Téma 2.: Průzmumová analýza jednorozměrných dat

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.

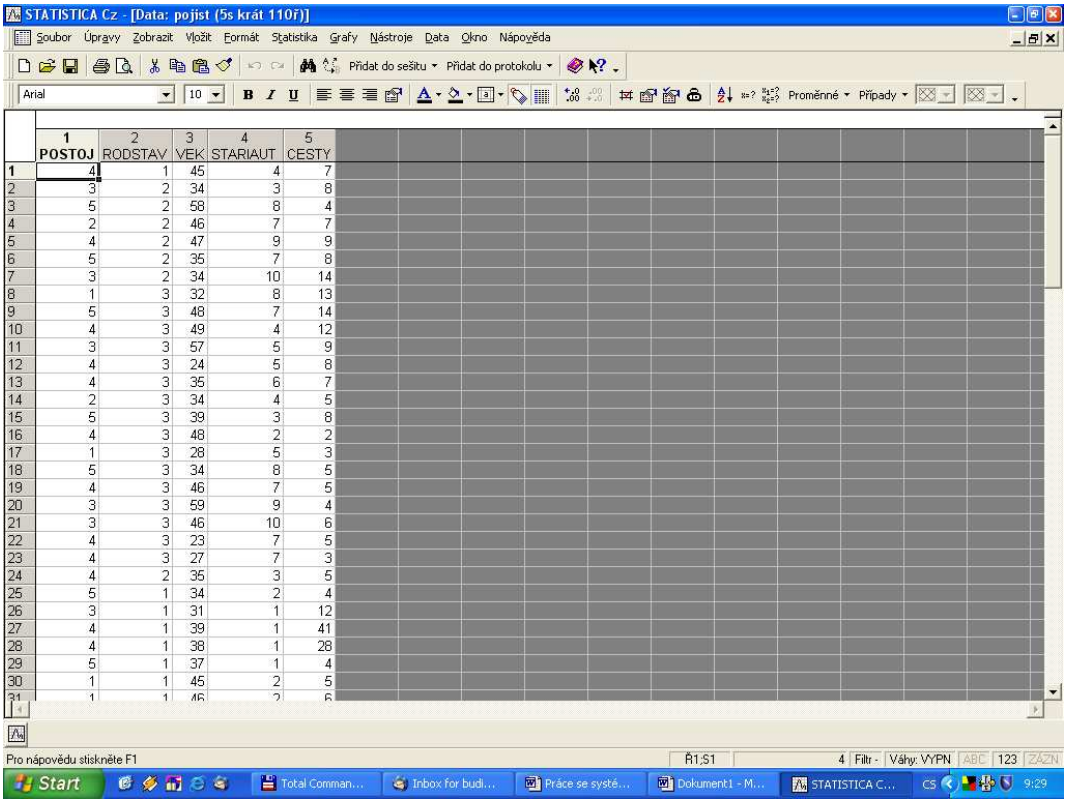
Náhodně bylo vybráno 110 současných zákazníků pojišťovny a ti byli telefonicky seznámeni s následujícím textem:

„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 pojist.sta a zakódovány takto:
POSTOJ ... postoj k novému typu pojištění (jednoznačný nezájem = 1, lehký nezájem = 2, neutrální postoj = 3, lehký zájem = 4, jednoznačný zájem = 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.

Ukázka části datového souboru:



	1 POSTOJ	2 RODSTAV	3 VEK	4 STARIAUT	5 CESTY
1	4	1	45	4	7
2	3	2	34	3	8
3	5	2	58	8	4
4	2	2	46	7	7
5	4	2	47	9	9
6	5	2	35	7	8
7	3	2	34	10	14
8	1	3	32	8	13
9	5	3	48	7	14
10	4	3	49	4	12
11	3	3	57	5	9
12	4	3	24	5	8
13	4	3	35	6	7
14	2	3	34	4	5
15	5	3	39	3	8
16	4	3	48	2	2
17	1	3	28	5	3
18	5	3	34	8	5
19	4	3	46	7	5
20	3	3	59	9	4
21	3	3	46	10	6
22	4	3	23	7	5
23	4	3	27	7	3
24	4	2	35	3	5
25	5	1	34	2	4
26	3	1	31	1	12
27	4	1	39	1	41
28	4	1	38	1	28
29	5	1	37	1	4
30	1	1	45	2	5
31	1	1	46	2	6

Úkol 1.: Datový soubor pojist.sta načtěte do systému STATISTICA. Všem proměnným vytvořte návěští a popište význam jednotlivých variant proměnných POSTOJ a RODSTAV.

Návod: Soubor – Otevřít – pojist.sta – Otevřít.
Názvy a vlastnosti proměnných se upravují v okně, do něhož vstoupíme, když 2x klikneme myší na název proměnné. Návěští se píše do Dlouhého jména, význam variant do Text. hodnot.

Úkol 2. Zjistěte absolutní a relativní četnosti a absolutní a relativní kumulativní četnosti proměnných POSTOJ a RODSTAV.
Návod: Statistiky – Základní statistiky/Tabulky – Tabulky četností – OK – Proměnné POSTOJ, RODSTAV – OK – Výpočet.
Tabulky se uloží do pracovního sešitu, listovat v nich můžeme pomocí stromové struktury v levé části okna.

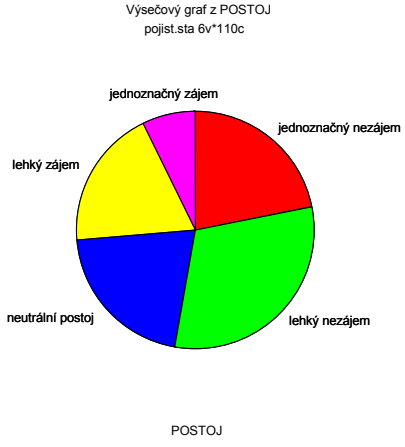
Tabulka četností pro POSTOJ

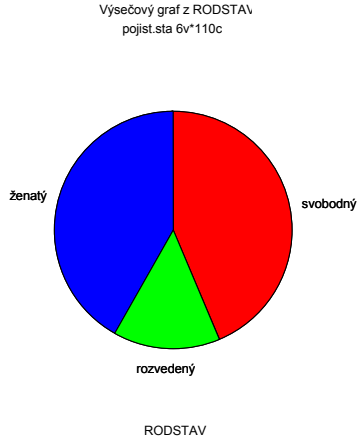
	Tabulka četností:POSTOJ: postoj k novému typu pojišt (pojist.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
jednoznačný nezájem	24	24	21,81818	21,8182
lehký nezájem	34	58	30,90909	52,7273
neutrální postoj	23	81	20,90909	73,6364
lehký zájem	21	102	19,09091	92,7273
jednoznačný zájem	8	110	7,27273	100,0000
ChD	0	110	0,00000	100,0000

Tabulka četností pro RODSTAV

	Tabulka četností:RODSTAV: rodinný stav zákazníka (pojist.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
svobodný	48	48	43,63636	43,6364
rozvedený	16	64	14,54545	58,1818
ženatý	46	110	41,81818	100,0000
ChD	0	110	0,00000	100,0000

Úkol 3. Absolutní četnosti proměnných POSTOJ a RODSTAV znázorněte graficky pomocí výsečového diagramu.
Návod: V menu zvolíme Grafy – 2D Grafy – Výsečové grafy.
Vybereme proměnné POSTOJ, RODSTAV a dostaneme následující grafy:



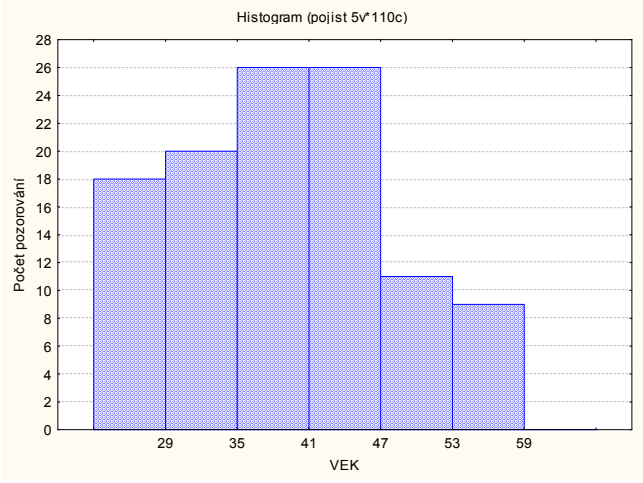


Z prvního diagramu je zřejmé, že nejméně zákazníků projevilo jednoznačný zájem o nový typ pojištění. Ostatní varianty jsou zastoupeny vcelku rovnoměrně.
Co se týká rodinného stavu zákazníků, vidíme, že v daném souboru jsou s přibližně stejnou četností zastoupeni ženatí a svobodní zákazníci. Rozvedených či ovdovělých je nejméně.

Úkol 4.: 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 proměnnou VEK vložíme novou proměnnou RVEK (Proměnné – Přidat – Za VEK, Jméno RVEK, Dlouhé jméno zakódovaný věk, OK). Nastavíme se kurzorem na RVEK. Data – Překódovat – Kategorie 1 Zahrnout pokud VEK >=23 and VEK <=29, do okénka Nová hodnota1 zapíšeme 1 atd. až Kategorie 6 Zahrnout pokud VEK > 53 and VEK <=59, do okénka Nová hodnota 6 zapíšeme 6, OK. Četnosti zjistíme analogicky jako v úkolu 2.

	Tabulka četností:RVEK: zakódovaný věk (pojist)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
<23,29>	18	18	16,36364	16,3636
(29,35>	20	38	18,18182	34,5455
(35,41>	26	64	23,63636	58,1818
(41,47>	26	90	23,63636	81,8182
(47,53>	11	101	10,00000	91,8182
(53,59>	9	110	8,18182	100,0000
ChD	0	110	0,00000	100,0000

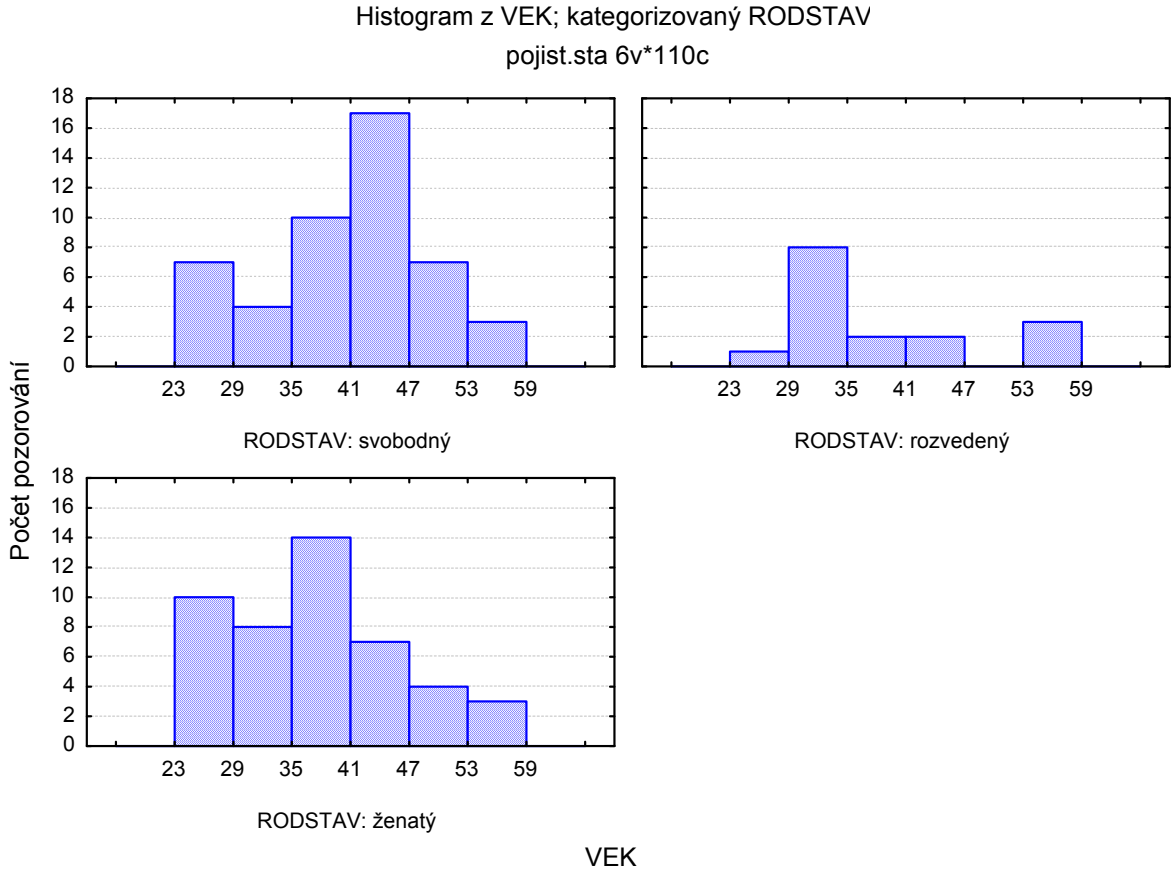
Úkol 5. Vytvořte histogram proměnné VEK se šesti třídicími intervaly <23,29>, (29,35>, (35,41>, (41,47>, (47,53>, (53,59>.
Návod: V menu vybereme Grafy – Histogramy – Proměnné VEK, OK, Detaily – zaškrtneme Hranice – Určit hranice – zaškrtneme Zadejte hraniční rozmezí, Minimum 23, Krok 6, Maximum 59 – OK – Vypneme normální proložení – OK. Dostaneme histogram v tomto tvaru:



Ze vzhledu histogramu lze soudit, že v souboru zákazníku jsou nejvíce zastoupeni lidé od 35 do 47 let. Soubor vykazuje kladné zešikmení, protože mladší věkové kategorie jsou zastoupeny s vyšší četností než starší věkové kategorie.

Úkol 6.: Vytvořte kategorizovaný histogram proměnné VEK podle proměnné RODSTAV.

Návod: Postupujeme stejně jako v předešlém případě a zvolíme Kategorizovaný – Kategorie X – Zapnuto – Změnit proměnnou RODSTAV – OK - OK.



Úkol 7.: Vypočtěte následující číselné charakteristiky: POSTOJ (ordinální proměnná) – modus, medián, dolní a horní kvartil, kvartilová odchylka. RODSTAV (nominální proměnná) – modus. VEK, STARIAUT, CESTY (poměrové proměnné) – průměr, směrodatná odchylka, koeficient variace, šikmost, špičatost.
Návod: Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK, Proměnné – zadáme název příslušné proměnné, Detailní výsledky – vybereme příslušné charakteristiky.

	Popisné statistiky (pojist.sta)					
Proměnná	Medián	Modus	Četnost modu	Spodní kvartil	Horní kvartil	Kvartilové rozpětí
POSTOJ	2	2	34	2	4	2

Vidíme, že medián, modus a dolní kvartil jsou stejné – je to varianta 2 „lehký nezájem“. Horním kvartilem je varianta 4 „lehký zájem“.

	Popisné statistiky (pojist.sta)	
Proměnná	Modus	Četnost modu
RODSTAV	1	48

V našem datovém souboru je nejčetnější variantou rodinného stavu varianta 1 „svobodný“.

Proměnná	Popisné statistiky (pojist.sta)				
	Průměr	Sm.odch.	Koef.prom.	Šikmost	Špičatost
VEK	39,58182	8,823844	22,29267	0,191625	-0,59532
STARIAUT	4,16364	2,359938	56,67974	0,905405	0,35924
CESTY	7,16364	5,304537	74,04811	3,150711	15,99807

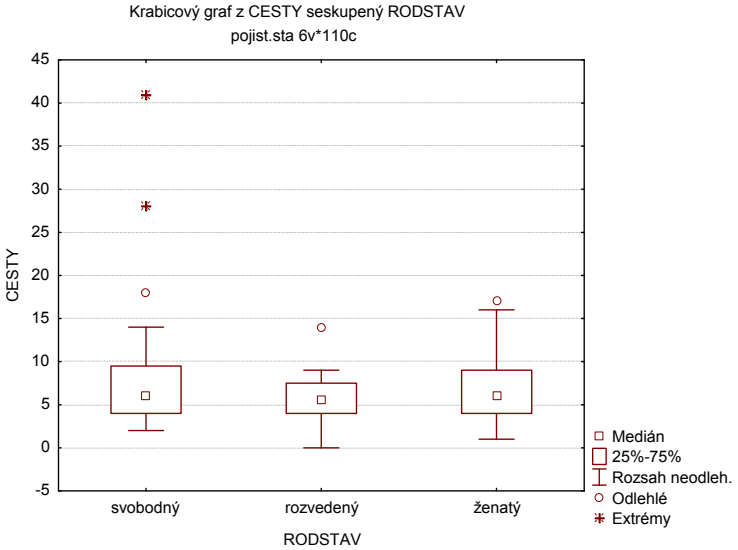
Průměrný věk zákazníka je 39 let a 7 měsíců se směrodatnou odchylkou 8 let a 10 měsíců. Rozložení věku vykazuje kladnou šikmost (podprůměrné hodnoty věku jsou četnější než nadprůměrné) a zápornou špičatost (rozložení věku je plošší než normální rozložení).
Průměrné stáří auta je 4 roky a 2 měsíce se směrodatnou odchylkou 2 roky a 4 měsíce. Rozložení stáří aut je kladně zešikmené a špičatější než normální rozložení.
Průměrný počet cest nad 300 km je 7,2 se směrodatnou odchylkou 5,3. Rozložení počtu cest na 300 km je značně kladně zešikmené a podstatně špičatější než normální rozložení.
Z porovnání variability uvedených tří proměnných pomocí koeficientů variace (koeficient variace je podíl směrodatné odchylky a průměru, často se udává v procentech) vyplývá, že nejvyšší variabilitu má proměnná CESTY, nejnižší VEK.

Úkol 8.: Zjistěte, jaký je průměrný počet cest nad 300 km pro svobodné, rozvedené , ženaté zákazníky pojišťovny. Výpočet doplňte krabicovým diagramem.
Návod: Statistiky – Základní statistiky/tabulky – Rozklad&jednofakt. ANOVA – OK – Proměnné – Závisle proměnné CESTY, Grupovací proměnná RODSTAV – OK – OK – Popisné statistiky – ponecháme jen N platných – Výpočet

Rozkladová tabulka popisných statistik (pojist.sta) N=110 (V seznamu záv. prom. nejsou ChD)		
RODSTAV	CESTY průměr	CESTY N
svobodný	7,895833	48
rozvedený	5,750000	16
ženatý	6,891304	46
Vš.skup.	7,163636	110

Vidíme, že nejvyšší průměrný počet cest nad 300 km mají svobodní zákazníci pojišťovny.

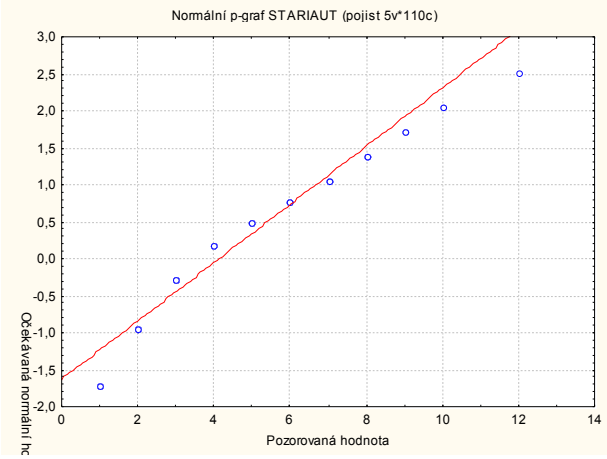
Vytvoření krabicového grafu: Grafy – 2D Grafy – Krabicové grafy – Proměnné – Závisle proměnné CESTY, Grupovací proměnná RODSTAV – OK – OK



Ve všech třech variantách rodinného stavu se vyskytují odlehlé hodnoty, u svobodných zákazníků pojišťovny jsou dokonce i extrémní hodnoty.

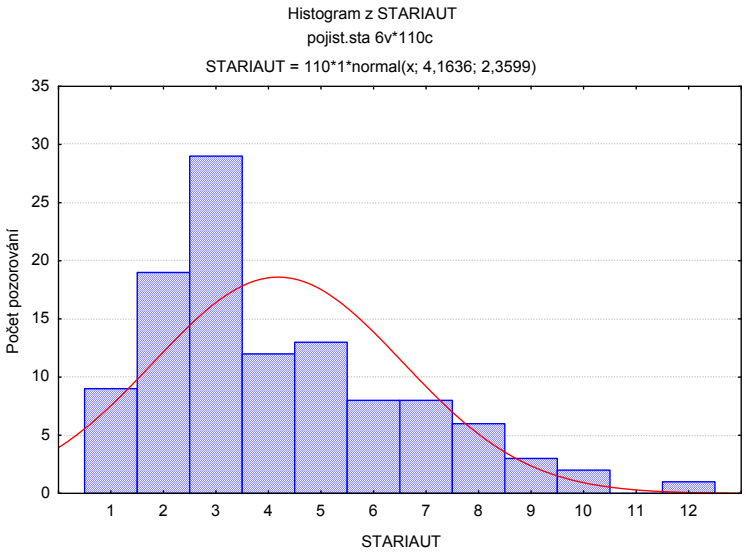
Úkol 9.: Pro proměnnou STARIAUT sestrojte N-P graf a s jeho pomocí posuďte normalitu této proměnné.

Návod: Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné STARIAUT – OK.



Tečky v NP grafu se značně odchylují od zakreslené přímky a řadí se do konkávního tvaru. Datový soubor vykazuje kladné zešikmení, nejedná se tedy o normální rozložení.

Úkol 10.: Pro proměnnou STARIAUT nakreslete histogram s proloženou hustotou normálního rozložení. Ponechejte implicitní počet třídicích intervalů.
Návod: Grafy – Histogramy – Proměnné STARIAUT – OK.

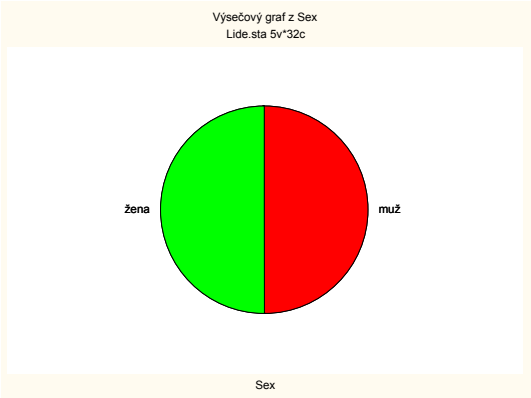


Tvar histogramu svědčí o kladně zešikmeném rozložení, jehož hustota neodpovídá hustotě normálního rozložení.

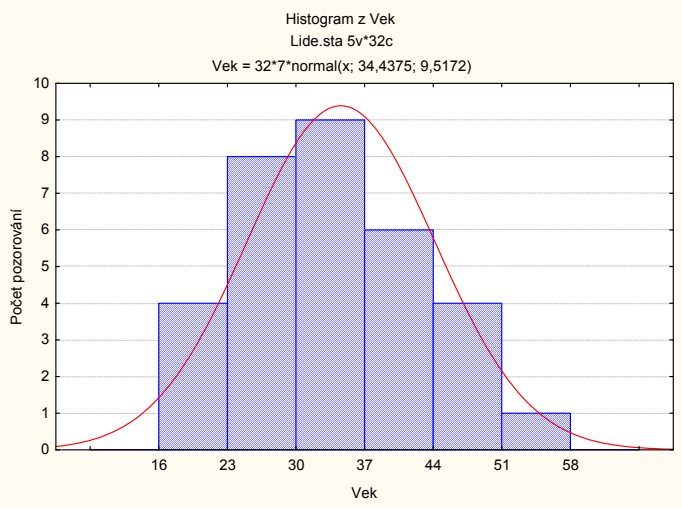
Příklad k samostatnému řešení:
Načtěte datový soubor lide.sta, s nímž jste pracovali v 1. cvičení.

1. Vytvořte tabulku absolutních a relativních četností proměnné SEX. Četnosti znázorněte pomocí výsečového diagramu.

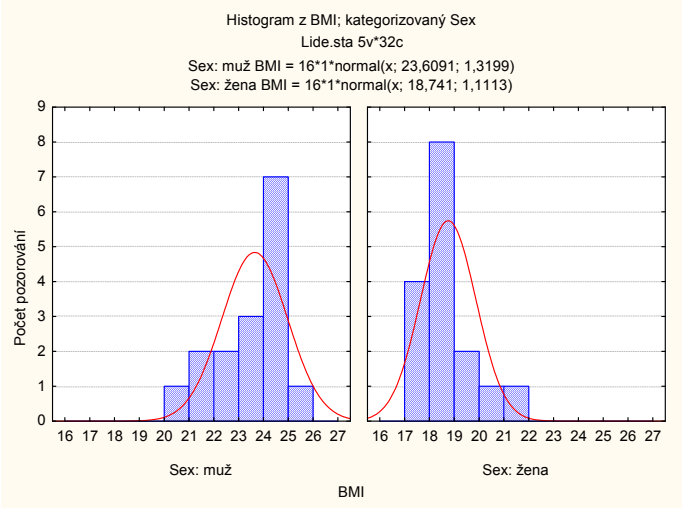
	Tabulka četností:Sex (Lide.sta)	
Kategorie	Četnost	Rel.četnost
muž	16	50
žena	16	50



2. Vytvořte histogram proměnné VEK se šesti třídicími intervaly (16,23>, (23,30>, (30,37>, (37,43>, (43,50>, (50,57> a zakreslenou Gaussovou křivkou.



3. Vytvořte kategorizované histogramy proměnné BMI pro muže a pro ženy.



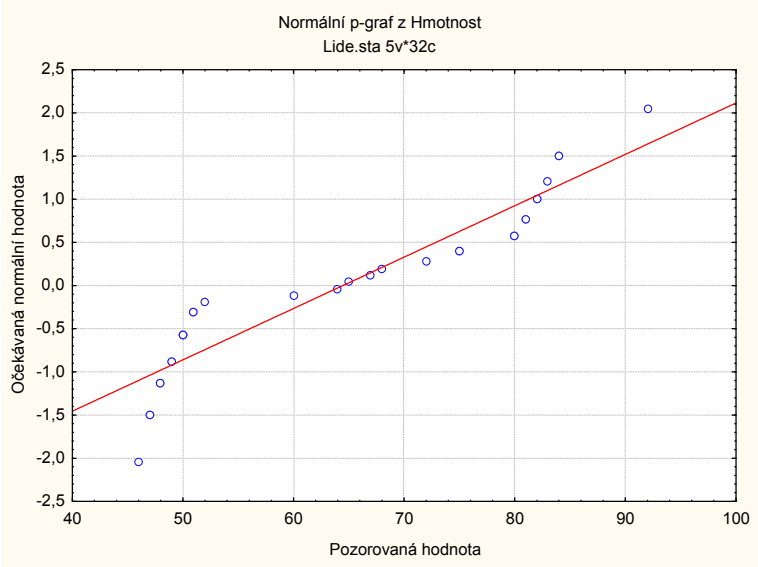
4. Vypočtěte průměr, směrodatnou odchylku, koeficient variace, šikmost a špičatost proměnné BMI pro muže a pro ženy. Výsledky udávejte na dvě destinná místa.
Pro muže:

	Popisné statistiky (Lide.sta) Zhrnout podmínku: Sex=1					
Proměnná	N platných	Průměr	Sm.odch.	Koef.prom.	Šikmost	Špičatost
BMI	16	23,61	1,32	5,59	-0,78	-0,25

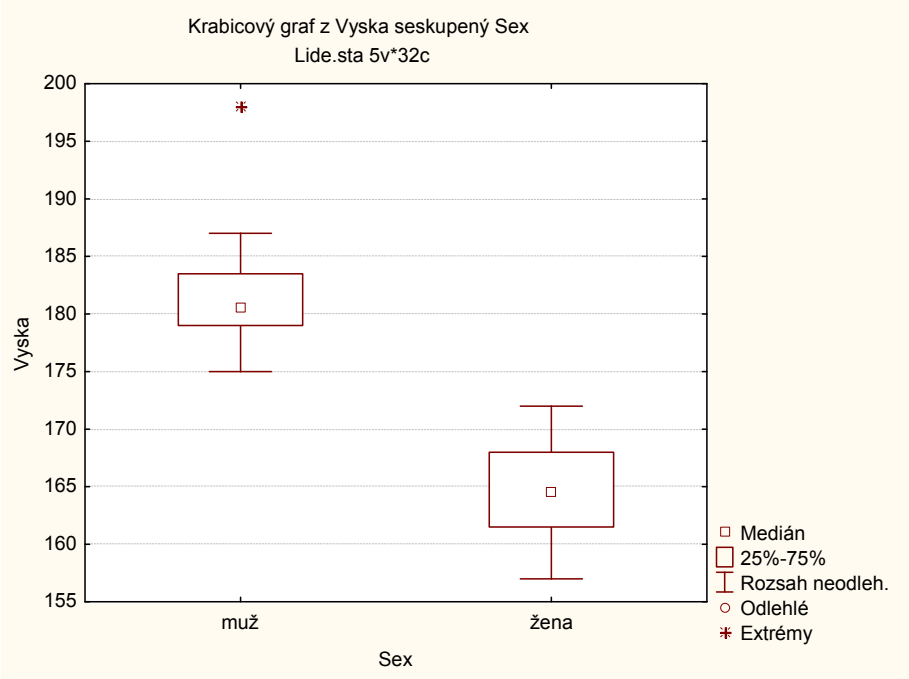
Pro ženy

	Popisné statistiky (Lide.sta) Zhrnout podmínku: Sex=2					
Proměnná	N platných	Průměr	Sm.odch.	Koef.prom.	Šikmost	Špičatost
BMI	16	18,74	1,11	5,93	1,39	2,65

5. Sestrojte N-P plot pro proměnnou Hmotnost.



6. Vytvořte kategorizované krabicové diagramy pro proměnnou Vyska pro muže a pro ženy.



7. K extrémní hodnotě výšky umístěte jméno muže, kterému tato výška přísluší.
(Jan)

8. Sestrojte bag plot pro proměnné Vyska, Hmotnost. K odlehlým hodnotám umístěte jména osob, kterým tato pozorování příslušejí.