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.

Funkcionální charakteristiky datového souboru

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.

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 $x_{[j]}$

$p_j = \frac{n_j}{n}$ – 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

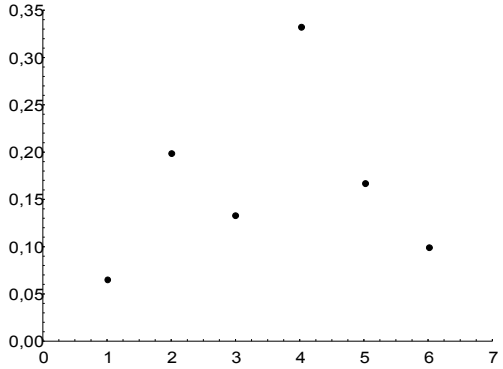
Vytvořte tabulku rozložení četností. Nakreslete grafy četnostní funkce a empirické distribuční funkce. Dále nakreslete sloupkový diagram a polygon četností počtu členů domácnosti.

Řešení:

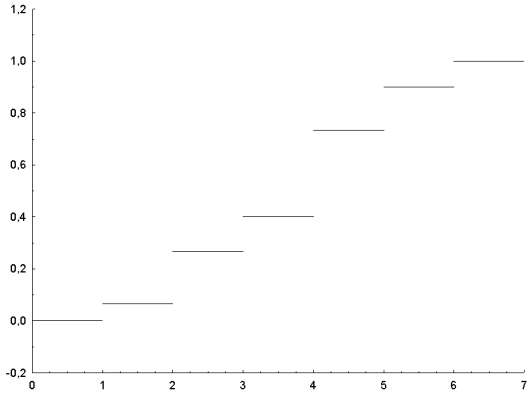
Tabulka rozložení četností

$x_{[j]}$	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

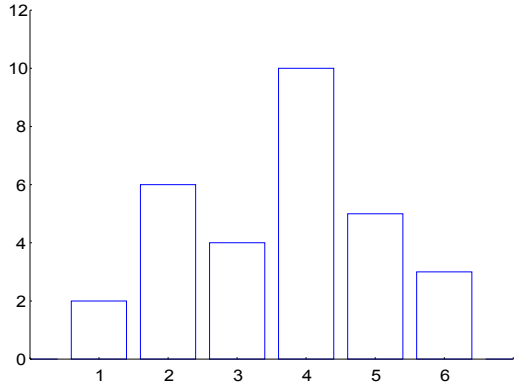
Graf četnostní funkce



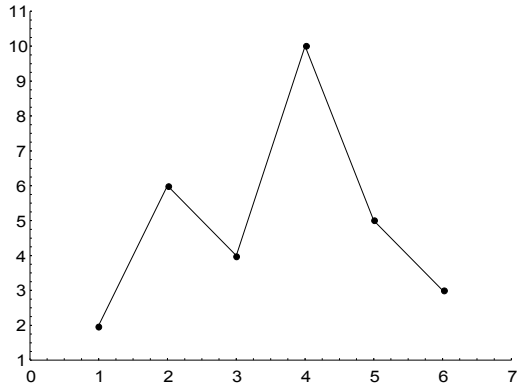
Graf empirické distribuční funkce



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

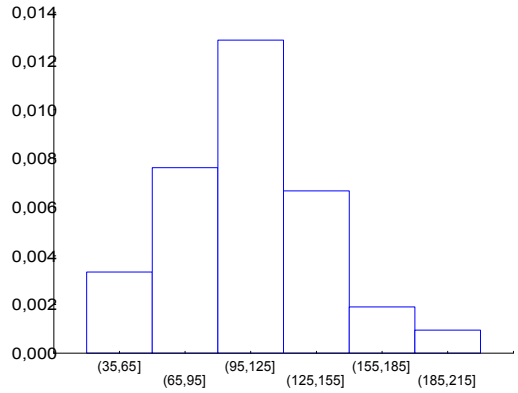
Výdaje	$\langle 35,65\rangle$	$\langle 65,95\rangle$	$\langle 95,125\rangle$	$\langle 125,155\rangle$	$\langle 125,155\rangle$	$\langle 185,215\rangle$
Počet dom.	7	16	27	14	4	2

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

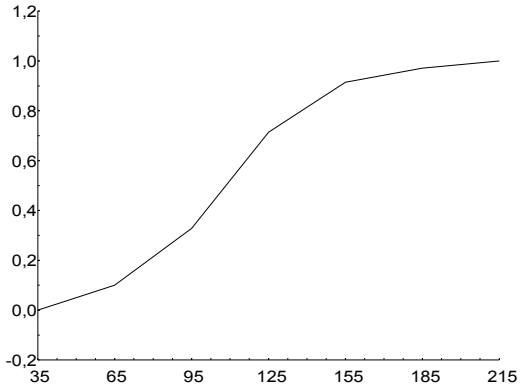
Řešení:
Tabulka rozložení četností

$(u_j,u_{j+1}]$	n_j	p_j	f_j	N_j	F_j
$\langle 35,65\rangle$	7	7/70	7/2100	7	7/70
$\langle 65,95\rangle$	16	16/70	16/2100	23	23/70
$\langle 95,125\rangle$	27	27/70	27/2100	50	50/70
$\langle 125,155\rangle$	14	14/70	14/2100	64	64/70
$\langle 155,185\rangle$	4	4/70	4/2100	68	68/70
$\langle 185,215\rangle$	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 $\alpha \in (0;1)$, 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	$n\alpha$	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

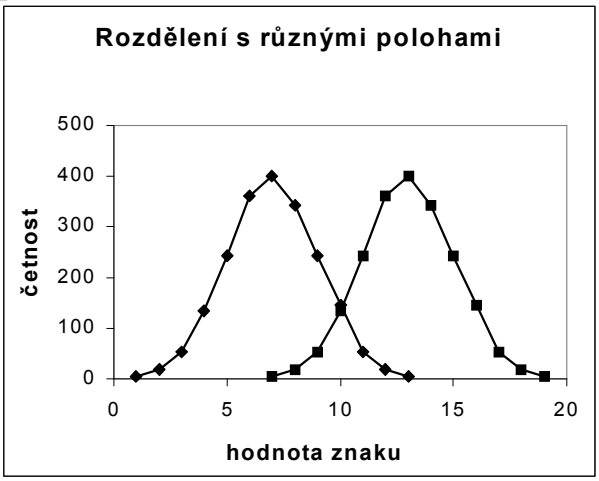
Výpočet pomocí systému STATISTICA:
Otevřeme nový datový soubor o 2 proměnných a 11 případech. První proměnnou nazveme X, druhou cetnost a zapíšeme do nich počet bodů a odpovídající absolutní četnosti.
Statistiky – Základní statistiky/tabulky – Popisné statistiky – zapneme proměnnou vah cetnost – OK – OK – Proměnné X – OK – Detailní výsledky – vybereme Medián, Dolní a horní kvartily, Kvantilové hranice – Výpočet – ve výstupní tabulce upravíme počet desetinných míst.

	Popisné statistiky (pocet bodu.sta)					
Proměnná	N platných	Medián	Spodní kvartil	Horní kvartil	Kvantil 10,00000	Kvantil 90,00000
X	101	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

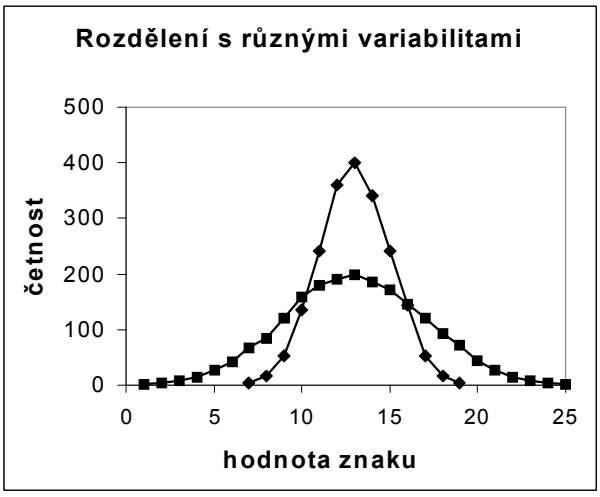
$$\frac{1}{n}\sum_{i=1}^{n}(x_i - m) = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} m = m - \frac{1}{n}\cdot n \cdot m = 0 \; = 0.$$

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.

Aritmetický průměr je silně ovlivněn extrémními hodnotami.
Aritmetický průměr je vhodné použít, pokud je rozložení dat přibližně symetrické.

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 $s^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^{\,2} - m^2$.) 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 $\frac{s}{m}$.
Znázornění rozložení četností dvou datových souborů, které se liší rozptylem:



Vlastnosti rozptylu a směrodatné odchylky:
Směrodatná odchylka je nulová pouze tehdy, když jsou všechny hodnoty stejné, jinak je kladná.
Rozptyl centrovaných hodnot je roven původnímu rozptylu, neboť
$$\frac{1}{n}\sum_{i=1}^{n}[(x_i - m) - 0]^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = s^2$$
Rozptyl standardizovaných hodnot je 1, protože
$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m}{s} - 0\right)^2 = \frac{1}{s^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{s^2}{s^2} = 1$$
Směrodatná odchylka je stejně jako průměr silně ovlivněna extrémními hodnotami.
Směrodatná odchylka se nehodí jako charakteristika variability, je-li rozložení dat zešikmené.

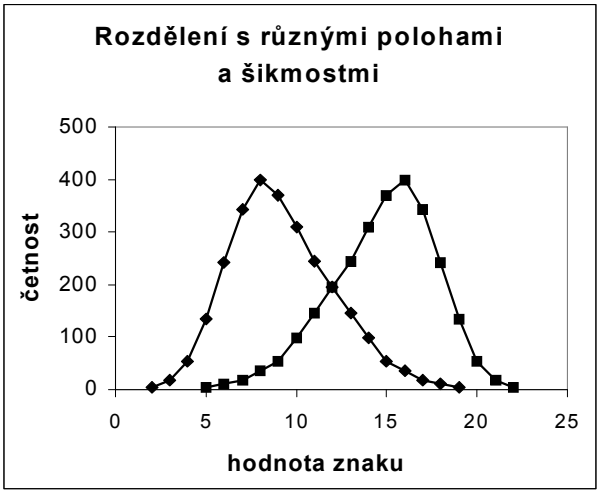
Známe-li absolutní či relativní četnosti variant $x_{[1]}, ..., x_{[r]}$, můžeme spočítat vážený průměr $m = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}$ či vážený rozptyl : $s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j (x_{[j]} - m)^2$. (Vážený rozptyl se zpravidla počítá podle vzorce $s^2 = \frac{1}{n}\sum_{j=1}^{r}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

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.

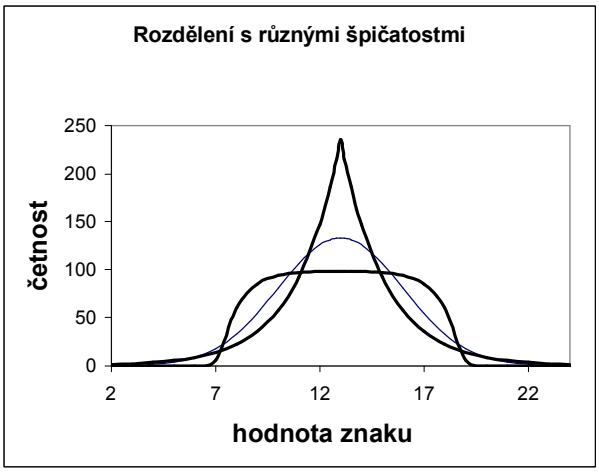
Je-li rozložení dat symetrické kolem aritmetického průměru, pak $\alpha_3 = 0$.
Má-li rozložení dat prodloužený pravý konec, jde o kladně zešikmené rozložení, $\alpha_3 > 0$.
Má-li rozložení dar prodloužený levý konec, jde o záporně zešikmené rozložení, $\alpha_3 < 0$.

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



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

Řešení:

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

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

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

Výpočetní tvar vzorce pro vážený rozptyl: $s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}^2 - m^2$, tedy

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

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o 2 proměnných a 6 případech. První proměnnou nazveme X, druhou cetnost a zapíšeme do nich počet členů domácnosti a odpovídající absolutní četnosti.

Statistiky – Základní statistiky/tabulky – Popisné statistiky – zapneme proměnnou vah cetnost – OK – OK – Proměnné X – OK – Detailní výsledky – vyberem Průměr. Rozptyl – Výpočet.

	Popisné statistiky (domacnosti.sta)		
Proměnná	N platných	Průměr	Rozptyl
X	30	3,633333	2,033333

Rozptyl vyjde jinak, protože STATISTICA používá ve jmenovateli n-1, nikoliv n. Provedeme tedy přepočet: k výsledné tabulce přidáme novou proměnnou a do jejího Dlouhého jména napíšeme =(29/30)*v3

	Popisné statistiky (domacnosti.sta)			
Proměnná	N platných	Průměr	Rozptyl	Jpraveny
X	30	3,633333	2,033333	1,965556

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 $y_1, ..., y_n$.

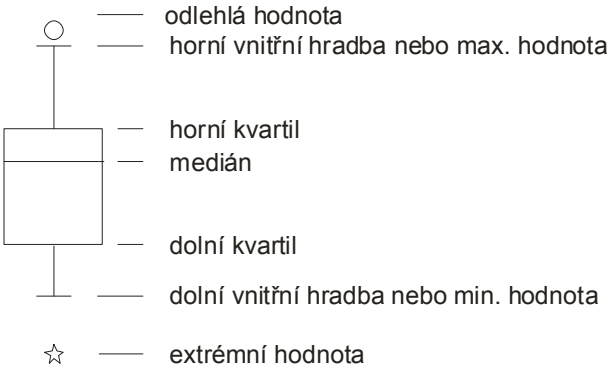
Řešení: $$m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = a + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + bm_1$$

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

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

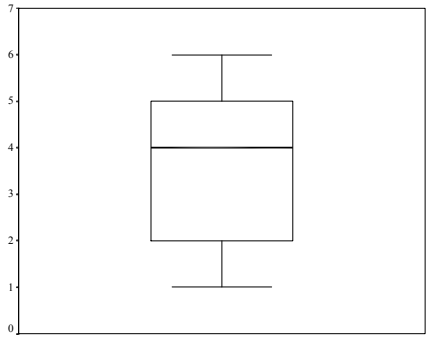
Rozsah souboru n = 30. Výpočty potřebných kvantilů uspořádáme do tabulky.

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: $x_{0,25} - 1,5q = 2 - 1,5.3 = -2,5$

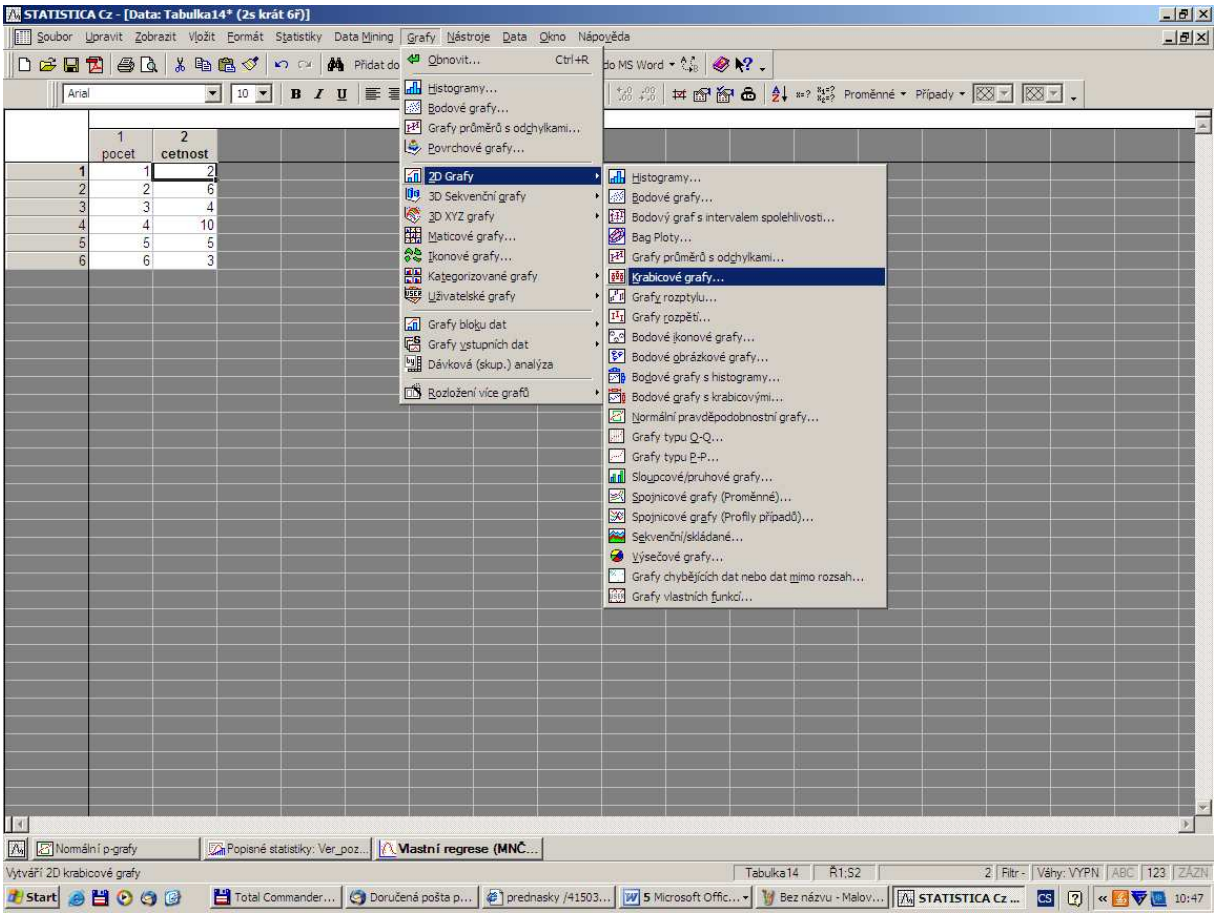
Horní vnitřní hradba: $x_{0,75} + 1,5q = 5 + 1,5.3 = 9,5$



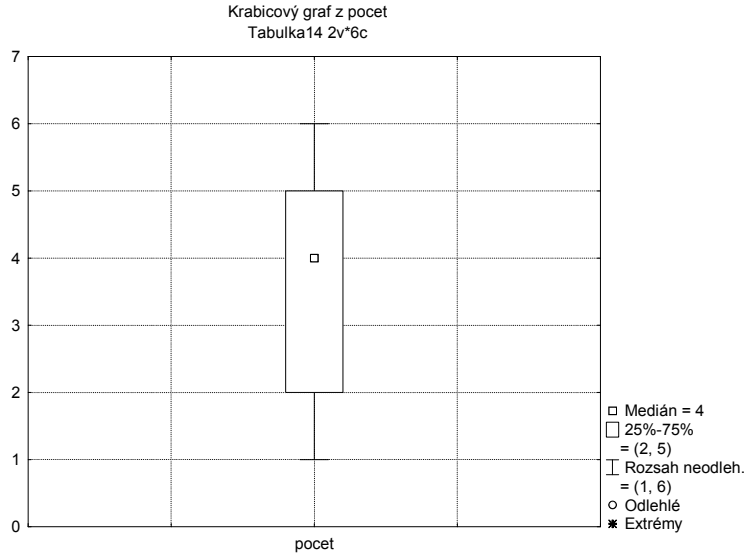
Vidíme, že datový soubor vykazuje určitou nesymetrii – medián je posunut směrem k hornímu kvartilu, soubor je tedy záporně sešikmen. V souboru se nevyskytují žádné odlehlé ani extrémní hodnoty.

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.

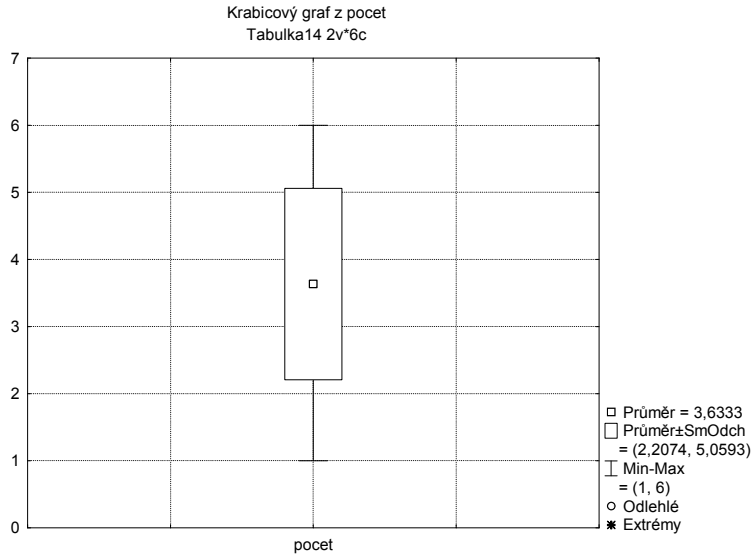


Zapneme proměnnou vah cetnost, zadáme závisle proměnnou pocet a dostaneme krabicový diagram:



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 $x_1, \ldots, x_n$ je posloupnost reálných čísel.
a) Jsou-li čísla navzájem různá, pak pořadím R_i čísla x_i rozumíme počet těch čísel $x_1, \ldots, x_n$, 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 $x_{(1)} \le ... \le x_{(n)}$ a na svislou osu kvantily u_{α_j}, kde $\alpha_j = \frac{3j-1}{3n+1}$, 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 $\left(x_{(j)}, u_{\alpha_j}\right)$ budou ležet na přímce.

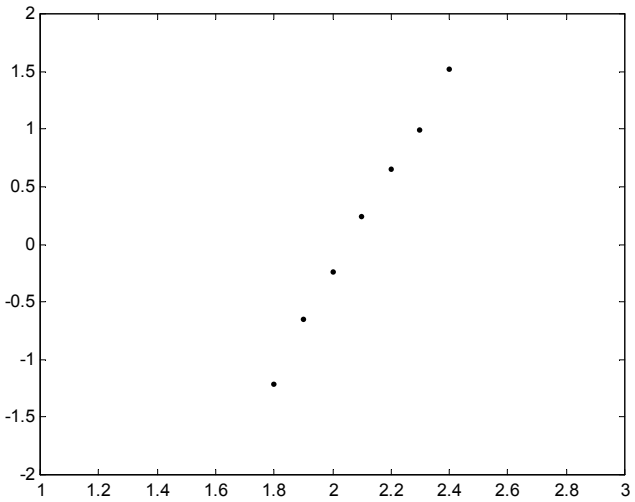
Pro data z rozložení s kladnou šikmostí se dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou řadit do konkávní křivky, zatímco pro data z rozložení se zápornou šikmostí se dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou řadit do konvexní křivky.

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 $\alpha_j = \frac{3j-1}{3n+1} = (0{,}1129; 0{,}2581; 0{,}4032; 0{,}5968; 0{,}7419; 0{,}8387; 0{,}9355)$,
vektor kvantilů $u_{\alpha_j} = (-1{,}2112; -0{,}6493; -0{,}245; 0{,}245; 0{,}6493; 0{,}9892; 1{,}5179)$.
Normální pravděpodobnostní graf

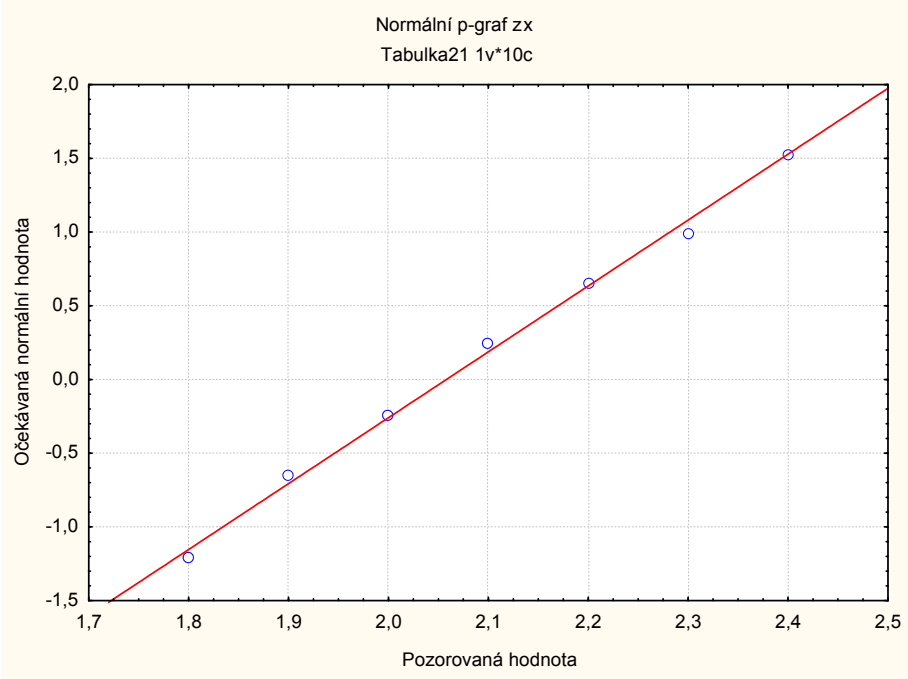


Protože dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ téměř leží na přímce, lze usoudit, že data pocházejí z normálního rozložení.

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 $x_{(1)} \le \ldots \le x_{(n)}$ 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 $x_{(1)} \le \ldots \le x_{(n)}$ 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 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.

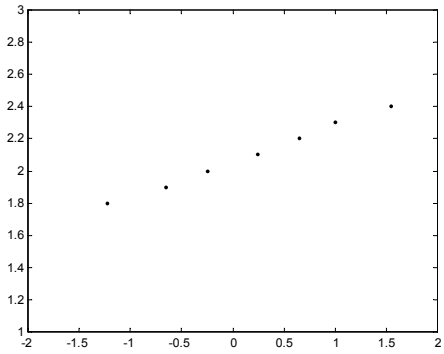
Řešení:

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

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

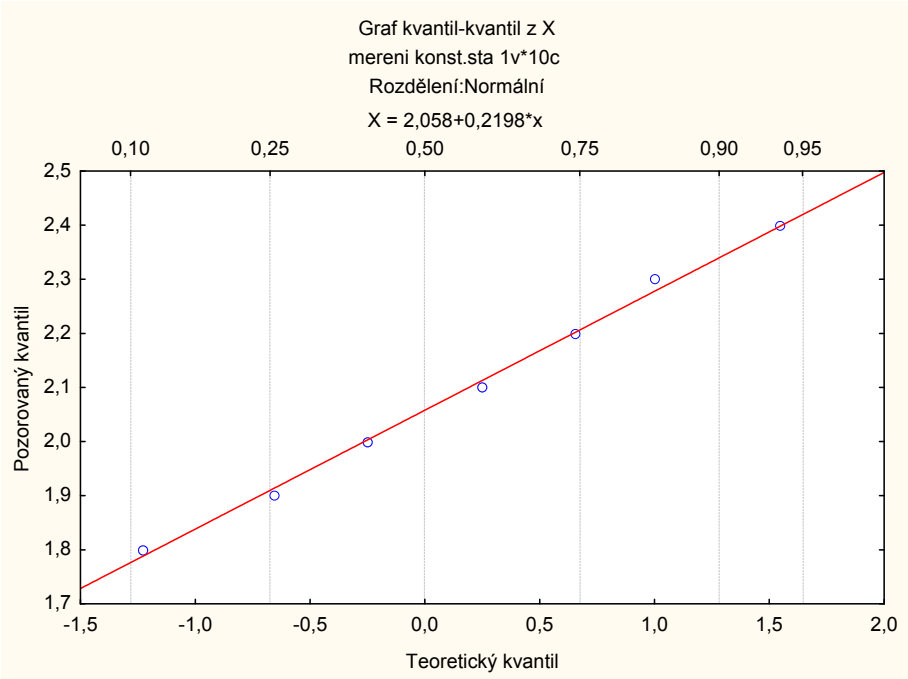


Vzhled grafu nasvědčuje tomu, že data pocházejí z normálního rozložení.

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.

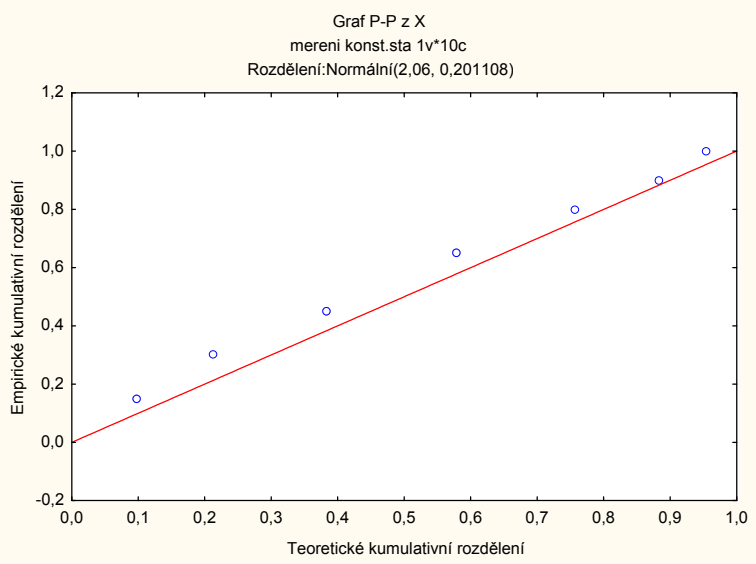
Způsob konstrukce: spočtou se standardizované hodnoty $z_{(j)} = \frac{x_{(j)} - m}{s}$, j = 1, ..., n. Na vodorovnou osu se vynesou hodnoty teoretické distribuční funkce $\Phi(z_{(j)})$ a na svislou osu hodnoty empirické distribuční funkce $F(z_{(j)}) = j/n$. (Jsou-li některé hodnoty $x_{(1)} \leq ... \leq x_{(n)}$ 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.

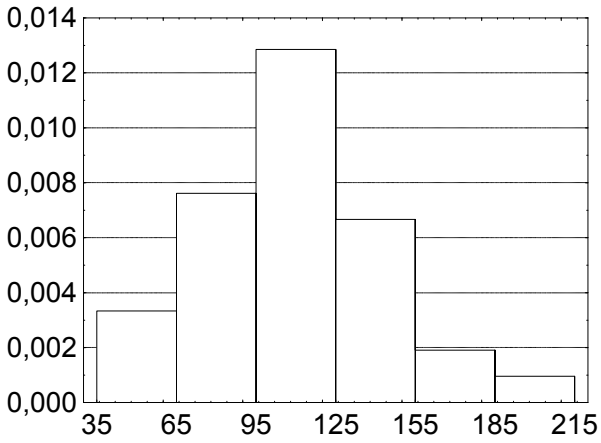
Příklad na konstrukci histogramu:
U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

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

Nakreslete histogram.
Řešení:
Nejprve sestavíme tabulku rozložení četností:

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

S pomocí této tabulky sestrojíme histogram:

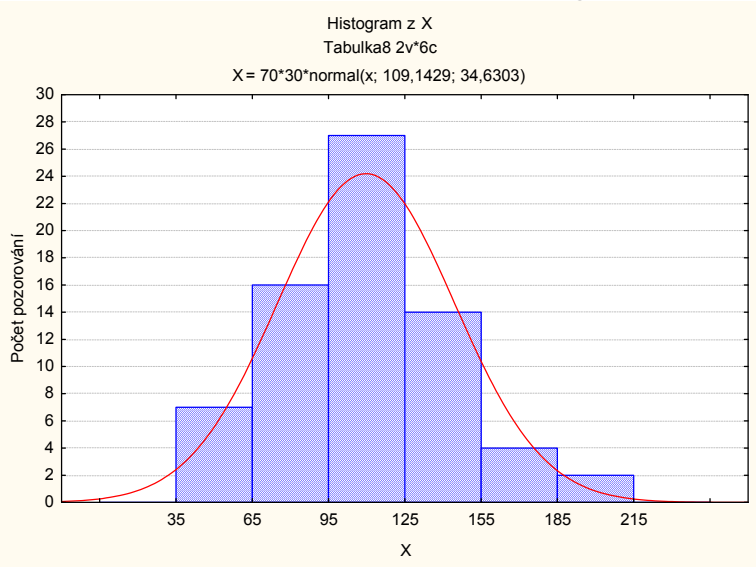


Výpočet pomocí systému STATISTICA:

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 cetnost
1	50	7
2	80	16
3	110	27
4	140	14
5	170	4
6	200	2

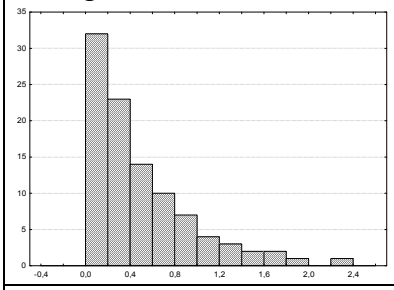
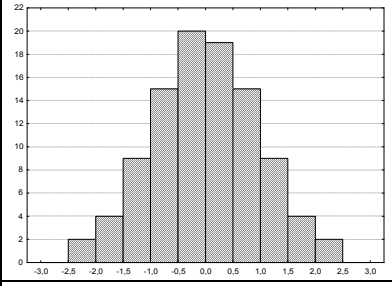
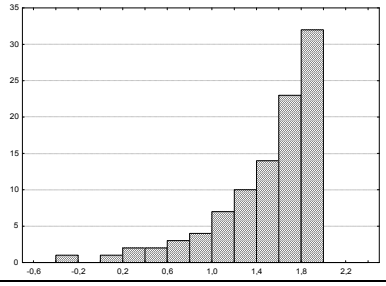
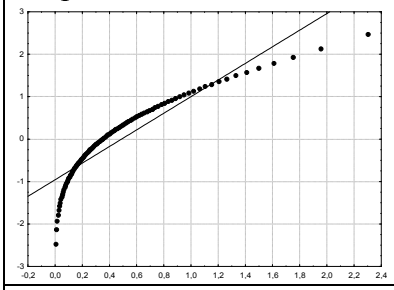
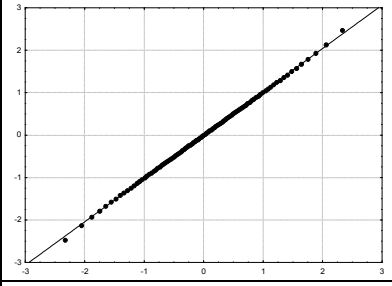
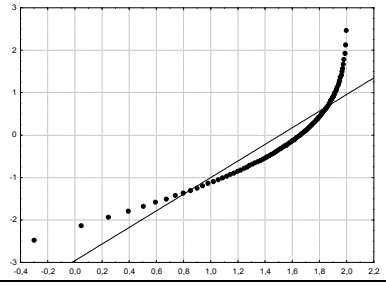
Grafy – Histogramy – zadáme proměnnou vah cetnost – Proměnná X - zaškrtneme Hranice – Určit hranice – zaškrtneme Zadejte hraniční rozmezí: Minimum 35, Krok 30, Maximum 215 – OK – OK. Dostaneme graf:



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