

2 EXPLORATORNÍ ANALÝZA

2.1. Níže uvedená data představují částečný výsledek zaznamenaný při průzkumu zatížení jedné z ostravských křižovatek, a to barvu projíždějících automobilů. Data vyhodnot'te a graficky znázorn'ete.

červená	modrá	červená	zelená
modrá	červená	červená	bílá
zelená	zelená	modrá	červená

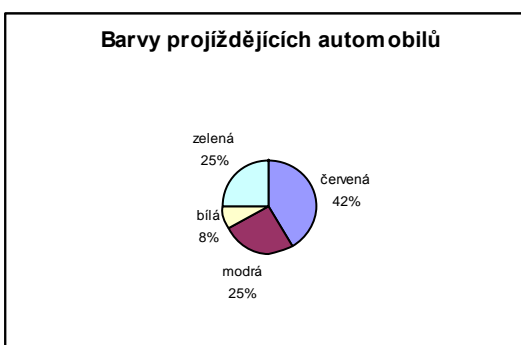
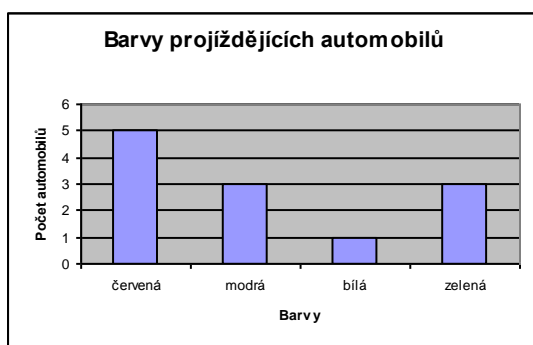
Řešení:

Je zřejmé, že se jedná o kvalitativní (slovní) proměnnou a vzhledem k tomu, že barvy automobilů nemá smysl seřazovat ani porovnávat, můžeme konstatovat, že se jedná o proměnnou nominální.

Pro její popis tedy zvolíme tabulku četností, určíme modus a barvu projíždějících automobilů znázorníme prostřednictvím histogramu a výsečového grafu.

TABULKA ROZDĚLENÍ ČETNOSTI		
Barvy projíždějících automobilů	Absolutní četnost	Relativní četnost
	n_i	p_i
červená	5	$5/12 = 0,42$
modrá	3	$3/12 = 0,25$
bílá	1	$1/12 = 0,08$
zelená	3	$3/12 = 0,25$
Celkem	12	1,00

Modus = červená (tj. v zaznamenaném vzorku se vyskytlo nejvíce červených automobilů)

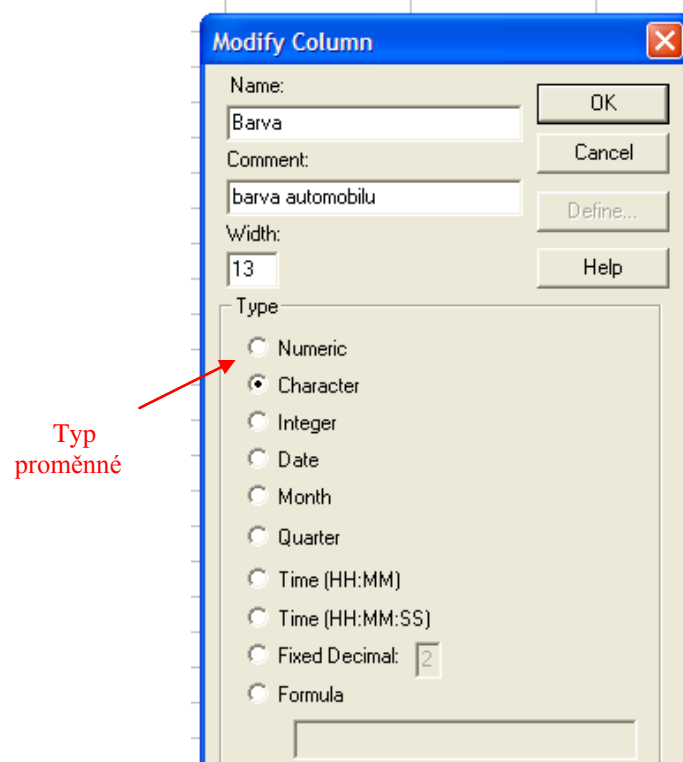


Celkem bylo sledováno 12 automobilů

Řešení daného problému ve Statgraphicsu:

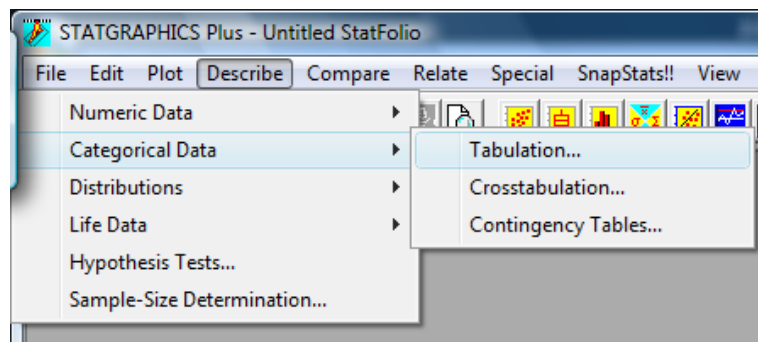
Zadání proměnné:

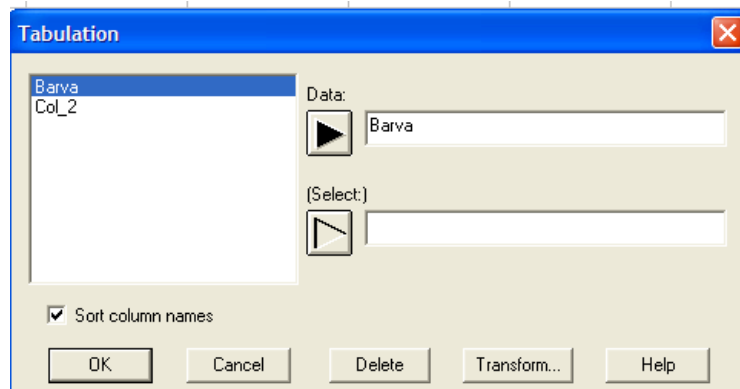
Chceme-li zadávat ručně novou proměnnou, provedeme DC (dvojklik) na hlavičku sloupce a zadáme parametry proměnné (název, popis (nepovinné), šířku a typ). Přednastavený typ je Numeric, proto je nutno nastavení typu proměnné ohlídat zejména při zadávání proměnné kategoriální.



Exploratorní analýza pro kategoriální proměnnou:

Touto analýzou získáme tabulku četnosti, histogram a výšečový graf.





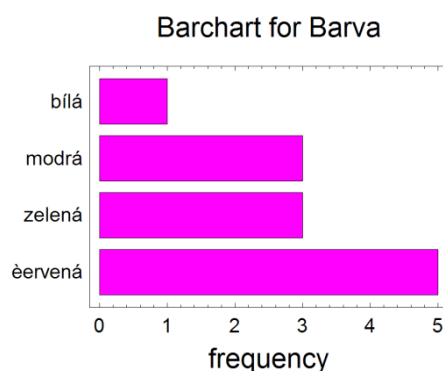
Datový výstup analýzy:

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	bílá	1	0,0833	1	0,0833
2	modrá	3	0,2500	4	0,3333
3	zelená	3	0,2500	7	0,5833
4	červená	5	0,4167	12	1,0000

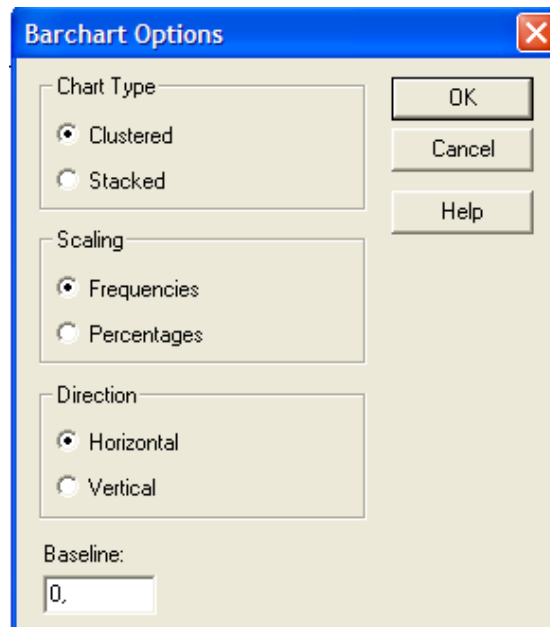
↑
názvy kategorií
↑
četnost
↑
relativní četnost
↑
kumulativní četnost
↑
kumulativní relativní četnost

Všimněte si, že **Statgraphics** automaticky určuje kumulativní četnosti a kumulativní relativní četnosti i pro nominální proměnnou (je tedy na uživateli, aby určil, zda mají tyto charakteristiky v konkrétním případě smysl).

Histogram:

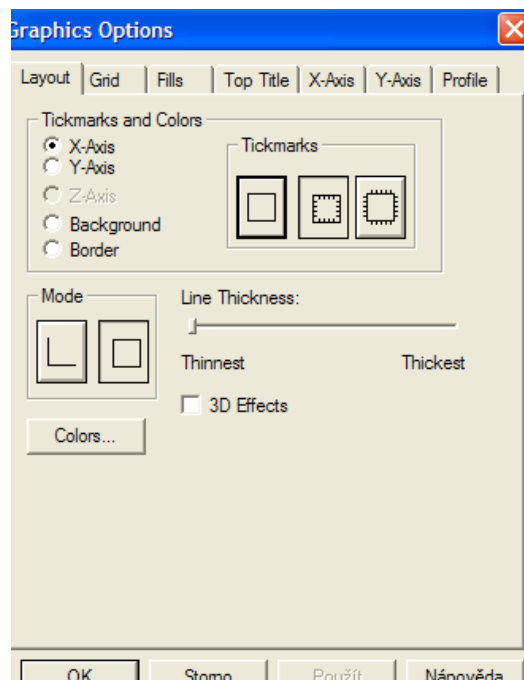


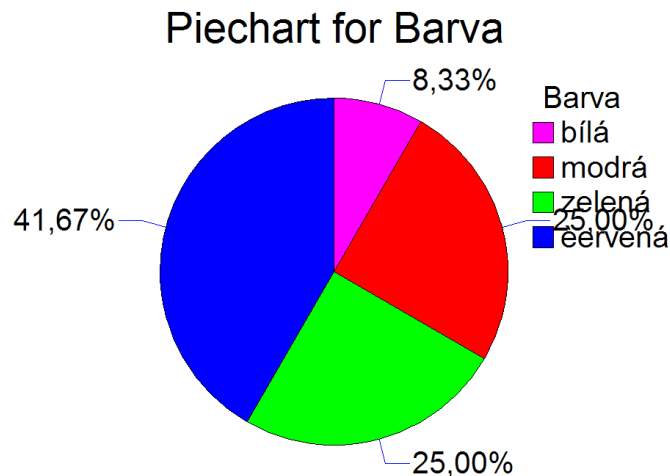
Formát grafu změním tak, že provedeme RC (klikneme pravým tlačítkem myši) na oblast grafu a zvolíme **Pane Option**.



V okně **Barchart Option** pak volíme formátování histogramu.

Grafické parametry histogramu (nadpisy, barvy...) nastavíme v okně Graphics Option, které získáme po RC na oblast grafu a volbě **Graphics Option**.



Výšečový graf:

Při úpravě výšečového grafu postupujeme obdobně jako při úpravě histogramu. (**Pane Option, Graphics option**).

2.2. Následující data představují velikosti triček prodaných při výprodeji firmy TRIKO.

S, M, L, S, M, L, XL, XL, M, XL, XL, L, M, S, M, L, L, XL, XL, XL, L, M

a) Data vyhodnoťte a graficky znázorněte.

b) Určete kolik procent lidí si koupilo tričko velikosti nejvýše L.

Řešení:

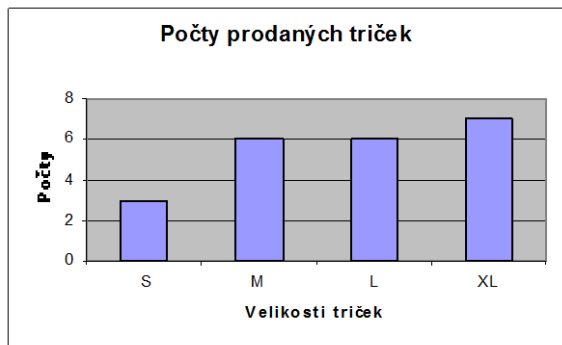
ada) Zřejmě se jedná o kvalitativní (slovní) proměnnou a vzhledem k tomu, že velikosti triček lze seřadit, jde o proměnnou ordinální. Pro její popis proto použijeme tabulku četností pro ordinální proměnnou, v níž varianty velikosti triček budou seřazeny od nejmenší po největší (S, M, L, XL) a modus.

TABULKA ROZDĚLENÍ ČETNOSTI				
Velikosti triček	Absolutní četnost	Kumulativní četnost	Relativní četnost	Relativní kum.četnost
	n_i	m_i	p_i	F_i
S	3	3	$3/22 = 0,14$	$3/22 = 0,14$
M	6	$3 + 6 = 9$	$6/22 = 0,27$	$9/22 = 0,41$
L	6	$9 + 6 = 15$	$6/22 = 0,27$	$15/22 = 0,68$
XL	7	$15 + 7 = 22$	$7/22 = 0,32$	$22/22 = 1,00$
Celkem	22	----	1,00	----

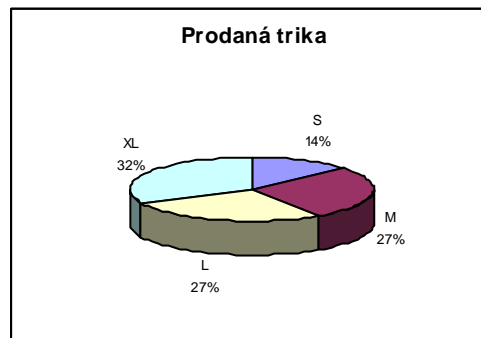
Modus = XL (nejvíce lidí si koupilo tričko velikosti XL)

Grafický výstup bude tvořit histogram, výšečový graf a polygon kumulativních četností (jelikož se nejedná o technická data, Paretův graf vytvářet nebudeme).

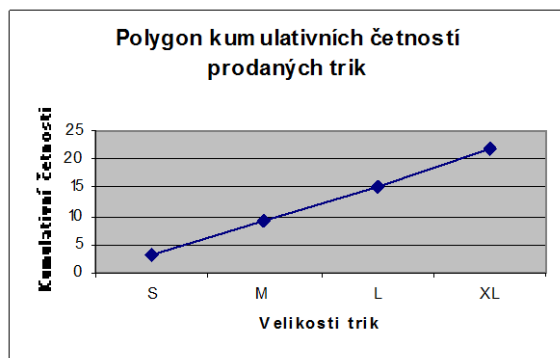
Grafický výstup:



Histogram



Celkem bylo prodáno 22 triček



Galtonova ogiva, S-křivka

adb) Na tuto otázku nám dá odpověď relativní kumulativní četnost pro variantu L, která určuje jaká část prodaných triček byla velikosti L a nižších. Tj. 68% zákazníků si koupilo tričko velikosti L a menší.

2.3. Následující data představují věk hudebníků vystupujících na přehlídce dechových orchestrů. Proměnnou věk považujte za spojitou. Určete průměr, shorth a modus věku hudebníků.

22 82 27 43 19 47 41 34 34 42 35

Řešení:

a) Určení průměru:

V tomto případě jednoznačně použijeme aritmetický průměr (zdůvodnění snad není nutné):

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{22 + 82 + 27 + 43 + 19 + 47 + 41 + 34 + 34 + 42 + 35}{11} = 38,7 \text{ let}$$

Průměrný věk hudebníka vystupujícího na přehlídce dechových orchestrů je 38,7 let.

Prohlédněte si ještě jednou zadaná data a promyslete si nakolik je průměrný věk reprezentativní statistikou daného výběru (odlehlá pozorování).

b) Určení shorthu:

Náš výběrový soubor má 11 hodnot, z čehož vyplývá, že v shorthu bude ležet 6 z nich (rozsah souboru je 11 (lichý počet hodnot), 50% z toho je 5,5 (5,5 hodnoty se špatně určuje, že?) a nejbližší vyšší přirozené číslo je 6 – neboli: $n/2 + 1/2 = 11/2 + 1/2 = 12/2 = 6$).

A další postup?

- Proměnnou seřadíme
- Určíme délky všech 6-ti členných intervalů, v nichž $x_i < x_{i+1} < \dots < x_{i+5}$
- Nejkratší z těchto intervalů prohlásíme za shorth (délka intervalu = $x_{i+5} - x_i$)

Originální data	Seřazená data	Délky 6-ti členných intervalů
22	19	16 (= 35 – 19)
82	22	19 (= 41 – 22)
27	27	15 (= 42 – 27)
43	34	9 (= 43 – 34)
19	34	13 (= 47 – 34)
47	35	47 (= 82 – 35)
41	41	
34	42	
34	43	
42	47	
35	82	

Z tabulky je zřejmé, že nejkratší interval má délku 9, čemuž odpovídá jediný interval: $\langle 34;43 \rangle$.

Shorth = $\langle 34;43 \rangle$, což můžeme interpretovat např. tak, že polovina hudebníků je ve věku 34 až 43 let (jde přitom o nejkratší interval ze všech možných).

c) Určení modu:

Modus je definován jako střed shorthu:

$$\hat{x} = \frac{34 + 43}{2} = 38,5$$

Modus = 38,5 let, tj. typický věk hudebníka vystupujícího na přehlídce dechových orchestrů je 38,5 let.

2.4. Pro data z předcházejícího příkladu určete:

- a) všechny kvartily,
- b) interkvartilové rozpětí
- c) MAD
- d) zakreslete empirickou distribuční funkci

Řešení:

ada) Naším úkolem je určit dolní kvartil $x_{0,25}$, medián $x_{0,5}$ a horní kvartil $x_{0,75}$. Budeme-li dodržovat postup doporučený pro určování kvantilů, znamená to – data seřadit a přiřadit jim pořadí. Splnění prvních dvou bodů postupu ukazuje následující tabulka:

Originální data	Seřazená data	Pořadí
22	19	1
82	22	2
27	27	3
43	34	4
19	34	5
47	35	6
41	41	7
34	42	8
34	43	9
42	47	10
35	82	11

A můžeme přejít k bodu 3, tj. stanovit pořadí hodnot proměnné pro jednotlivé kvartily a tím i jejich hodnoty:

Dolní kvartil $x_{0,25}$: $p = 0,25; n = 11 \Rightarrow z_p = 11 \cdot 0,25 + 0,5 = 3,25$,

Dolní kvartil je tedy průměrem prvků s pořadím 3 a 4 - $x_{0,25} = \frac{27+34}{2} = 30,5$ let.

Tj. 25% hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 30,5 let (75% z nich má 30,5 let a více).

Medián $x_{0,5}$: $p = 0,5; n = 11 \Rightarrow z_p = 11 \cdot 0,5 + 0,5 = 6 \Rightarrow x_{0,5} = 35$

Tj. polovina hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 35 let (50% z nich má 35 let a více).

Horní kvartil $x_{0,75}$: $p = 0,75; n = 11 \Rightarrow z_p = 11 \cdot 0,75 + 0,5 = 8,75$

Horní kvartil je tedy průměrem prvků s pořadím 8 a 9 - $x_{0,75} = \frac{42+43}{2} = 42,5$ let.

Tj. 75% hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 42,5 let (25% z nich má 42,5 let a více).

adb) **Interkvartilové rozpětí IQR:**

$$\mathbf{IQR} = x_{0,75} - x_{0,25} = 42,5 - 30,5 = 12$$

adc) **MAD**

Chceme-li určit tuto statistiku, budeme postupovat přesně podle toho co nám říká definice (medián absolutních odchylek od mediánu), tudíž dodržíme výše uvedený postup, jehož aplikaci vám ukazuje následující tabulka.

$$x_{0,5} = 35$$

Originální data x_i	Seřazená data y_i	Absolutní hodnoty odchylek seřazených dat od jejich mediánu $ y_i - x_{0,5} $	Seřazené absolutní hodnoty odchylek seřazených dat od jejich mediánu M_i
22	19	16 = $ 19 - 35 $	0
82	22	13 = $ 22 - 35 $	1
27	27	8 = $ 27 - 35 $	1
43	34	1 = $ 34 - 35 $	6
19	34	1 = $ 34 - 35 $	7
47	35	0 = $ 35 - 35 $	8
41	41	6 = $ 41 - 35 $	8
34	42	7 = $ 42 - 35 $	12
34	43	8 = $ 43 - 35 $	13
42	47	12 = $ 47 - 35 $	16
35	82	47 = $ 82 - 35 $	47

$$MAD = M_{0,5}$$

$$p = 0,5; n = 11 \Rightarrow z_p = 11 \cdot 0,5 + 0,5 = 6 \Rightarrow M_{0,5} = 8$$

(MAD je medián absolutních odchylek od mediánu, tj. 6. hodnota seřazeného souboru absolutních odchylek od mediánu). $MAD = 8$.

add) Zbývá nám poslední úkol – sestavit **empirickou distribuční funkci**. Připomeňme si proto její definici – a postupujme podle ní:

$$F(x) = \begin{cases} 0 & \text{pro } x \leq x_1 \\ \sum_{i=1}^j p(x_i) & \text{pro } x_j < x \leq x_{j+1}, 1 \leq j \leq n-1 \\ 1 & \text{pro } x_n < x \end{cases}$$

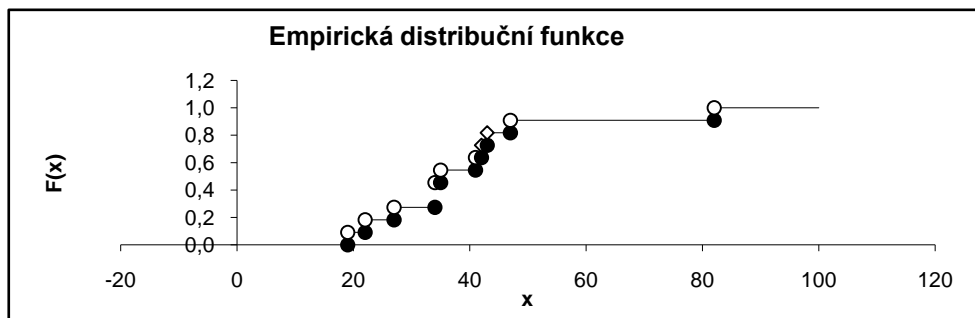
- do tabulky si zapíšeme seřazené hodnoty proměnné, jejich četnosti, relativní četnosti a z nich odvodíme empirickou distribuční funkci:

Originální data x_i	Seřazené hodnoty a_i	Absolutní četnosti seřazených hodnot n_i	Relativní četnosti seřazených hodnot p_i	Empirická dist. funkce $F(a_i)$
22	19	1	1/11	0
82	22	1	1/11	1/11
27	27	1	1/11	2/11
43	34	2	2/11	3/11
19	35	1	1/11	5/11
47	41	1	1/11	6/11
41	42	1	1/11	7/11
34	43	1	1/11	8/11
34	47	1	1/11	9/11
42	82	1	1/11	10/11
35				

Z definice emp. dist. funkce $F(x)$ tedy plyne, že pro všechna x menší než 19 je $F(x)$ rovna nule, pro x větší než 19 a menší nebo rovna 22 je $F(x)$ rovna 1/11, pro x větší než 22 a menší nebo rovna 27 je $F(x)$ rovna 1/11 + 1/11, atd.

x	$(-\infty; 19)$	$(19; 22)$	$(22; 27)$	$(27; 34)$	$(34; 35)$
$F(x)$	0	1/11	2/11	3/11	5/11

x	$(35; 41)$	$(41; 42)$	$(42; 43)$	$(43; 47)$	$(47; 82)$	$(82; \infty)$
$F(x)$	6/11	7/11	8/11	9/11	10/11	11/11



2.5. Firma vyrábějící tabulové sklo vyvinula méně nákladnou technologii pro zlepšení odolnosti skla vůči žáru. Pro testování bylo vybráno 5 tabulí skla a rozřezáno na polovinu. Jedna polovina pak byla ošetřena novou technologií, zatímco druhá byla ponechána jako kontrolní. Obě poloviny pak byly vystaveny zvyšujícímu se působení tepla, dokud nepraskly. Výsledky byly následující:

Mezní teplota (sklo prasklo) [°C]	
Stará technologie x_i	Nová technologie y_i
475	485
436	390
495	520
483	460
426	488

Porovnejte obě technologie pomocí základních charakteristik exploratorní statistiky (průměru a rozptylu, popř. směrodatné odchylky).

Řešení:

- Nejprve se pokusíme porovnat obě technologie pouze za pomoci průměru:

Průměr pro starou technologii:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{475 + 436 + \dots + 426}{5} = 463,0 \quad [^{\circ}C]$$

Průměr pro novou technologii:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{485 + 390 + \dots + 488}{5} = 468,6 \quad [^{\circ}C]$$

Na základě vypočtených průměrů bychom mohli říci, že novou technologii doporučujeme, poněvadž mezní teplota je při nové technologii téměř o 6^oC vyšší.

A co na to míry variability?

Stará technologie:

Výběrový rozptyl:

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{(475 - 463,0)^2 + (436 - 463,0)^2 + \dots + (426 - 463,0)^2}{5-1} = 916,3 \quad [^{\circ}C^2]$$

Výběrová směrodatná odchylka:

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{s_x^2} = \sqrt{916,3} = 30,3 \quad [^{\circ}C]$$

Nová technologie:

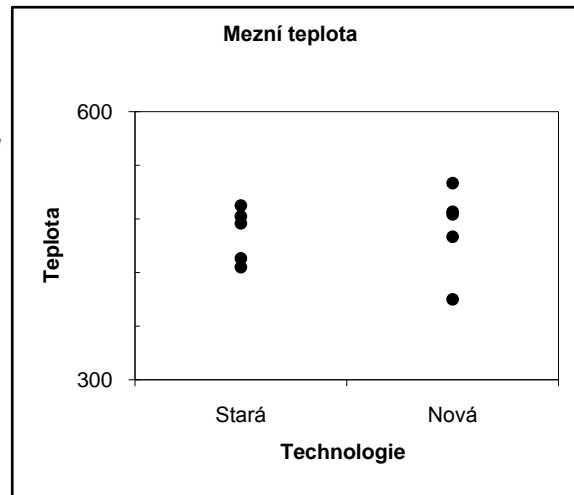
Výběrový rozptyl:

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{(485 - 468,6)^2 + (390 - 468,6)^2 + \dots + (488 - 468,6)^2}{5-1} = 2384,4 \quad [^{\circ}C^2]$$

Výběrová směrodatná odchylka:

$$s_x = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = \sqrt{s_y^2} = \sqrt{2384,4} = 48,8 \text{ } [^{\circ}C]$$

Tady pozor. Výběrový rozptyl (výběrová směrodatná odchylka) vyšel pro novou technologii mnohem vyšší než pro technologii starou. Co to znamená? Podívejte se na grafické znázornění naměřených dat.



Mezní teploty pro novou technologii jsou mnohem rozptýlenější, tzn. že tato technologie není ještě dobře zvládnutá a její použití nám nezaručí zkvalitnění výroby. V tomto případě může dojít k silnému zvýšení, ale také k silnému snížení mezní teploty – proto by se měla nová technologie ještě vrátit do vývoje.

Zdůrazněme, že tyto závěry jsou stanoveny pouze na základě exploratorní analýzy, statistika nám nabízí exaktnější metody pro rozhodnutí takovýchto případů (testování hypotéz), s nimiž se seznámíte později.

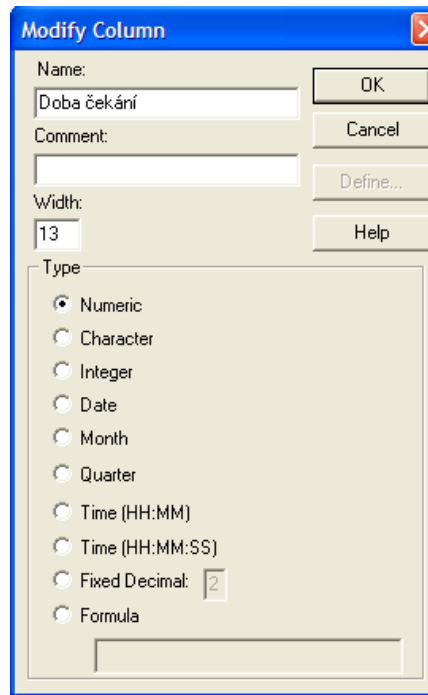
2.6. Následující data představují dobu čekání [min] zákazníka na obsluhu. Proved'te explorační analýzu pomocí Statgraphicsu.

120	80	100	90
150	5	140	130
100	70	110	100

Řešení daného problému ve Statgraphicsu:

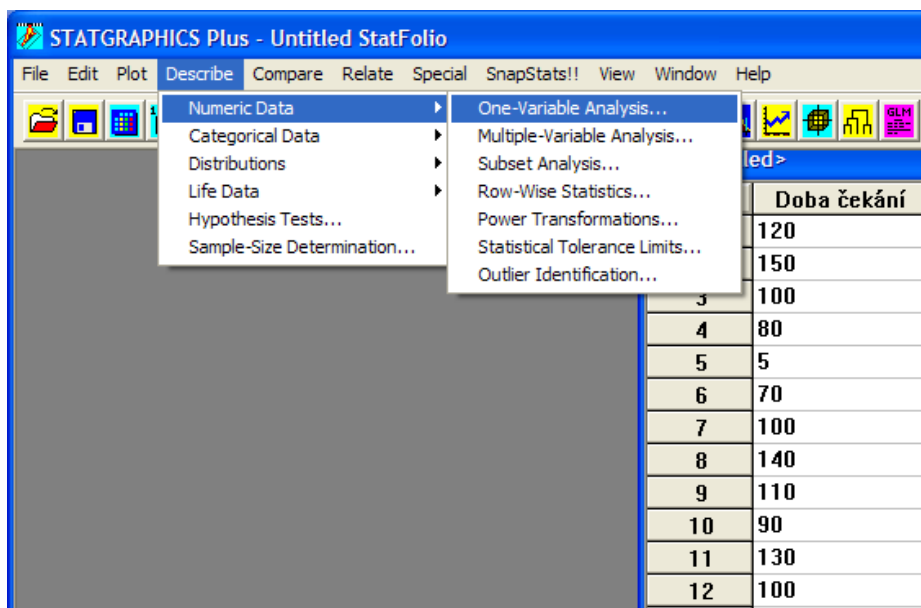
Zadání proměnné:

Chceme-li zadávat ručně novou proměnnou, provedeme DC (dvojklik) na hlavičku sloupce a zadáme parametry proměnné (název, popis (nepovinné), šířku a typ). Přednastavený typ je Numeric, tudíž jej nemusíme měnit.

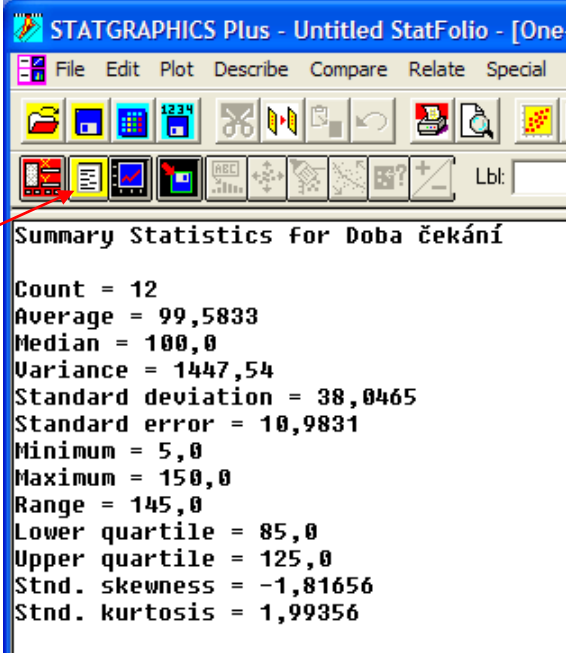


Exploratorní analýza pro numerickou proměnnou:

Textové i grafické výstupy popisné (exploratorní) statistiky získáme obdobně jako u kategoriální proměnné.



Opět si projdeme jednotlivé výstupy exploratorní analýzy.

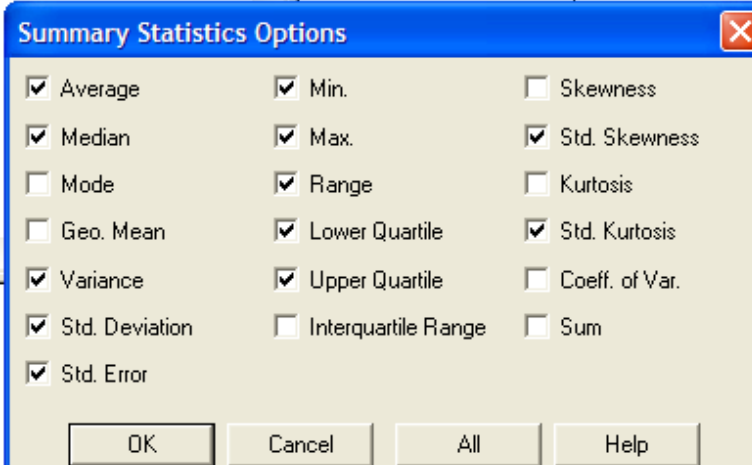


Summary Statistics for Doba čekání

Count = 12
 Average = 99,5833
 Median = 100,0
 Variance = 1447,54
 Standard deviation = 38,0465
 Standard error = 10,9831
 Minimum = 5,0
 Maximum = 150,0
 Range = 145,0
 Lower quartile = 85,0
 Upper quartile = 125,0
 Std. skewness = -1,81656
 Std. kurtosis = 1,99356

Tabular Option

V levém dolním okně najdeme souhrnnou statistiku – tj. vybrané charakteristiky příslušné numerické proměnné (doby čekání). Výběr základních charakteristik, které mají být zobrazeny nám umožní RC na oblast souhrnné statistiky. Po jeho provedení se nám objeví následující okno, v němž zvolíme požadované charakteristiky.



Summary Statistics Options

<input checked="" type="checkbox"/> Average	<input checked="" type="checkbox"/> Min.	<input type="checkbox"/> Skewness
<input checked="" type="checkbox"/> Median	<input checked="" type="checkbox"/> Max.	<input checked="" type="checkbox"/> Std. Skewness
<input type="checkbox"/> Mode	<input checked="" type="checkbox"/> Range	<input type="checkbox"/> Kurtosis
<input type="checkbox"/> Geo. Mean	<input checked="" type="checkbox"/> Lower Quartile	<input checked="" type="checkbox"/> Std. Kurtosis
<input checked="" type="checkbox"/> Variance	<input checked="" type="checkbox"/> Upper Quartile	<input type="checkbox"/> Coeff. of Var.
<input checked="" type="checkbox"/> Std. Deviation	<input type="checkbox"/> Interquartile Range	<input type="checkbox"/> Sum
<input checked="" type="checkbox"/> Std. Error		

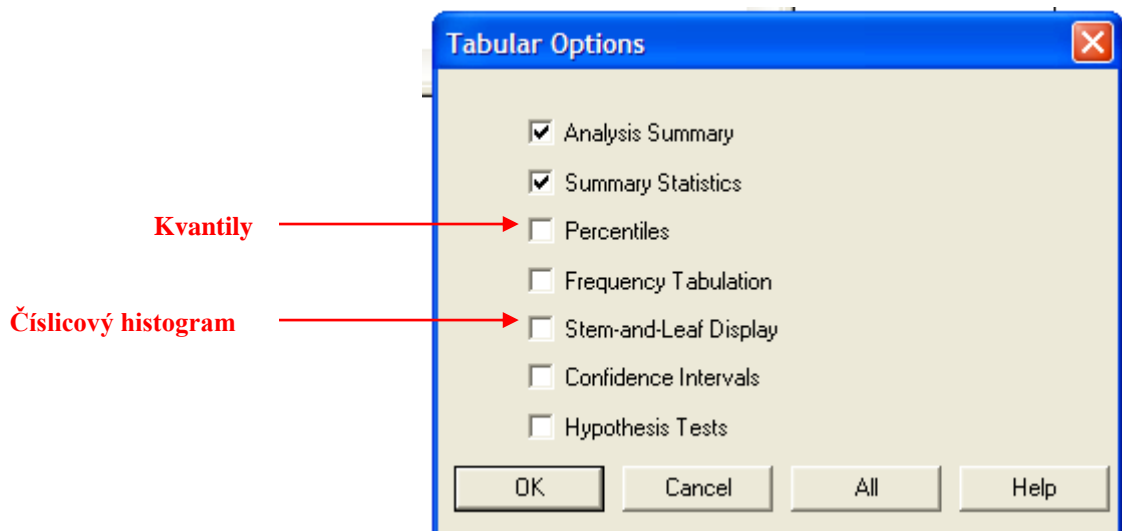
OK Cancel All Help

Slovník názvů jednotlivých charakteristik:

Count	Rozsah souboru (počet hodnot)
Average	Průměr
Median	Medián
Mode	Modus

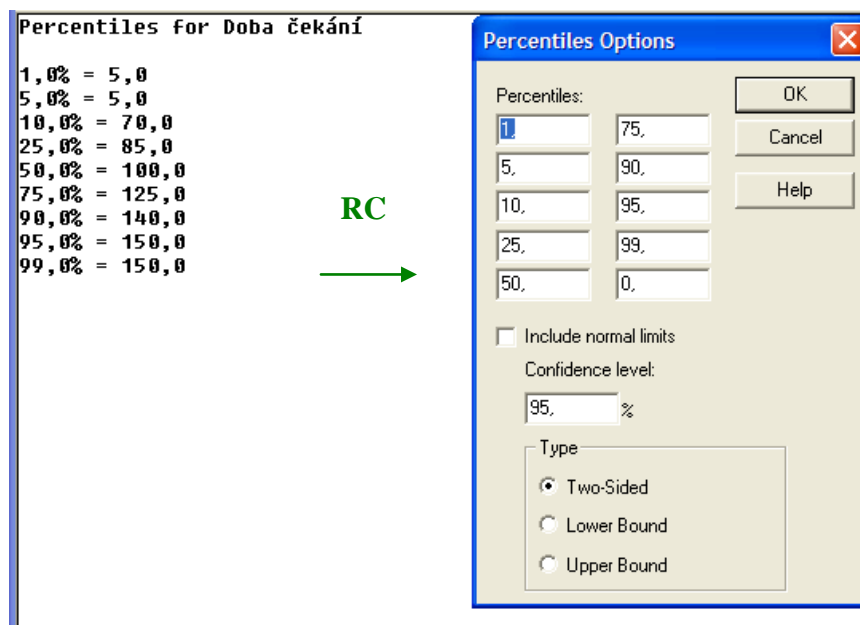
Geo. Mean	Geometrický průměr
Variance	Rozptyl (výběrový)
Std. Deviation	Směrodatná odchylka (výběrová)
Std. Error	Standardní chyba (s/\sqrt{n})
Min.	Minimum
Max.	Maximum
Range	Rozpětí (maximum – minimum)
Lower Quartile	Dolní kvartil
Upper Quartile	Horní kvartil
Interquartile range	Interkvartilové rozpětí (IQR)
Skewness	Šikmost
Std. Skewness	Standardizovaná šikmost
Kurtosis	Špičatost
Std. Kurtosis	Standardizovaná špičatost
Coeff. Of Var.	Variační koeficient (s/\bar{x})
Sum	Součet hodnot

Kliknutím na ikonu **Tabular Options** (žlutá ikona, 2. řádek, 2. zleva) se nám objeví nabídka dalších textových výstupu.

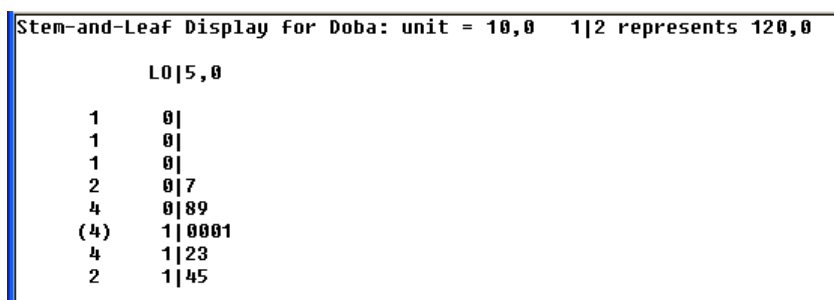


Při popisné statistice nás z této nabídky zajímá pouze možnost volby zobrazení kvantilů a číslcového histogramu.

Zvolíme-li si zobrazení kvantilů, objeví se nám textový výstup s hodnotami deseti přednastavených kvantilů. Jejich výběr můžeme změnit provedeme-li RC na oblast, v níž jsou kvantily zobrazeny a zvolíme-li **Pane Option**.

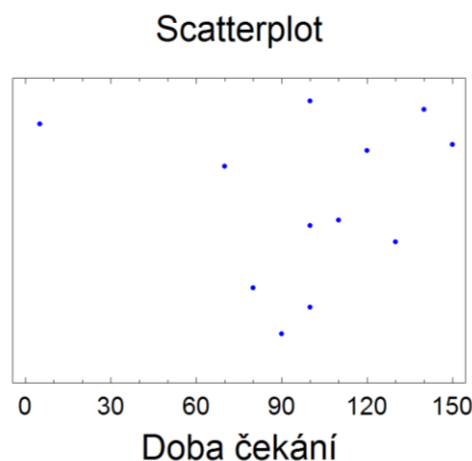


Zvolíme-li v **Tabular Options - Stem and Leaf Display**, získáme Číslcový histogram

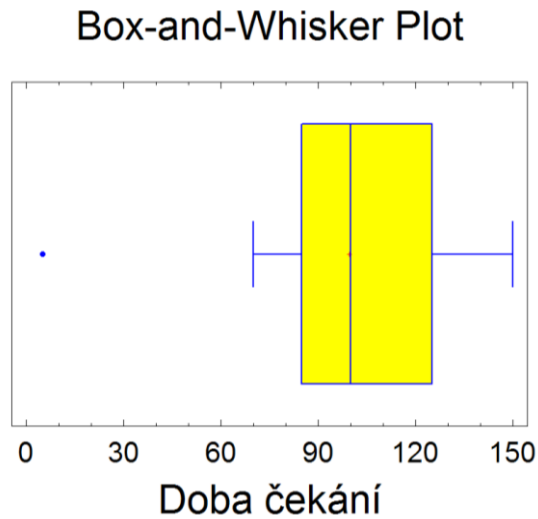


Nyní se zaměříme na pravé horní okno, v němž najdeme tzv. Bodový graf (nazývaný také **rozptylogram**, anglicky Scatterplot). Na ose x jsou v něm vyneseny hodnoty numerické proměnné, na ose y je pořadí, v němž byly hodnoty proměnné zapsány. Je tedy zřejmé, že bodový graf nám umožňuje vizuální posouzení rozptylu proměnné.

Chceme-li změnit grafické parametry bodového grafu, provedeme RC na oblast grafu a požadované parametry nastavíme v menu **Graphics Option**.



V pravém dolním rohu najdeme [Krabicový graf](#). Jeho grafické parametry můžeme obdobně jako u Bodového grafu nastavit v menu **Graphics Option**.



Použité zkratky:

- DC** dvojklik levým tlačítkem myši
- RC** kliknutí pravým tlačítkem myši