Průzkumová analýza jednorozměrných dat, diagnostické grafy

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.

Data zkoumáme pomocí funkcionálních a číselných charakteristik a pomocí diagnostických grafů.

Osnova:

- datový soubor
- bodové a intervalové rozložení četností
- typy znaků, číselné charakteristiky znaků
- krabicový diagram, N-P plot, P-P plot, Q-Q plot, histogram

Funkcionální charakteristiky datového souboru

Označení

Na množině objektů $\varepsilon_1, \ldots, \varepsilon_n$ zjišťujeme hodnoty znaku X (např. u 6 domácností zjišťujeme počet členů).
Hodnotu znaku X na objektu ε_i označíme x_i, i = 1, ..., n.

Tyto hodnoty zaznamenáme do jednorozměrného datového souboru $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ (např. $\begin{pmatrix} 2 \\ 1 \\ 2 \\ 3 \\ 1 \\ 2 \end{pmatrix}$).

Uspořádané hodnoty $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ tvoří uspořádaný datový soubor $\begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{pmatrix}$, v našem případě $\begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \end{pmatrix}$.

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, v našem případě $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$.

Bodové rozložení četností

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

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

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

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

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

Absolutní a relativní četnosti zapisujeme do tabulky rozložení četností nebo je 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: $F(x) = \begin{cases} 0 \text{ pro } x < x_{[1]} \\ F_j \text{ pro } x_{[j]} \le x < x_{[j+1]},\ j = 1, \ldots, r\text{-}1 \\ 1 \text{ pro } x \ge x_{[r]} \end{cases}$

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

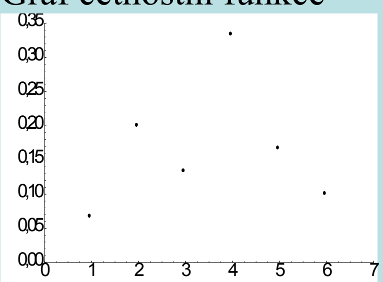
Počet členů	1	2	3	4	5	6
Počet domácností	2	6	4	10	5	3

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

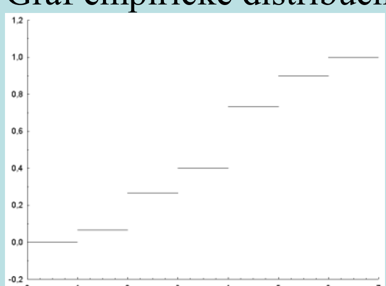
Řešení:
Tabulka rozložení četností

$x_{[j]}$	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

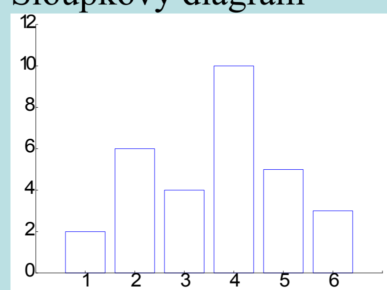
Graf četnostní funkce



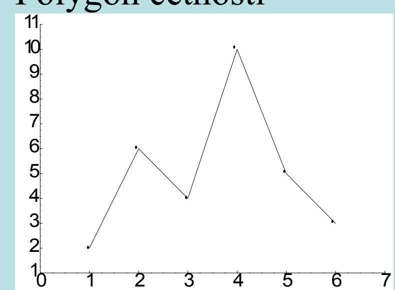
Graf empirické distribuční funkce



Sloupkový diagram



Polygon četností



Intervalové rozložení četností

Je-li počet variant znaku X 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 u bodového rozložení četností, 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: $f(x) = \begin{cases} f_j & \text{pro } u_j < x \le u_{j+1}, j = 1, \dots, r \\ 0 & \text{jinak} \end{cases}$ (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	(35,65]	(65,95]	(95,125]	(125,155]	(155,185]	(185,215]
Počet dom.	7	16	27	14	4	2

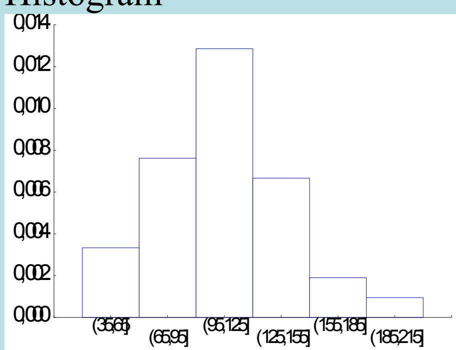
Sestavte tabulku rozložení četností, nakreslete histogram a graf intervalové empirické distribuční funkce.

Řešení:

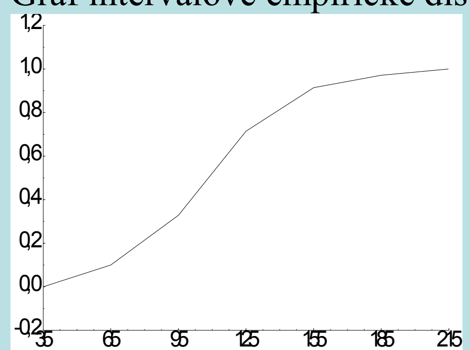
Tabulka rozložení četností

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

Histogram



Graf intervalové empirické distribuční funkce



Číselné charakteristiky datového souboru

Znaky nominálního typu

Tyto znaky umožňují obsahovou interpretaci pouze u relace rovnosti.
Příklady nominálních znaků: lékařská diagnóza, typ profese, barva očí, rodinný stav, národnost, …
Charakteristikou polohy je modus, tj. nejčetnější varianta či střed nejčetnějšího intervalu.

Znaky ordinálního typu

Lze u nich navíc obsahově interpretovat relaci uspořádání.
Příklad ordinálního znaku: školní klasifikace vyjadřuje menší nebo větší znalosti zkoušených žáků – jedničkář je lepší než dvojkař, ale intervaly mezi známkami nemají obsahovou interpretaci. Nelze tvrdit, že rozdíl ve znalostech mezi jedničkářem a dvojkařem je stejný jako mezi trojkařem a čtyřkařem.
Další příklady: Různá bodování ve sportovních a uměleckých soutěžích, posuzování různých rysů sociálního chování, posuzování stavu pacientů, hodnocení postojů respondentů k různým otázkám, …

Charakteristikou polohy je α-kvantil. Je-li $\alpha \in$ …, 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:

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

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	$n\alpha$	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

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o 2 proměnných a 11 případech. První proměnnou nazveme X, druhou cetnost a zapíšeme do nich počet bodů a odpovídající absolutní četnosti.

Statistiky – Základní statistiky/tabulky – Popisné statistiky – zapneme proměnnou vah cetnost – OK – OK – Proměnné X – OK – Detailní výsledky – vybereme Medián, Dolní a horní kvartily, Kvantilové hranice – Výpočet – ve výstupní tabulce upravíme počet desetinných míst.

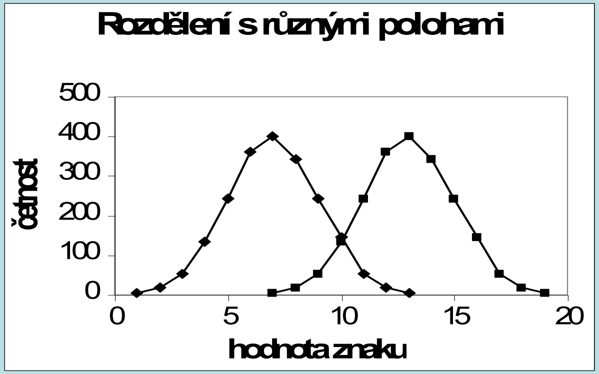
	Popisne statistiky (pocet bodu.sta)					
Promě	N platn	Media	Spod kvarti	Horn kvarti	Kvant 10,00	Kvant 90,00
X	10	6	4	7	2	8

Znaky intervalového a poměrového typu

U těchto znaků lze navíc obsahově interpretovat operaci rozdílu resp. podílu.
Příklad intervalového znaku: teplota měřená ve stupních Celsia. Např. naměříme-li ve čtyřech po sobě jdoucích dnech polední teploty 0, 2, 4, 6 ºC, znamená to, že každým dnem stouply teploty o 2 ºC. Nelze však říci, že z druhého na třetí den vzrostla teplota dvojnásobně, kdežto ze třetího na čtvrtý den pouze jeden a půl krát.
Další příklady: kalendářní systémy, směr větru, inteligenční kvocient, …
Společný znak intervalových znaků: nula byla stanovena uměle, pouhou konvencí.
Příklad poměrového znaku: délka předmětu měřená v cm. Má-li jeden předmět délku 8 cm a druhý 16 cm, má smysl prohlásit, že druhý předmět je dvakrát delší než první předmět.
Další příklady: počet dětí v rodině, výška kapesného v Kč, hmotnost osoby, …
Společný znak poměrových znaků: poměrový znak má přirozený počátek, ke kterému jsou vztahovány všechny další hodnoty znaku.

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 $\sqrt[n]{x_1 \cdot \ldots \cdot x_n}$.
Pomocí průměru zavedeme i-tou centrovanou hodnotu $x_i - m$ (podle znaménka poznáme, zda i-tá hodnota je podprůměrná či nadprůměrná).
Znázornění rozložení četností dvou datových souborů, které se liší aritmetickým průměrem



Vlastnosti aritmetického průměru

- Aritmetický průměr si lze představit jako těžiště dat – součet podprůměrných hodnot je stejný jako součet nadprůměrných hodnot – oba součty jsou v rovnováze.

- Průměr centrovaných hodnot je nulový, protože $\frac{1}{n}\sum^{n} \ldots = \frac{1}{n}\sum^{n} \ldots \frac{1}{n}\sum^{n} \ldots - \ldots \frac{1}{n}\, n \ldots = 0$.

- Výraz $\sum_{i=1}^{n} (x_i - a)^2$ (tzv. kvadratická odchylka) nabývá svého minima pro a = m. Uvedený výraz charakterizuje celkovou chybu, které se dopustíme, když datový soubor nahradíme jedinou hodnotou a. Tato chyba je tedy nejmenší, když datový soubor nahradíme aritmetickým průměrem, přičemž za míru chyby považujeme kvadratickou odchylku.

- Pokud každou hodnotu x_i podrobíme lineární transformaci $y_i = a + bx_i$, pak průměr transformovaných hodnot je roven lineární transformaci původního průměru, tj. $m_2 = a + bm_1$.

- Mají-li znaky X, Y průměry m_1, m_2, pak znak Z = X + Y má průměr $m_1 + m_2$.

- Aritmetický průměr je silně ovlivněn extrémními hodnotami.

- Aritmetický průměr je vhodné použít, pokud je rozložení dat přibližně symetrické.

Příklad na vlastnosti aritmetického průměru:

U skupiny 20 pracovníků v určité dílně byly zjišťovány měsíční mzdy. Průměr mezd činil 15 500 Kč. Určete průměr mezd, jestliže mzdy všech pracovníků se zvýší

a) o 300 Kč, b) 1,1 krát, c) o 20%.

Řešení:

Označme m_1 průměr hodnot $x_1, \ldots, x_n$ a m_2 průměr hodnot $y_1, \ldots, y_n$, přičemž $y_i = a + bx_i$, $i = 1, \ldots, n$. Pak $m_2 = a + bm_1$.

ad a) $m_2 = 300 + m_1 = 15\ 800$

Průměr se zvýšil o 300 Kč na 15 800 Kč.

ad b) $m_2 = 1{,}1.m_1 = 17\ 050$

Průměr se zvýšil na 17 050 Kč.

ad c) $m_2 = 1{,}2.m_1 = 18\ 600$

Průměr se zvýšil na 18 600 Kč.

Charakteristiky variability intervalových a poměrových znaků

Variační rozpětí $R = x_{(n)} - x_{(1)}$ (nevýhoda – bere v úvahu pouze nejmenší a největší hodnotu datového souboru),

průměrná odchylka $o = \frac{1}{n}\sum_{i=1}^{n} |x_i - m|$ (udává, o kolik jednotek se data liší od průměru)

rozptyl $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - m)^2$ (nevýhoda – vychází ve druhých mocninách jednotek, v nichž byl měřen znak X)

směrodatná odchylka $s = \sqrt{s^2}$.

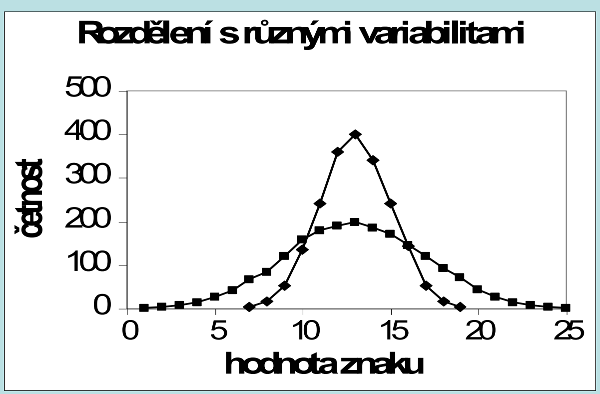
Pomocí směrodatné odchylky zavedeme i-tou standardizovanou hodnotu $\frac{x_i - m}{s}$ (vyjadřuje, o kolik směrodatných odchylek se i-tá hodnota odchýlila od průměru).

U poměrových znaků se jako charakteristika variability používá též:

koeficient variace $\frac{s}{m}$ (často se udává v procentech a udává, kolika procent průměru dosahuje směrodatná odchylka),

relativní průměrná odchylka $\frac{o}{m}$ (při vyjádření v procentech udává, kolika procent průměru dosahuje průměrná odchylka)

Znázornění rozložení četností dvou datových souborů, které se liší rozptylem:



Vlastnosti rozptylu:

- Rozptyl je nulový pouze tehdy, když jsou všechny hodnoty stejné, jinak je kladný.

- Rozptyl centrovaných hodnot je roven původnímu rozptylu, neboť $\frac{1}{n}\sum_{i=1}^{n}\left((x_i - m) - 0\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = s^2$.

- Rozptyl standardizovaných hodnot je 1, protože $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m}{s} - 0\right)^2 = \frac{1}{s^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{s^2}{s^2} = 1$.

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

- Pokud každou hodnotu x_i podrobíme lineární transformaci $y_i = a + bx_i$, pak rozptyl transformovaných hodnot je roven původnímu rozptylu vynásobenému b^2, tj. $s_2^2 = b^2 s_1^2$.

- Rozptyl je stejně jako průměr silně ovlivněn extrémními hodnotami.

- Rozptyl se nehodí jako charakteristika variability, je-li rozložení dat nesymetrické.

Příklad 4.: Kurzy akcií společnosti AAA Auto Group v průběhu 23 dní v měsíci srpnu 2010 byly následující: 17,75; 17,74; 17,85; 17,59; 17,92; 17,98; 18,39; 18,25; 18,30; 18,00; 18,15; 18,15; 18,22; 18,40; 18,25; 17,95; 18,25; 18,23; 17,95; 17,90; 17,80; 17,87; 17,87. Vypočtěte charakteristiky variability.

Řešení:

Nejprve vypočítáme variační rozpětí: $R = x_{max} - x_{min} = 18,40 - 17,59 = 0,81$.

Před výpočtem dalších charakteristik variability musíme získat aritmetický průměr: $\bar{x} = \frac{1}{23}(17,75 + 17,74 + \dots + 17,87) = 18,03$.

Průměrná odchylka: $d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{23}(|17,75 - 18,03| + |17,74 - 18,03| + \dots + |17,87 - 18,03|) = 0,1965$

Relativní průměrná odchylka: $\frac{d}{\bar{x}}100\% = \frac{0,1965}{18,03}100\% = 1,09\%$

Rozptyl: $s^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 = \frac{1}{23}(17,75^2 + 17,74^2 + \dots + 17,87^2) - 18,03^2 = 0,049$

Směrodatná odchylka: $s = \sqrt{s^2} = \sqrt{0,049} = 0,221$

Koeficient variace: $\frac{s}{\bar{x}}100\% = \frac{0,2213}{18,03}100\% = 1,23\%$

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné X a 23 případech. Do proměnné X zapíšeme zjištěné kurzy akcií. Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – vybereme Průměr, Rozptyl, Rozpětí – Výpočet. Ve výstupní tabulce přidáme za proměnnou Rozptyl tři nové proměnné nazvané rozptyl, směr. odch. a koef. variace. Do Dlouhého jména proměnné rozptyl napíšeme =v3*22/23, Dlouhého jména proměnné směr. odch. napíšeme =sqrt(v4) a do Dlouhého jména proměnné koef. variace napíšeme =100*v5/v1.

Proměn	Prům	Rozp	Rozpě	rozpty =v3*2	směr. od =sqrt(v	koef. var =100*v5
x	18,03	0,810	0,051	0,049	0,22136	1,22750

Pro výpočet průměrné odchylky a relativní průměrné odchylky je zapotřebí přidat k původnímu datovému souboru dvě nové proměnné nazvané Průměr a Odchylka. Do Dlouhého jména proměnné Průměr napíšeme =18,033 a do Dlouhého jména proměnné Odchylka napíšeme =abs(v1-v2). Nyní spočteme průměr proměnné Odchylka: Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné Odchylka – OK – Detailní výsledky – vybereme Průměr – Výpočet.
Ve výstupní tabulce přejmenujeme proměnnou Průměr na prům. odch. a za tuto proměnnou přidáme proměnnou rel. prům. odch. Do jejího Dlouhého jména napíšeme =100*v1/18,033.

Proměn	odchy	rel. prům. =100*v1/1
Odchy	0,196	1,08954

Vážené číselné charakteristiky

Známe-li absolutní četnosti $n_1, \ldots, n_r$ či relativní četnosti $p_1, \ldots, p_r$ variant $x_{[1]}, \ldots, x_{[r]}$, můžeme spočítat

vážený průměr $m = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]} = \sum_{j=1}^{r} p_j x_{[j]}$,

vážený rozptyl $s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j \left(x_{[j]} - m\right)^2 = \sum_{j=1}^{r} p_j \left(x_{[j]} - m\right)^2$ (výpočetní vzorec: $s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}^2 - m^2 = \sum_{j=1}^{r} p_j x_{[j]}^2 - m^2$),

váženou průměrnou odchylku $o = \frac{1}{n}\sum_{j=1}^{r} n_j \left|x_{[j]} - m\right| = \sum_{j=1}^{r} p_j \left|x_{[j]} - m\right|$.

Příklad 5.: U 35 zaměstnanců byl zjištěn počet odpracovaných hodin za měsíc.

Počet odpracovaných hodin	184	185	186	187	188	189
Počet zaměstnanců	4	6	7	6	7	5

Vypočtěte průměr, průměrnou odchylku, relativní průměrnou odchylku, směrodatnou odchylku a koeficient variace počtu odpracovaných hodin.

Řešení:

Vážený průměr: $m = \frac{1}{n}\sum_{i=1}^{r} x_i n_i = \frac{1}{35}(4\cdot 184 + 6\cdot 185 + 7\cdot 186 + 6\cdot 187 + 7\cdot 188 + 5\cdot 189) = 186{,}6$

Vážená průměrná odchylka:

$o = \frac{1}{n}\sum_{i=1}^{r} |x_i - m| n_i = \frac{1}{35}(4\cdot|184-186{,}6| + 6\cdot|185-186{,}6| + 7\cdot|186-186{,}6| + 6\cdot|187-186{,}6| + 7\cdot|188-186{,}6| + 5\cdot|189-186{,}6|) = 1{,}38$ h = 1h 23 min

Vážený rozptyl: $s^2 = \frac{1}{n}\sum_{i=1}^{r} x_i^2 n_i - m^2 = \frac{1}{35}(4\cdot 184^2 + 6\cdot 185^2 + 7\cdot 186^2 + 6\cdot 187^2 + 7\cdot 188^2 + 5\cdot 189^2) - 186{,}6^2 = 2{,}525$

Vážená směrodatná odchylka: $s = \sqrt{s^2} = \sqrt{2{,}525} = 1{,}59$ h = 1h 35 min

Relativní průměrná odchylka: $\frac{o}{m}100\% = \frac{1{,}38}{186{,}6}100\% = 0{,}74\%$

Koeficient variace: $\frac{s}{m}100\% = \frac{1{,}59}{186{,}6}100\% = 0{,}85\%$

Vidíme, že zaměstnanci odpracovali za měsíc v průměru 186,6 h, přičemž průměrná odchylka dosahuje 0,74 % průměrné odpracované doby a směrodatná odchylka dosahuje 0,85 % průměrné odpracované doby.

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o 2 proměnných a 6 případech. První proměnnou nazveme X, druhou četnost a zapíšeme do nich počet odpracovaných hodin a odpovídající počty zaměstnanců.

Statistiky – Základní statistiky/tabulky – Popisné statistiky – zapneme proměnnou vah četnost – OK – OK – Proměnné X – OK – Detailní výsledky – vybereme Průměr, Rozptyl – Výpočet. Ve výstupní tabulce přidáme za proměnnou Rozptyl dvě nové proměnné nazvané směr. odch. a koef. variace. Do Dlouhého jména proměnné směr. odch. napíšeme =sqrt(v2*34/35) a do Dlouhého jména proměnné koef. variace napíšeme =100*v3/v1.

Promě	Prům	Rozp	směr.od =sqrt(v2*	koef. va =100*v3
X	186	2,	1,589	0,8516

Pro výpočet průměrné odchylky a relativní průměrné odchylky je zapotřebí přidat k původnímu datovému souboru dvě nové proměnné nazvané Průměr a Odchylka. Do Dlouhého jména proměnné Průměr napíšeme =186,6 a do Dlouhého jména proměnné Odchylka napíšeme =abs(v1-v3). Nyní spočteme průměr proměnné Odchylka: Statistiky – Základní statistiky/tabulky – Popisné statistiky – zapneme proměnnou vah četnost – OK – OK – Proměnné Odchylka – OK – Detailní výsledky – vybereme Průměr – Výpočet. Ve výstupní tabulce přejmenujeme proměnnou Průměr na prům. odch. a za tuto proměnnou přidáme proměnnou rel. prům. odch. Do jejího Dlouhého jména napíšeme =100*v1/186,6.

Promě	prům odch	rel. prům odch. =100*v1/
Odchy	1,382	0,74108

Převod desetinných částí hodiny na minuty můžeme provést např. pomocí aplikace na adrese http://www.prevody-jednotek.cz/.

Počáteční a centrální momenty

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, ... ,

k-tý centrální moment

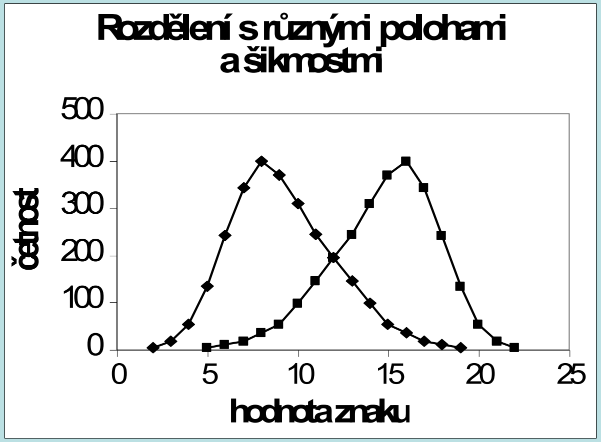
$m_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - m)^k$, k = 1, 2, ...

Pomocí 3. a 4. počátečního momentu se definuje šikmost a špičatost.

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

Je-li rozložení dat symetrické kolem aritmetického průměru, pak $\alpha_3 = 0$.
Má-li rozložení dat prodloužený pravý konec, jde o kladně zešikmené rozložení, $\alpha_3 > 0$.
Má-li rozložení dar prodloužený levý konec, jde o záporně zešikmené rozložení, $\alpha_3 < 0$.

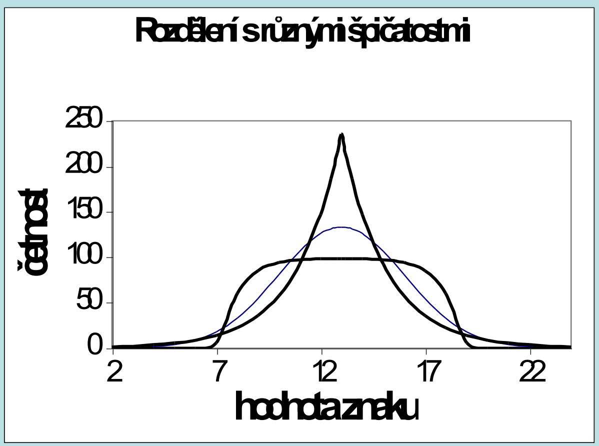
Znázornění rozložení četností dvou datových souborů, které se liší aritmetickým průměrem a šikmostí



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

Je-li rozložení dat normální (Gaussovo), pak $\alpha_4 = 0$.
Je-li rozložení dat strmé, pak $\alpha_4 > 0$.
Je-li rozložení dat ploché, pak $\alpha_4 < 0$.

Znázornění rozložení četností dvou datových souborů, které se liší špičatostí

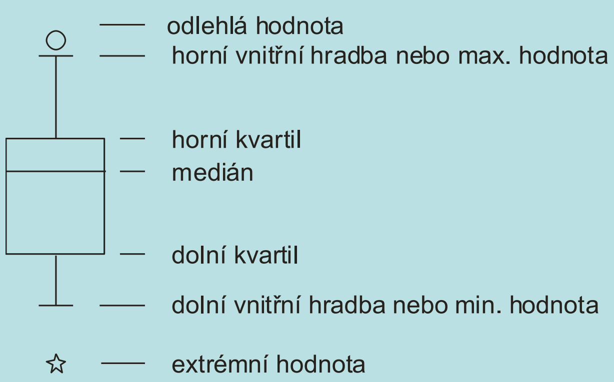


Diagnostické grafy

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 $(-\infty,\ x_{0,25} - 3q)$.

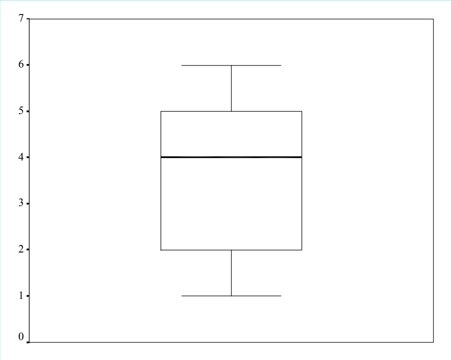
Příklad 6.: Pro údaje z příkladu 1 sestrojte krabicový diagram.
Řešení:

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

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

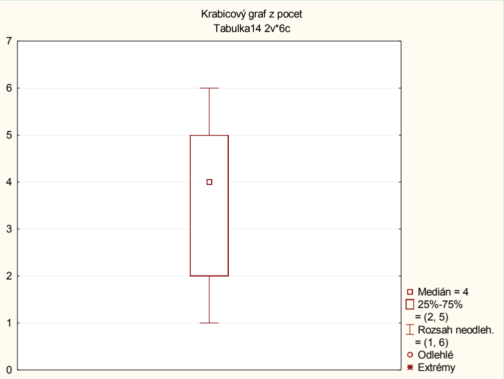
α	$n\alpha$	c		x_α
0,25	7,5	8	$x_{(c)}=x_{(8)}$	2
0,50	15	15	$\frac{x_{(15)}+x_{(16)}}{2}$	4
0,75	22,5	23	$x_{(c)}=x_{(23)}$	5

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

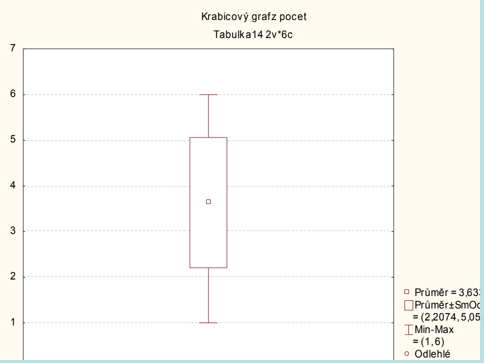


Vidíme, že datový soubor vykazuje určitou nesymetrii – medián je posunut směrem k hornímu kvartilu, soubor je tedy záporně sešikmen. V souboru se nevyskytují žádné odlehlé ani extrémní hodnoty.

Výpočet pomocí systému STATISTICA:
Otevřeme nový datový soubor o 2 proměnných a 6 případech. První proměnnou nazveme počet, druhou cetnost a zapíšeme do nich počet členů domácnosti a odpovídající absolutní četnosti. Zvolíme Grafy – 2D Grafy – Krabicové grafy.
Zapneme proměnnou vah cetnost, zadáme závisle proměnnou pocet a dostaneme krabicový diagram:



Upozornění: Máme-li data intervalového či poměrového charakteru, o nichž lze předpokládat, že pocházejí z nějakého symetrického rozložení (například normálního), je možné použít jinou variantu krabicového diagramu: bod či čára uvnitř krabice reprezentuje průměr, vodorovné hrany krabice jsou ve výšce průměr ± směrodatná odchylka a svorky končí v minimu či maximu.
V našem případě dostaneme krabicový diagram:



Před uvedením dalších diagnostických grafů je nutné zavést pojem pořadí čísla v posloupnosti čísel.

Pojem pořadí

Nechť $x_1, \dots, x_n$ je posloupnost reálných čísel.

a) Jsou-li čísla navzájem různá, pak pořadím R_i čísla x_i rozumíme počet těch čísel $x_1, \dots, x_n$, která jsou menší nebo rovna číslu x_i.

b) Vyskytují-li se mezi danými čísly skupinky stejných čísel, pak každé takové skupince přiřadíme průměrné pořadí.

Příklad na stanovení pořadí

a) Jsou dána čísla 9, 4, 5, 7, 3, 1. Stanovte pořadí těchto čísel.

b) Jsou dána čísla 6, 7, 7, 9, 6, 10, 8, 6, 6, 9.

Řešení

ad a)

usp. čísla	1	3	4	5	7	9
pořadí	1	2	3	4	5	6

ad b)

usp. čísla	6	6	6	6	7	7	8	9	9	10
pořadí	1	2	3	4	5	6	7	8	9	10
prům. pořadí	2,5	2,5	2,5	2,5	5,5	5,5	7	8,5	8,5	10

Normální pravděpodobnostní graf (N-P plot)

N- P plot umožňuje graficky posoudit, zda data pocházejí z normálního rozložení.

Způsob konstrukce:

Na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$,

na svislou osu kvantily u_α standardizovaného normálního rozložení, kde $\alpha = \frac{3j-1}{3n+1}$, přičemž j je pořadí j-té uspořádané hodnoty (jsou-li některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající takové skupince).

Pocházejí-li data z normálního rozložení, pak všechny dvojice $(x_{(j)}, u_\alpha)$ budou ležet na přímce.

Pro data z rozložení s kladnou šikmostí se dvojice $(x_{(j)}, u_\alpha)$ budou řadit do konkávní křivky,

pro data z rozložení se zápornou šikmostí se dvojice $(x_{(j)}, u_\alpha)$ budou řadit do konvexní křivky.

Příklad na konstrukci N – P plotu:
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 normálního pravděpodobnostního grafu posuďte, zda se tato data řídí normálním rozložením.

Řešení:

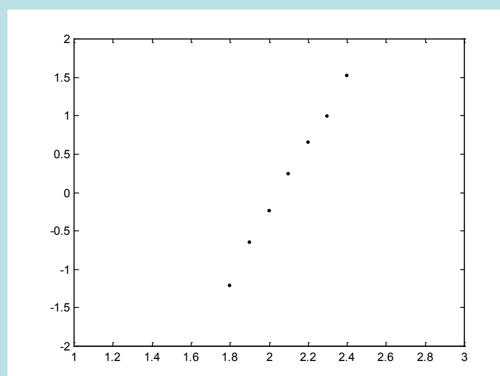
usp. hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

Vektor hodnot průměrného pořadí: j = (1,5 3 4,5 6,5 8 9 10),

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

vektor kvantilů $u_\alpha = (-1{,}211\ \ -0{,}649\ \ -0{,}245\ \ 0{,}245\ \ 0{,}649\ \ 0{,}989\ \ 1{,}517)$.

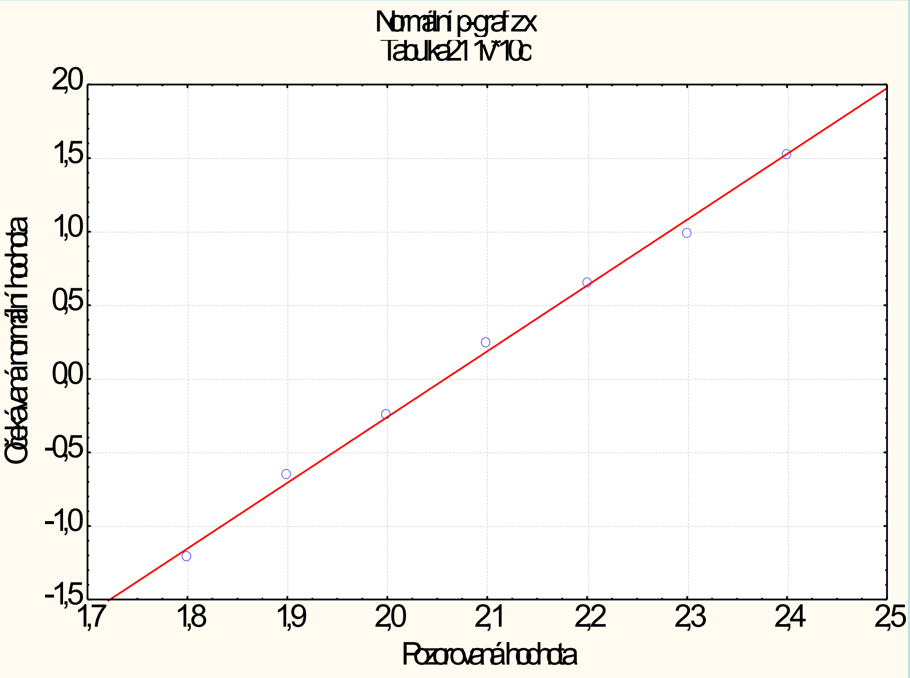
Normální pravděpodobnostní graf



Protože dvojice $(x_{(j)}, u_\alpha)$ téměř leží na přímce, lze usoudit, že data pocházejí z normálního rozložení.

Výpočet pomocí systému STATISTICA:
Otevřeme nový datový soubor o jedné proměnné a 10 případech. Zjištěné hodnoty zapíšeme do proměnné X.

Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnná X – OK - odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK.



Quantile - quantile plot (Q-Q plot)

Umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení (např. STATISTICA nabízí 8 typů rozložení: beta, exponenciální, Gumbelovo, gamma, log-normální, normální, Rayleighovo a Weibulovo).

Způsob konstrukce:

na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \leq ... \leq x_{(n)}$,

na vodorovnou osu kvantily $K_{\alpha}(X)$ vybraného rozložení, kde $\alpha = \frac{j - r_{adj}}{n + n_{adj}}$, přičemž r_{adj} a n_{adj} jsou korigující faktory $\leq 0{,}5$, implicitně $r_{adj} = 0{,}375$ a $n_{adj} = 0{,}25$. (Jsou-li některé hodnoty $x_{(1)} \leq ... \leq x_{(n)}$ 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 $\left(K_{\alpha}(X), x_{(j)}\right)$ se metodou nejmenších čtverců proloží přímka. Čím méně se body odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.

Příklad na konstrukci Q-Q plotu: 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 Q-Q plotu ověřte, zda se tato data řídí normálním rozložením.

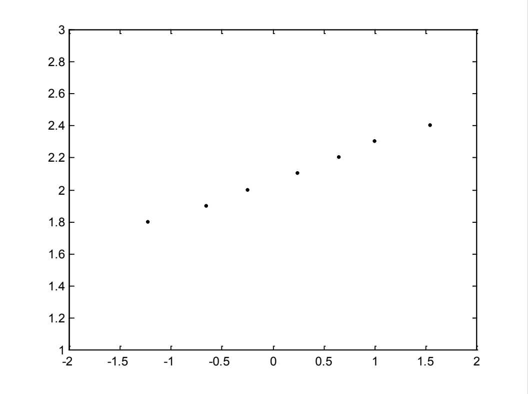
Řešení:

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

Vektor hodnot průměrného pořadí: j = (1,5 3 4,5 6,5 8 9 10)

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

vektor kvantilů $u_\alpha = (-1,227\ \ -0,655\ \ -0,247\ \ 0,247\ \ 0,655\ \ 1,000\ \ 1,56)$

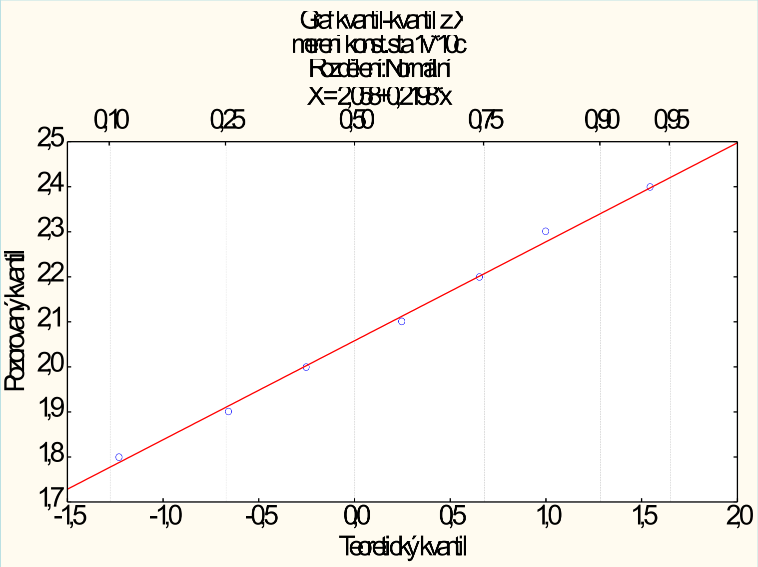


Vzhled grafu nasvědčuje tomu, že data pocházejí z normálního rozložení.

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné a 10 případech. Zjištěné hodnoty zapíšeme do proměnné X.

Grafy – 2D Grafy – Grafy typu Q-Q– Proměnná X – OK - odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK.



Probability - probability plot (P-P plot)

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

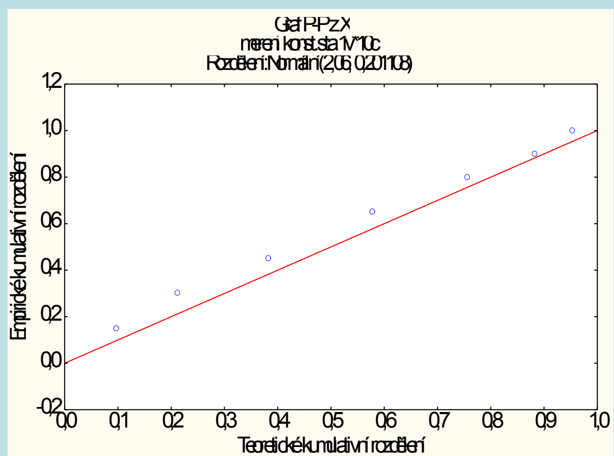
Způsob konstrukce: spočtou se standardizované hodnoty $z_{(j)} = \frac{x_{(j)} - \bar{x}}{s}$, j = 1, ..., n. Na vodorovnou osu se vynesou hodnoty teoretické distribuční funkce $\Phi(z_{(j)})$ a na svislou osu hodnoty empirické distribuční funkce $F(z_{(j)}) = j/n$. (Jsou-li některé hodnoty $x_{(1)} \leq ... \leq x_{(n)}$ 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] x [0,1], lze usuzovat na dobrou shodu empirického a teoretického rozložení.

Příklad na konstrukci P-P plotu pomocí systému STATISTICA: 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 P-P plotu ověřte, zda se tato data řídí normálním rozložením.

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné a 10 případech. Zjištěné hodnoty zapíšeme do proměnné X.

Grafy – 2D Grafy – Grafy typu P-P – Proměnná X – OK - odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK.



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: na vodorovnou osu vynášíme meze třídicích intervalů. Nad každým třídicím intervalem sestrojíme obdélník o ploše odpovídající relativní četnosti příslušného třídicího intervalu, tj. výška obdélníku je rovna četnostní hustotě třídicího intervalu (četnostní hustota je relativní četnost třídicího intervalu dělená délkou tohoto intervalu).

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.

Příklad na konstrukci histogramu:

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

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

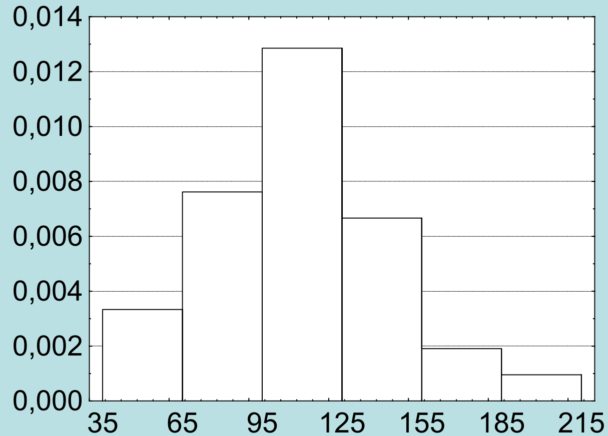
Nakreslete histogram.

Řešení:

Nejprve sestavíme tabulku rozložení četností:

$\langle u_j, u_{j+1})$	$x_{[j]}$	d_j	n_j	p_j	N_j	F_j	f_j
$\langle 35,65)$	50	30	7	7/70=0,1	7	7/70=0,1	7/2100=0,0033
$\langle 65,95)$	80	30	16	16/70=0,23	23	23/70=0,33	16/2100=0,0076
$\langle 95,125)$	110	30	27	27/70=0,38	50	50/70=0,71	23/2100=0,0109
$\langle 125,155)$	140	30	14	14/70=0,2	64	64/70=0,91	14/2100=0,0067
$\langle 155,185)$	170	30	4	4/70=0,06	68	68/70=0,97	4/2100=0,0019
$\langle 185,215)$	200	30	2	2/70=0,03	70	70/70=1	2/2100=0,00010

S pomocí této tabulky sestrojíme histogram:

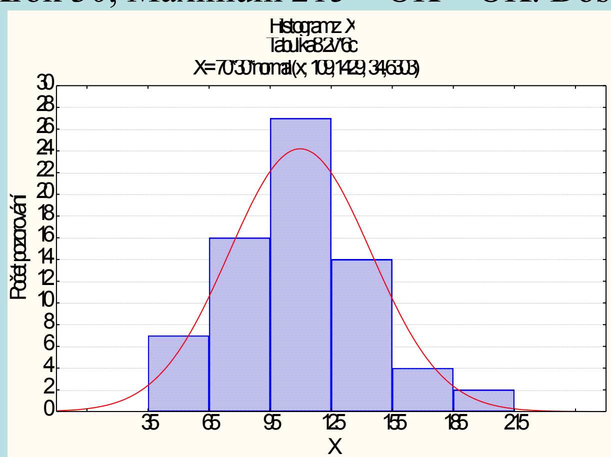


Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných a 6 případech. První proměnnou nazveme X, druhou cetnost. Do proměnné X napíšeme středy třídicích intervalů, do proměnné cetnost odpovídající absolutní četnosti:

	1 X	2 cetnos
1	5	7
2	8	1
3	11	2
4	14	1
5	17	4
6	20	2

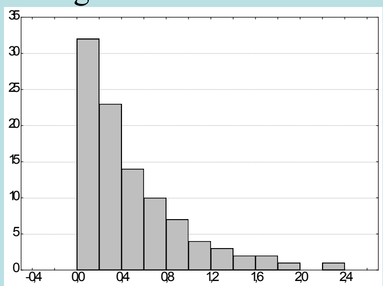
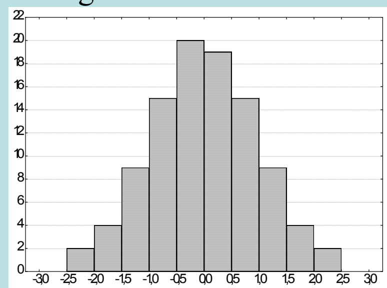
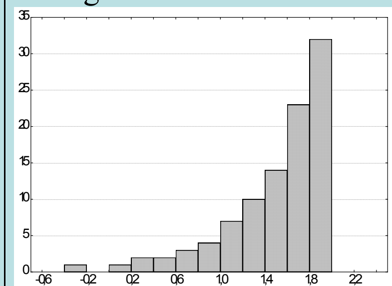
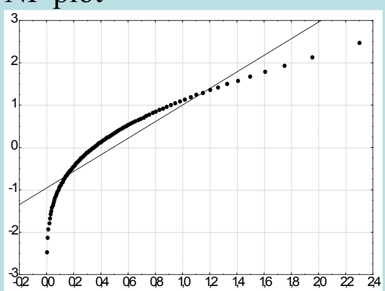
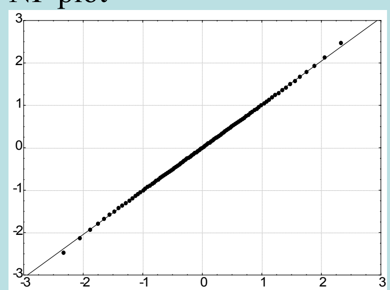
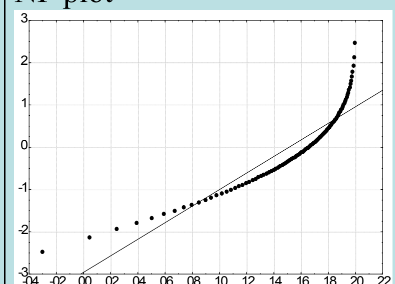
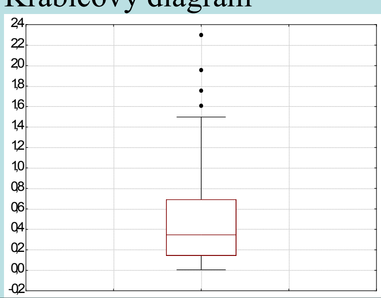
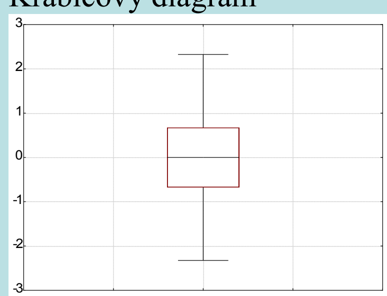
Grafy – Histogramy – zadáme proměnnou vah cetnost – Proměnná X - zaškrtneme Hranice – Určit hranice – zaškrtneme Zadejte hraniční rozmezí: Minimum 35, Krok 30, Maximum 215 – OK – OK. Dostaneme graf:



Na rozdíl od histogramu konstruovaného ručně jsou na svislé ose absolutní četnosti, nikoliv četnostní hustoty. V porovnání s grafem hustoty normálního rozložení je vidět, že naše rozložení četností je lehce kladně zešikmené. Naše data tedy nepocházejí z normálního rozložení.

Vzhled diagnostických grafů pro rozložení s různou šikmostí

Pro ilustraci se podívejme, jak se různá šikmost rozložení projeví na histogramu, N-P plotu a na krabicovém diagramu.

Rozložení s kladnou šikmostí	Normální rozložení	Rozložení se zápornou šikmostí
Histogram 	Histogram 	Histogram 
NP plot 	NP plot 	NP plot 
Krabicový diagram 	Krabicový diagram 	Krabicový diagram