průměr $\overline{x}$ jsou k němu více nahuštěny než hodnoty větší. Pro $A_3 < 0$ je rozdělení sešikmené vlevo, větší hodnoty jsou více nahuštěny k průměru než hodnoty nižší.

Je-li A_4 blízké nule, říkáme, že jedná o soubor s normální špičatostí. Při $A_4 < 0$ mluvíme o souborech plochých a při $A_4 > 0$ mluvíme o souborech špičatých. Podrobněji pojednáme o těchto charakteristikách později v souvislosti s náhodnou veličinou a jejím rozdělením.

11.9. Písmenkové charakteristiky V některých aplikacích se používají označení charakteristik polohy a variability pomocí písmen. Označujeme tak kvantily, které mají po řadě hodnoty $p = \frac{1}{2^n}$ a některé veličiny, které charakterizují rozptýlení hodnot souboru.
M − medián $\tilde{x} = x_{0,5}$, tedy $0,5-$kvantil;
F − kvartily; F_D dolní kvartil $x_{0,25}$; F_H horní kvartil $x_{0,75}$;
E − oktily; E_D dolní oktil, kvantil $x_{1/8}$; E_H horní oktil, kvantil $x_{7/8}$;
D − sedecily; D_D dolní sedecil, kvantil $x_{1/16}$; D_H horní sedecil, kvantil $x_{15/16}$.
$R_F = F_H - F_D = IQR$ je mezikvartilové rozpětí.
$B_D,\ B_H$ vnitřní hradby souboru, kde $B_D = F_D - 1,5R_F,\ B_H = F_H + 1,5R_F$.
(I_D, I_H) interval spolehlivosti pro medián, kde $I_D = M - \frac{1,57R_F}{\sqrt{n}}$ a
$I_H = M + \frac{1,57R_F}{\sqrt{n}}$, přičemž n je počet prvků v souboru.

11.10. Grafická znázornění

I. Graf dat

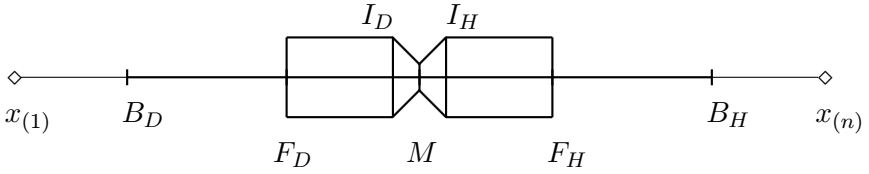


Obr. 11.12

II. Krabicový graf Šířku obdélníka volíme úměrnou hodnotě $\sqrt{n}$

Obr. 11.13

III. Vrubový krabicový graf Šířku obdélníka volíme úměrnou hodnotě $\sqrt{n}$



Obr. 11.14

Krabicové grafy jsou vhodné pro porovnání dvojice souborů, kdy případné rozdíly jsou okamžitě patrné z rozměrů „krabic“.

IV. Histogram

V. Graf polosum k testování symetrie.

Na osu x vynášíme hodnoty $x(i)$ a na osu y hodnoty „polosum“ $y_i = \frac{1}{2}(x_{(i)} + x_{(n+1-i)})$. Pro symetrické rozdělení leží body kolem přímky $y = M$.

VI. Kvantil=kvantilový $Q - Q$ graf je grafem kvantilové funkce.

Na osu x vynášíme hodnoty P_i- kvantilů $Q(P_i),\ P_i = \frac{i}{n+1}$ a na osu y hodnoty $y = x_{(i)}$.

VII. Pravděpodobnostní $P - P$ graf je grafem distribuční funkce.

Na osu x vynášíme hodnoty $x_{(i)}$ a na osu y hodnoty $P_i = \frac{i}{n+1}$.

Oba grafy slouží k testování shody rozdělení, kde porovnáváme průběhy pro dva soubory. Používáme je ve dvojici, kdy využíváme toho, že $Q-Q$ graf je citlivější na chyby v okrajových datech souboru a $P - P$ graf je naopak citlivý na chyby v okolí mediánu.

VIII. Rankitový graf

je kvantilový $Q - Q$ graf, ve kterém porovnáváme rozdělení s normálním rozdělením. Na osu x vynášíme P_i kvantil x_{P_i} normálního rozdělení a na osu y hodnoty $y = x(i)$. Parametry příslušnéo normálního rozdělení odhadneme pomocí hodnot

$$\hat{\mu} = M, \quad \hat{\sigma} = \frac{3}{4}(F_H - F_D).$$

Odpovídající kvantily určíme pomocí vzorců

$$U_i = \Phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right), \quad x_{P_i} = \Phi^{-1}\left(\frac{1}{2}(U_{i-1} + U_{i+1})\right),\ U_0 = 0,\ U_{n+1} = 1.$$

V případě normálního rozdělení leží body na přímce.

11.11. Vícerozměrné soubory

Sledujeme-li dva znaky, pak soubor dat má charakter uspořádaných dvojic $\{(x_i, y_i),\ 1 \leq i \leq n\}$. První otázkou, kterou obvykle řešíme je popis závislosti prvního a druhého znaku. Jako charakteristiku polohy volíme dvojici $(\overline{x}, \overline{y})$. Za charakteristiku variability obvykle volíme směrodatné odchylky $s_x,\ s_y$. Jako míru statistické závislosti volíme *koeficient korelace.*

11.12. Koeficient korelace (*covariance, coefficient of variation*) r_{xy} dvou souborů $\{x_i\}$ a $\{y_i\},\ 1 \leq i \leq n$ je definován vztahem

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x.s_y}$$

Vlastnosti koeficientu korelace

a) $r_{xy} = \left((\frac{1}{n}\sum_{i=1}^{n} x_i y_i) - \overline{xy}\right)/(s_x.s_y)$;

b) $r_{xy} = r_{yx}; \quad r_{xx} = 1$;

c) $|r_{xy}| \leq 1$;

d) pro $y_i = ax_i + b$ je $r_{xy} = \mathrm{sgn} a$.

e) $r_{xy} = \pm 1 \Rightarrow y = ax + b$.