1. Průzkumová analýza jednorozměrných dat, diagnostické grafy

1.1. Motivace

Průzkumová analýza dat je odvětví statistiky, které pomocí různých postupů odhaluje zvláštnosti v datech. Při zpracování dat se často používají metody, které jsou založeny na předpokladu, že data pocházejí z nějakého konkrétního rozložení, nejčastěji normálního. Tento předpoklad nemusí být vždy splněn, protože data

- mohou pocházet z jiného rozložení
- mohou být zatížena hrubými chybami
- mohou pocházet ze směsi několika rozložení.

Proto je důležité provést průzkumovou analýzu dat, abychom se vyvarovali neadekvátního použití statistických metod.

1.2. Funkcionální charakteristiky datového souboru

1.2.1. Označení

Na množině objektů $\{\varepsilon_1,\ldots,\varepsilon_n\}$ zjišťujeme hodnoty znaku X. Hodnotu znaku X na objektu ε_i označíme x_i, i = 1, ..., n. Tyto hodnoty zaznamenáme do jednorozměrného datového souboru $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. Uspořádané hodnoty $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ tvoří uspořádaný datový soubor $\begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{pmatrix}$. Vektor $\begin{pmatrix} x_{[1]} \\ \vdots \\ x_{[r]} \end{pmatrix}$, kde $x_{[1]} < \ldots < x_{[r]}$ jsou navzájem různé hodnoty znaku X, se nazývá vektor variant.

1.2.2. Bodové rozložení četností

Je-li počet variant malý, přiřazujeme četnosti jednotlivým variantám a hovoříme o bodovém rozložení četností.

n_j – absolutní četnost varianty $x_{[j]}$

$p_j = \frac{n_j}{n}$ – relativní četnost varianty $x_{[j]}$

$N_j = n_1 + \ldots + n_j$ – absolutní kumulativní četnost prvních j variant

$F_j = \frac{N_j}{n} = p_1 + \ldots + p_j$ – relativní kumulativní četnost prvních j variant

Absolutní či relativní četnosti znázorňujeme graficky např. pomocí sloupkového diagramu či polygonu četností.

Četnostní funkce: $p(x) = \begin{cases} p_j \text{ pro } x = x_{[j]},\ j = 1,\ \ldots, r \\ 0 \text{ jinak} \end{cases}$

Empirická distribuční funkce: $F(x) = \begin{cases} 0 \text{ pro } x < x_{[1]} \\ F_j \text{ pro } x_{[j]} \le x < x_{[j+1]},\ j = 1, \ldots, r-1 \\ 1 \text{ pro } x \ge x_{[r]} \end{cases}$

Příklad 1.: U 30 domácností byl zjišťován počet členů.

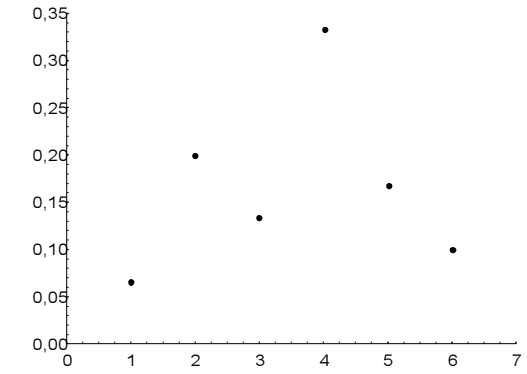
Počet členů	1	2	3	4	5	6
Počet domácností	2	6	4	10	5	3

Vytvořte tabulku rozložení četností. Nakreslete grafy četností funkce a empirické distribuční funkce. Dále nakreslete sloupkový diagram a polygon četností počtu členů domácnosti.

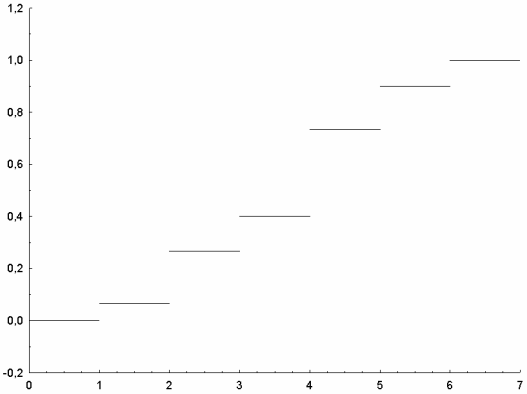
Řešení:
Tabulka rozložení četností

$x_{[j]}$	n_j	p_j	N_j	F_j
1	2	2/30	2	2/30
2	6	6/30	8	8/30
3	4	4/30	12	12/30
4	10	10/30	22	22/30
5	5	5/30	27	27/30
6	3	3/30	30	1

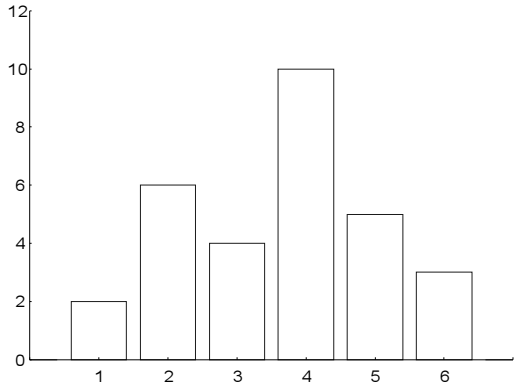
Graf četnostní funkce



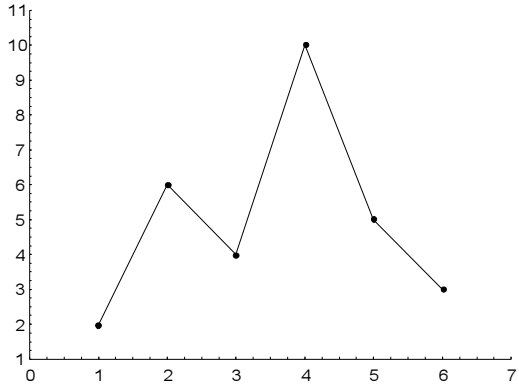
Graf empirické distribuční funkce



Sloupkový diagram



Polygon četností



1.2.3. Intervalové rozložení četností

Je-li počet variant velký, přiřazujeme četnosti nikoli jednotlivým variantám, ale třídicím intervalům $\langle u_1, u_2 \rangle$, ..., $\langle u_r, u_{r+1} \rangle$ a hovoříme o intervalovém rozložení četností. Názvy četností jsou podobné jako v bodě 1.2.2., navíc zavádíme četnostní hustotu j-tého třídicího intervalu $f_j = \frac{p_j}{d_j}$, kde $d_j = u_{j+1} - u_j$. Stanovení počtu třídicích intervalů je dosti subjektivní záležitost. Často se doporučuje volit r blízké $\sqrt{n}$.

Hustota četnosti: $f(x) = \begin{cases} f_j \text{ pro } u_j < x \le u_{j+1},\ j = 1, \cdots, r \\ 0 \text{ jinak} \end{cases}$ (grafem hustoty četnosti je histogram)

Intervalová empirická distribuční funkce: $F(x) = \int_{-\infty}^{x} f(t)dt$.

Příklad 2.: U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

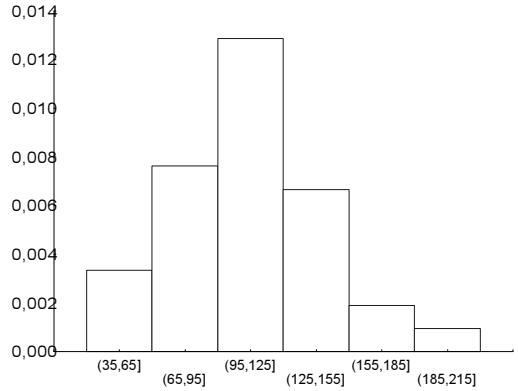
Výdaje	$\langle 35, 65 \rangle$	$\langle 65, 95 \rangle$	$\langle 95, 125 \rangle$	$\langle 125, 155 \rangle$	$\langle 125, 155 \rangle$	$\langle 185, 215 \rangle$
Počet dom.	7	16	27	14	4	2

Sestavte tabulku rozložení četností, nakreslete histogram a graf intervalové empirické distribuční funkce.

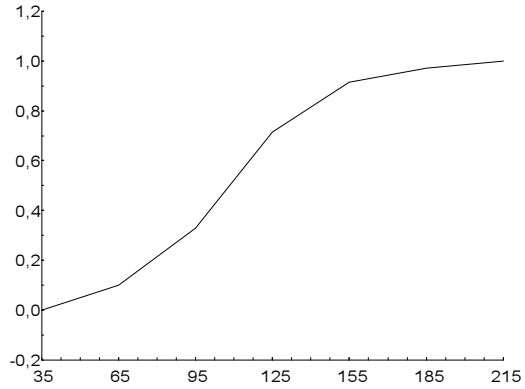
Řešení:
Tabulka rozložení četností

$(u_j,u_{j+1}]$	n_j	p_j	f_j	N_j	F_j
$\langle 35,65)$	7	7/70	7/2100	7	7/70
$\langle 65,95)$	16	16/70	16/2100	23	23/70
$\langle 95,125)$	27	27/70	27/2100	50	50/70
$\langle 125,155)$	14	14/70	14/2100	64	64/70
$\langle 155,185)$	4	4/70	4/2100	68	68/70
$\langle 185,215)$	2	2/70	2/2100	70	1

Histogram



Graf intervalové empirické distribuční funkce



1.3. Číselné charakteristiky datového souboru

1.3.1. Znaky nominálního typu

Tyto znaky umožňují obsahovou interpretaci pouze u relace rovnosti. Charakteristikou polohy je modus, tj. nejčetnější varianta či střed nejčetnějšího intervalu.

1.3.2. Znaky ordinálního typu

Lze u nich navíc obsahově interpretovat relaci uspořádání. Charakteristikou polohy je α-kvantil. Je-li $\alpha \in (0;1)$, pak α-kvantil x_α je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl 1 – α všech dat. Pro výpočet α-kvantilu slouží algoritmus:

$$n\alpha = \begin{cases} \text{celé číslo } c \Rightarrow x_\alpha = \dfrac{x_{(c)} + x_{(c+1)}}{2} \\ \text{necelé číslo} \Rightarrow \text{zaokrouhlíme nahoru na nejbližší celé číslo } c \Rightarrow x_\alpha = x_{(c)} \end{cases}$$

Pro speciálně zvolená α užíváme názvů: $x_{0,50}$ – medián, $x_{0,25}$ – dolní kvartil, $x_{0,75}$ – horní kvartil, $x_{0,1}$, ..., $x_{0,9}$ – decily, $x_{0,01}$, ..., $x_{0,99}$ – percentily. Jako charakteristika variability slouží kvartilová odchylka: $q = x_{0,75} - x_{0,25}$.

Příklad 3.: Během semestru se studenti podrobili písemnému testu z matematiky, v němž bylo možno získat 0 až 10 bodů. Výsledky jsou uvedeny v tabulce:

Počet bodů	0	1	2	3	4	5	6	7	8	9	10
Počet studentů	1	4	6	7	11	15	19	17	12	6	3

Zjistěte modus, medián, 1.decil, 9. decil a kvartilovou odchylku počtu bodů.

Řešení: Modus je nejčetnější varianta znaku, v tomto případě tedy 6.
Pro výpočet kvantilů musíme znát rozsah datového souboru: n = 1 + 4 + ... + 3 = 101. Výpočty uspořádáme do tabulky.

α	nα	c	$x_\alpha = x_{(c)}$
0,50	50,5	51	6
0,10	10,1	11	2
0,90	90,9	91	8
0,25	25,25	26	4
0,75	75,75	76	7

$q = 7 - 4 = 3$

1.3.3. Znaky intervalového a poměrového typu

U těchto znaků lze navíc obsahově interpretovat operaci rozdílu resp. podílu.

Charakteristika polohy: aritmetický průměr $m = \frac{1}{n}\sum_{i=1}^{n} x_i$, u poměrových znaků, které nabývají pouze kladných hodnot, lze použít geometrický průměr $\sqrt[n]{x_1 \cdot \ldots \cdot x_n}$.

Charakteristika variability: rozptyl $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$ či směrodatná odchylka $s = \sqrt{s^2}$.

(Rozptyl se zpravidla počítá podle vzorce $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - m^2$.)

U poměrových znaků se jako charakteristika variability používá též koeficient variace $\frac{s}{m}$.

Známe-li absolutní či relativní četnosti variant $x_{[1]}$, ..., $x_{[r]}$, můžeme spočítat vážený průměr $m = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}$ či vážený rozptyl : $s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j (x_{[j]} - m)^2$. (Vážený rozptyl se zpravidla počítá podle vzorce $s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}^{\,2} - m^2$.)

Aritmetický průměr a rozptyl jsou speciální případy momentů. Zavedeme

k-tý počáteční moment $m_k' = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,k}$, k = 1, 2, ... a k-tý centrální moment

$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^k$, k = 1, 2, ... Pomocí 3. a 4. počátečního momentu se definuje šikmost a špičatost.

Šikmost: $\alpha_3 = \frac{m_3}{s^3}$ - měří nesouměrnost rozložení četností kolem průměru.

Špičatost: $\alpha_4 = \frac{m_4}{s^4} - 3$ - měří koncentraci rozložení četností kolem průměru.

Příklad 4.: Pro údaje z příkladu 1 vypočtěte průměr a rozptyl počtu členů.

Řešení: $m = \frac{1}{30}(1\cdot 2 + 2\cdot 6 + 3\cdot 4 + 4\cdot 10 + 5\cdot 5 + 6\cdot 3) = \frac{109}{30} = 3{,}6\bar{3}$

$$s^2 = \frac{1}{30}(1^2\cdot 2 + 2^2\cdot 6 + 3^2\cdot 4 + 4^2\cdot 10 + 5^2\cdot 5 + 6^2\cdot 3) - \left(\frac{109}{30}\right)^2 = \frac{1769}{900} = 1{,}96\bar{5}$$

Příklad 5.: Nechť m_1 je průměr a s_1^2 rozptyl hodnot $x_1, ..., x_n$. Nechť a, b jsou reálné konstanty. Položme $y_i = a + bx_i$, i = 1, ..., n. Vypočtěte průměr m_2 a rozptyl s_2^2 hodnot $y_1, ..., y_n$.

Řešení: $m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = a + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + bm_1$

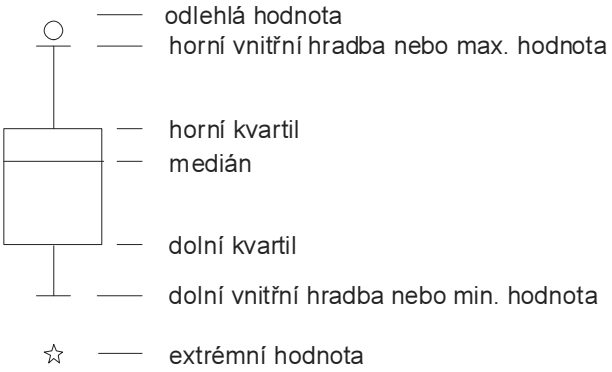
$$s_2^{\,2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - m_2)^2 = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - bm_1)^2 = b^2\frac{1}{n}\sum_{i=1}^{n}(x_i - m_1)^2 = b^2 s_1^{\,2}$$

1.4. Diagnostické grafy

1.4.1. Krabicový diagram

Umožňuje posoudit symetrii a variabilitu datového souboru a existenci odlehlých či extrémních hodnot.

Způsob konstrukce



Odlehlá hodnota leží mezi vnějšími a vnitřními hradbami, tj. v intervalu $(x_{0,75} + 1{,}5q, x_{0,75} + 3q)$ či v intervalu $(x_{0,25} - 3q, x_{0,25} - 1{,}5q)$.

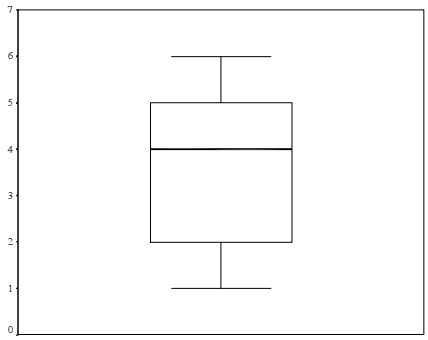
Extrémní hodnota leží za vnějšími hradbami, tj. v intervalu $(x_{0,75} + 3q, \infty)$ či v intervalu $(-\infty, x_{0,25} - 3q)$.

Příklad 6.: Pro údaje z příkladu 1 sestrojte krabicový diagram.

Řešení: Rozsah souboru n = 30. Výpočty potřebných kvantilů uspořádáme do tabulky.

α	$n\alpha$	c		x_α
0,25	7,5	8	$x_{(c)}=x_{(8)}$	2
0,50	15	15	$\frac{x_{(15)} + x_{(16)})}{2}$	4
0,75	22,5	23	$x_{(c)}=x_{(23)}$	5

$q = 5 - 2 = 3$
Dolní vnitřní hradba: $x_{0,25} - 1,5q = 2 - 1,5.3 = -2,5$
Horní vnitřní hradba: $x_{0,75} + 1,5q = 5 + 1,5.3 = 9,5$



1.4.2. Normal probability plot (NP-plot)

Umožňuje graficky posoudit, zda data pocházejí z normálního rozložení.
Způsob konstrukce: na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \le ... \le x_{(n)}$ a na svislou osu kvantily u_{α_j}, kde $\alpha_j = \frac{3j-1}{3n+1}$ (jsou-li některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající takové skupince). Pocházejí-li data z normálního rozložení, pak všechny dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou ležet na přímce.

1.4.3. Quantile - quantile plot (Q-Q plot)

Umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení (např. STATISTICA 6.0 nabízí 8 typů rozložení: beta, exponenciální, Gumbelovo, gamma, log-normální, normální, Rayleighovo a Weibulovo).
Způsob konstrukce: na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \le ... \le x_{(n)}$ a na vodorovnou osu kvantily $K_{\alpha_j}(X)$ vybraného rozložení, kde $\alpha_j = \frac{j - r_{adj}}{n + n_{adj}}$, přičemž r_{adj} a n_{adj} jsou korigující faktory $\le 0,5$, implicitně $r_{adj} = 0,375$ a $n_{adj} = 0,25$. (Jsou-li některé hodnoty $x_{(1)} \le ... \le x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.) Pokud vybrané rozložení závisí na nějakých parametrech, pak se tyto parametry odhadnou z dat nebo je může zadat uživatel. Body $\left(K_{\alpha_j}(X), x_{(j)}\right)$ se metodou nejmenších čtverců proloží přímka. Čím méně

se body odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.

Příklad 7.: Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Pomocí N-P plot a P-P plot ověřte, zda se tato data řídí normálním rozložením.

Řešení:

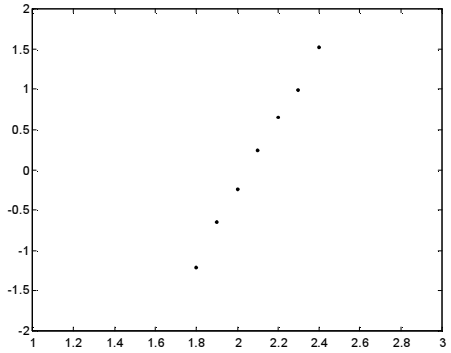
usp.hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

a) N-P plot

$j = (1,5\ \ 3\ \ 4,5\ \ 6,5\ \ 8\ \ 9\ \ 10)$

$$\alpha_j = \frac{3j-1}{3n+1} = (0,1129;0,2581;0,4032;0,5968;0,7419;0,8387;0,9355)$$

$$u_{\alpha_j} = (-1,2112;-0,6493;-0,245;0,245;0,6493;0,9892;1,5179)$$

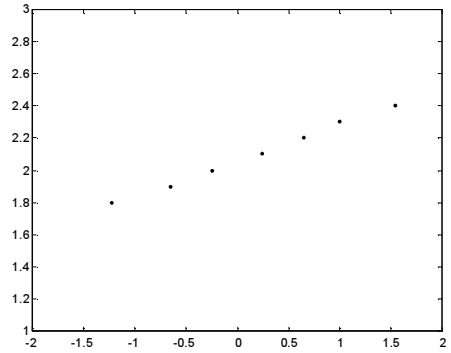


b) Q-Q plot

$j = (1,5\ \ 3\ \ 4,5\ \ 6,5\ \ 8\ \ 9\ \ 10)$

$$\alpha_j = \frac{j-0,375}{n+0,25} = (0,1098;0,2561;0,4024;0,5976;0,7439;0,8415;0,939)$$

$$u_{\alpha_j} = (-1,2278;-0,6554;-0,247;0,247;0,6554;1,0005;1,566)$$



Vzhled obou grafů nasvědčuje tomu, že data pocházejí z normálního rozložení.

1.4.4. Probability - probability plot (P-P plot)

Používá se ke stejným účelům jako Q-Q plot, ale jinak se konstruuje.

Způsob konstrukce: spočtou se standardizované hodnoty $z_{(j)} = \frac{x_{(j)} - m}{s}$, j = 1, ..., n. Na vodorovnou osu se vynesou hodnoty teoretické distribuční funkce $\Phi(z_{(j)})$ a na svislou osu hodnoty empirické distribuční funkce $F(z_{(j)}) = j/n$. (Jsou-li některé hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.)Pokud se body $(\Phi(z_{(j)}), F(z_{(j)}))$ řadí kolem hlavní diagonály čtverce $[0,1] \times [0,1]$, lze usuzovat na dobrou shodu empirického a teoretického rozložení.

1.4.5. Histogram

Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti vybraného teoretického rozložení. (Ve STATISTICE je pojem histogramu širší, skrývá se za ním i sloupkový diagram.)

Způsob konstrukce ve STATISTICE: na vodorovnou osu se vynášejí třídicí intervaly (implicitně 10, jejich počet lze změnit, stejně tak i meze třídicích intervalů) či varianty znaku a na svislou osu absolutní nebo relativní četnosti třídicích intervalů či variant. Do histogramu se zakreslí tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení. Kromě 8 typů rozložení uvedených u Q-Q plotu umožňuje STATISTICA použít ještě další 4 rozložení: Laplaceovo, logistické, geometrické, Poissonovo.

Příklady k 1. kapitole

Příklad 1. : U 20 studentů 1. ročníku byla zjišťována známka z matematiky na prvním zkušebním termínu.

Známka	1	2	3	4
Počet studentů	7	3	2	8

Vytvořte tabulku rozložení četností. Nakreslete grafy četností funkce a empirické distribuční funkce. Dále nakreslete sloupkový diagram a polygon četností počtu členů domácnosti.

Příklad 2. : U 60 vzorků oceli byl zjišťována mez plasticity.

Mez plasticity	$(30,50\rangle$	$(50,70\rangle$	$(70,90\rangle$	$(90,110\rangle$	$(110,130\rangle$	$(130,150\rangle$	$(150,170$
Počet vzorků	8	4	13	15	9	7	4

Sestavte tabulku rozložení četností, nakreslete histogram a graf intervalové empirické distribuční funkce.

Příklad 3. : Pro údaje z příkladu 2 vypočtěte průměr a rozptyl meze plasticity.
($m = 96,67$, $s^2 = 1148,89$)

Příklad 4. : V datovém souboru, z něhož byl vypočten průměr 110 a rozptyl 800, byly zjištěny 2 chyby: místo 85 má být 95 a místo 120 má být 150. Ostatních 18 údajů je správných. Opravte průměr a rozptyl.
($m = 112$, $s^2 = 851$)

Příklad 5. :Pro údaje z příkladu 1 sestrojte krabicový diagram.
(Pomocné výpočty: $x_{0,50} = 2,5$, $x_{0,25} = 1$, $x_{0,75} = 4$, $q = 3$, dolní vnitřní hradba = -3,5, horní vnitřní hradba = 8,5)

Práce se systémem STATISTICA

Téma: Grafické a tabulkové zpracování četností, výpočet číselných charakteristik, diagnostické grafy

Vedení pojišťovny (zaměřené na pojištění automobilů) požádalo manažera oddělení marketingového výzkumu o provedení průzkumu, který by ukázal názory zákazníků na uvažovaný nový systém pojištění aut.

Náhodně bylo vybráno 110 současných zákazníků pojišťovny a ti byli telefonicky seznámeni s následujícím textem:

„Naše pojišťovna nabízí nový systém pojištění aut výhradně pro cesty nad 300 km. Za roční poplatek 12 tisíc Kč budete pojištěni pro případ libovolných potíží s autem při všech cestách nad 300 km. V případě nehody pojišťovna uhradí opravu, cestovní náklady a popř. i některé další výlohy, jako je ubytování a stravování v hotelu, telefon atd.

Stupnicí od 1 (jednoznačný nezájem) do 5 (jednoznačný zájem) laskavě vyjádřete svůj postoj k nabízenému novému typu pojištění. Dále uveďte svůj věk, počet cest nad 300 km v loňském roce, stáří vašeho auta a váš rodinný stav. Děkujeme."

Získané odpovědi byly zaznamenány do datového souboru a zakódovány takto:

POSTOJ ... postoj k novému typu pojištění (ne = 1, asi ne = 2, nevím = 3, asi ano = 4, ano = 5).
RODSTAV ... rodinný stav (svobodný = 1, rozvedený, ovdovělý = 2, ženatý = 3).
VEK ... věk v dokončených letech.
STARIAUT ... stáří auta v letech.
CESTY ... počet cest nad 300 km v předešlém roce.

Úkoly:

1. Datový soubor pojist.sta načtěte do systému STATISTICA. Všem proměnným vytvořte návěští a popište význam jednotlivých variant proměnných POSTOJ a RODSTAV.
   Návod: File – Open – Soubory typu Data Files – pojist.sta – Otevřít.
   Názvy a vlastnosti proměnných se upravují v okně, do něhož vstoupíte, když 2x kliknete myší na název proměnné. Návěští se píše do Long Name, význam variant do Text labels.
2. Zjistěte absolutní a relativní četnosti a absolutní a relativní kumulativní četnosti proměnných POSTOJ a RODSTAV.
   Návod: Statistics – Basic Statistics/Tables – Frequency Tables – OK – Variables POSTOJ, RODSTAV – OK – Summary Frequency tables. Tabulky se uloží do workbooku, listovat v nich můžete pomocí stromové struktury v levém okně.
3. Proměnnou VEK zakódujte do 6 třídicích intervalů <23,29>, (29,35>, (35,41>, (41,47>, (47,53>, (53,59> a zjistěte jejich četnosti.
   Návod: Za VEK vložte novou proměnnou RVEK (Insert – Add Variables –After VEK, Name RVEK, Type Integer, Long Name zakódovaný věk, OK). Nastavte se kurzorem na RVEK. Data – Recode – Category 1 Include If VEK >=23 and VEK <=29, New Value1, value1 atd. až Category 7 Include If VEK > 53 and VEK <=59, New Value 7 value 7, OK. Četnosti zjistíte analogicky jako v bodě 2.
4. Vypočtěte následující číselné charakteristiky: POSTOJ (ordinální proměnná) – modus, medián, dolní a horní kvartil, kvartilová odchylka. RODSTAV (nominální proměnná) – modus. VEK, STARIAUT, CESTY (poměrové proměnné) – modus, medián, průměr, minimum, maximum, směrodatná odchylka, rozptyl, šikmost, špičatost.
   Návod: Statistics – Basic Statistics/Tables - Descriptive Statistics – OK, Variables – název proměnné, Advanced – vyberte příslušné charakteristiky (modus – Mode, medián – Median, průměr – Mean, směrodatná odchylka – Standard Deviation, rozptyl – Variance,

šikmost – Skewness, špičatost – Kurtosis, dolní a horní kvartil – Lower&upper quartiles, kvartilová odchylka – Quartile range).

5. Vytvořte sloupkový diagram, výsečový graf a polygon četností proměnných POSTOJ a RODSTAV.
   Návod: Sloupkový diagram: Graphs – Histograms – Variables POSTOJ, OK, Advanced – Fit type Off, zaškrtneme Breaks between Columns, Y Axis %&N, OK.
   Výsečový graf: Graphs – 2D Graphs – Pie Charts – Variables POSTOJ, OK, Advanced - Pie legend Text and Percent, OK.
   Polygon četností: ve workbooku vstupte do tabulky rozložení četností proměnné POSTOJ. Pomocí Edit – Delete - Cases vymažte řádek označený Missing. Nastavte se kurzorem na Count a kliknutím pravého tlačítka vstupte do menu Line Plot: Entire Columns. Vytvoří se polygon četností.
6. Vytvořte histogram proměnné VEK se šesti třídicími intervaly <23,29>, (29,35>, (35,41>, (41,47>, (47,53>, (53,59>.
   Návod: Graphs – Histograms – Variables VEK, OK, Advanced – zaškrtněte Boundaries – Specify Boundaries – Enter Upper Boundaries 29 35 41 47 53 59, OK.
7. Vytvořte kategorizovaný histogram proměnné VEK podle proměnné RODSTAV.
   Návod: Postupujte stejně jako v předešlém případě a zvolte Categorized – X-categoried ON – Change Variable RODSTAV, OK, Codes – Specify Codes All, OK, OK.
8. Zjistěte, jaký je průměrný počet cest nad 300 km pro svobodné, rozvedené , ženaté zákazníky pojišťovny.
   Návod: Postupujte stejně jako v úkolu č. 4, ale klikněte na SELECT CASES – zaškrtněte Enable Selection Conditions – Include cases – zaškrtněte Specific, selected by, By Expression RODSTAV = 1, OK. Pro rozvedené či ženaté zákazníky použijete RODSTAV = 2 či RODSTAV = 3,
9. Sestrojte krabicový diagram proměnné CESTY. S jeho pomocí zjistěte, zda proměnná CESTY obsahuje odlehlé či extrémní hodnoty.
   Návod: Graphs – 2D Graphs – Box Plots – Variables – Dependent variable CESTY – OK – OK.
   Interpretace: Medián je posunut k dolnímu kvartilu, což svědčí o kladně zešikmeném rozložení. Vyskytují se odlehlé i extrémní hodnoty, jedná se tedy o špičaté rozložení.
10. Pro proměnnou STARIAUT sestrojte NP plot a s jeho pomocí posuďte normalitu této proměnné.
    Návod: Graphs – 2D Graphs – Normal Probability Plots – Variables STARIAUT – OK.
    Interpretace: Vzhled NP plot svědčí o kladně zešikmeném rozložení, nejedná se tedy o normální rozložení.
11. Pro proměnnou STARIAUT nakreslete histogram s proloženou hustotou normálního rozložení. Ponechejte implicitní počet třídicích intervalů.
    Návod: Graphs – Histograms – Variables STARIAUT – OK.
    Interpretace: Tvar histogramu svědčí o kladně zešikmeném rozložení, jehož hustota neodpovídá hustotě normálního rozložení.