Výpočet základních statistických veličin; Vlastnosti směrodatné odchylky a střední chyby průměru; Změna charakteristik variability (disperze) vlivem odlehlých (extrémních) hodnot; Exploračí analýza dat - vizualizace dat frekvenčními histogramy a krabicovými diagramy, percentile plots; Test normality dat.

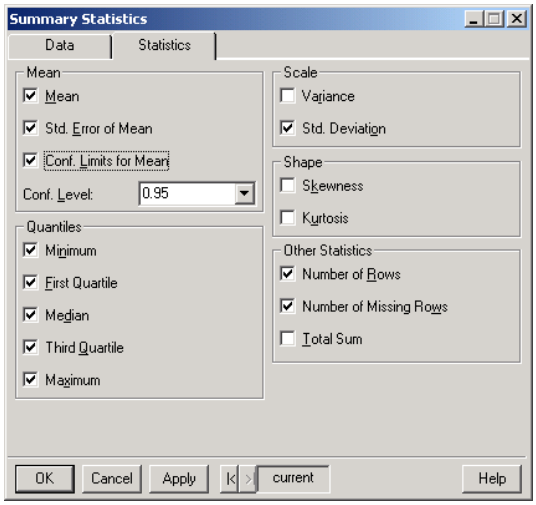
Datový soubor X1
1, 5, 7, 8, 9, 10, 11, 12, 12.5, 13, 14, 15, 16, 17, 18, 20, 24

Po vytvoření nové datové sestavy (SDF1) zadáme hodnoty nové proměnné X1 pomocí příkazového řádku.

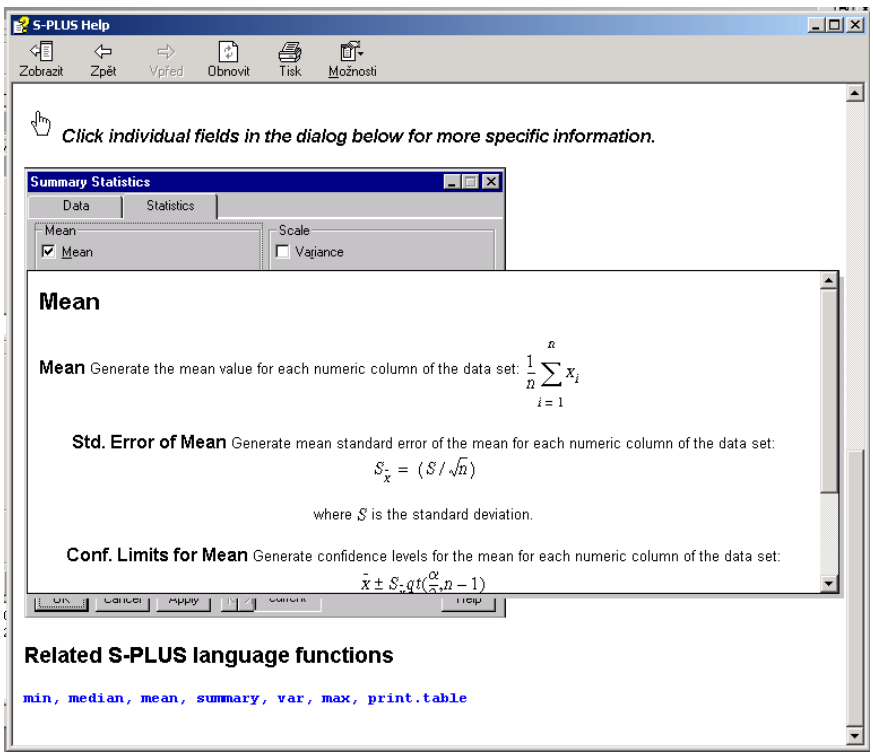
```
SDF1$X1<-c(1, 5, 7, 8, 9, 10, 11, 12, 12.5, 13, 14, 15, 16, 17, 18, 20, 24)
```

1. Základní statistické veličiny

Pro výpočet ***základních charakteristik*** volíme menu: Statistics>Data Summaries>Summary Statistics. Zvolíme jméno datové sestavy, vybereme proměnné a na záložce Statistics zaškrtneme požadované veličiny.



V případě nejasností klávesa F1 aktivuje nápovědu spojenou s daným dialogovým oknem a ke každému políčku je dostupné vysvětlení.



Výsledek se zobrazí v okně "Report Window"

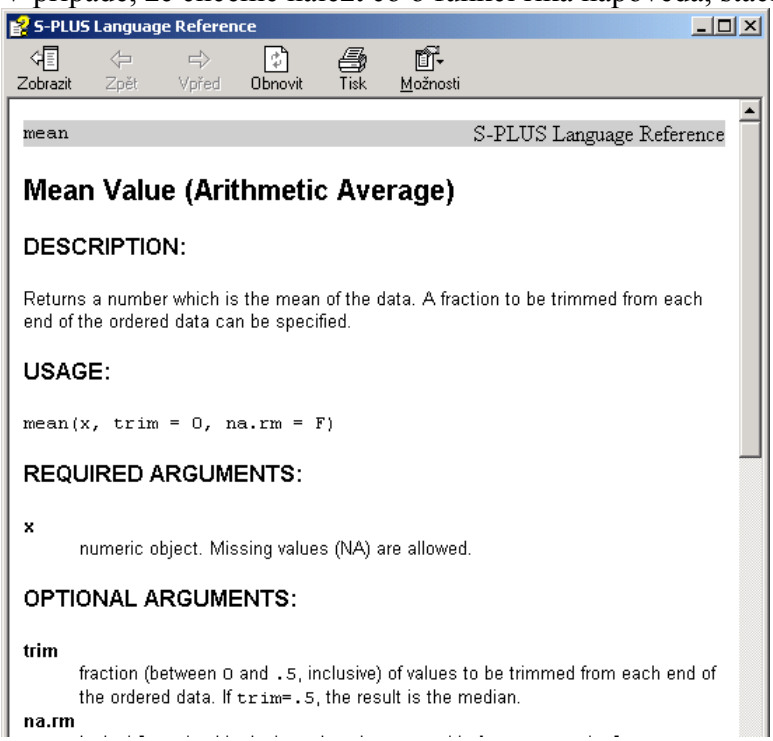
```
       ***  Summary Statistics for data in:  SDF1 ***

numeric matrix: 16 rows, 1 columns.
                    X1
     Min:   1.0000000
 1st Qu.:   9.0000000
    Mean:  12.5000000
  Median:  12.5000000
 3rd Qu.:  16.0000000
     Max:  24.0000000
 Total N:  17.0000000
   NA's :   0.0000000
Variance:  32.5000000
Std Dev.:   5.7008771
     Sum: 212.5000000
 SE Mean:   1.3826658
LCL Mean:   9.5688795 spodní hranice konfidenčního intervalu pro průměr
UCL Mean:  15.4311205 horní hranice konfidenčního intervalu pro průměr
Skewness:   0.0000000
Kurtosis:   0.1665207
```

Pokud nás zajímá pouze průměr, medián, suma či kvantily..., můžeme tyto funkce volat přímo z příkazového řádku. Pro průměr proměnné X1 píšeme `mean (SDF1$X1, na.rm=T)`, kde `na.rm=T` značí, aby S+ vyloučil z výpočtu všechna prázdná pole (pole NA). Pokud v příkazovém řádku napíšeme pouze jméno funkce, vypíše se nám její syntaxe, volitelné a přednastavené parametry. Ještě je třeba před zadáním příkazu dát vědět S+, že daná sestava existuje, použijeme příkaz attach (SDF1). To je třeba použít po každé změně sestavy. Pokud pracujeme s nabídkami z menu, S+ si aktualizuje sestavu sám.

```
> mean
function(x, trim = 0, na.rm = F)
```

V případě, že chceme nalézt co o funkci říká nápověda, stačí zadat např. `help(quantile)`.



Chování směrodatné odchylky a střední chyby průměru a tím i šíře konfidenčního intervalu pro průměr v ***závislosti na počtu měření*** si ukážeme na datovém souboru X2, což je zdvojený datový soubor X1. Data buď zkopírujeme pod sebe, či v příkazovém řádku napíšeme:
`SDF1$X2<-rep(SDF1$X1, 2);` používáme vestavěnou funkci pro opakování `rep`
s parametry: co opakuji, kolikrát. Opět nezapomeňte použít `attach (SDF1).`

```
      ***  Summary Statistics for data in:  SDF1 ***

numeric matrix: 15 rows, 2 columns.
                    X1            X2
     Min:   1.0000000    1.00000000
 1st Qu.:   9.0000000    9.00000000
    Mean:  12.5000000   12.50000000
  Median:  12.5000000   12.50000000
 3rd Qu.:  16.0000000   16.00000000
     Max:  24.0000000   24.00000000
 Total N:  17.0000000   34.00000000
   NA's :  17.0000000    0.00000000
Std Dev.:   5.7008771    5.61383572
     Sum: 212.5000000  425.00000000
 SE Mean:   1.3826658    0.96276488
LCL Mean:   9.5688795   10.54124012
UCL Mean:  15.4311205   14.45875988
Skewness:   0.0000000    0.00000000
Kurtosis:   0.1665207   -0.04713542
```

Charakteristiky polohy (aritmetický průměr, medián a kvartily) jsou u obou proměnných samozřejmě stejné. Směrodatná odchylka se také téměř nezměnila, ale při výpočtu se zdvojnásobila jak suma čtverců odchylek, tak i jmenovatel (z 16 na 33). Dvojnásobný počet měření však výrazně zmenšil střední chybu průměru (z 1.38 na 0.96) a tím i konfidenční interval pro průměr. Střední chyba průměru je směrodatnou odchylkou rozdělení výběrových průměrů.
Toto je pouze **PŘÍKLAD** a tedy **NIKDY** nezdvojujeme datový soubor za účelem zvýšení přesnosti odhadu!!!

2. Vliv extrémních a odlehlých hodnot na charakteristiky variability

Vytvoření nové proměnné s extrémní hodnotou (poslední hodnotu X1 24 jsme nahradili 40). Využijeme funkce volání určité hodnoty v rámci proměnné pomocí hranaté závorky. Nejdříve zkopírujeme X1 do X3 a pak přepíšeme poslední (desátou hodnotu) na 40. Píšeme tedy:

```
SDF1$X3<-SDF1$X1
SDF1$X3[10]<-40
```

Spočítáme souhrny pro X1 a X3

```
      ***  Summary Statistics for data in:  SDF1 ***

numeric matrix: 13 rows, 2 columns.
                  X1         X3
     Min:   1.000000   1.000000
 1st Qu.:   9.000000   9.000000
    Mean:  12.500000  13.441176
  Median:  12.500000  12.500000
 3rd Qu.:  16.000000  16.000000
     Max:  24.000000  40.000000
 Total N:  34.000000  17.000000
   NA's :  17.000000   0.000000
Std Dev.:   5.700877   8.399930
     Sum: 212.500000 228.500000
 SE Mean:   1.382666   2.037282
LCL Mean:   9.568879   9.122331
UCL Mean:  15.431121  17.760022
```

Souhrny ukazují, že charakteristiky polohy jsou ovlivněny jen málo. Medián a kvartily jsou stejné (symetrie souboru se nezměnila, změnili jsme jednu krajní hodnotu), průměr se trochu zvýšil. Variabilita souboru se však zvýšila výrazně, spolu se směrodatnou odchylkou vyrostla střední chyba průměru a rozšířil se tím i interval spolehlivosti pro průměr. Z toho vyplývá, že extrémní hodnota zvýšila variabilitu datového souboru a tedy snížila přesnost odhadu statistických parametrů.

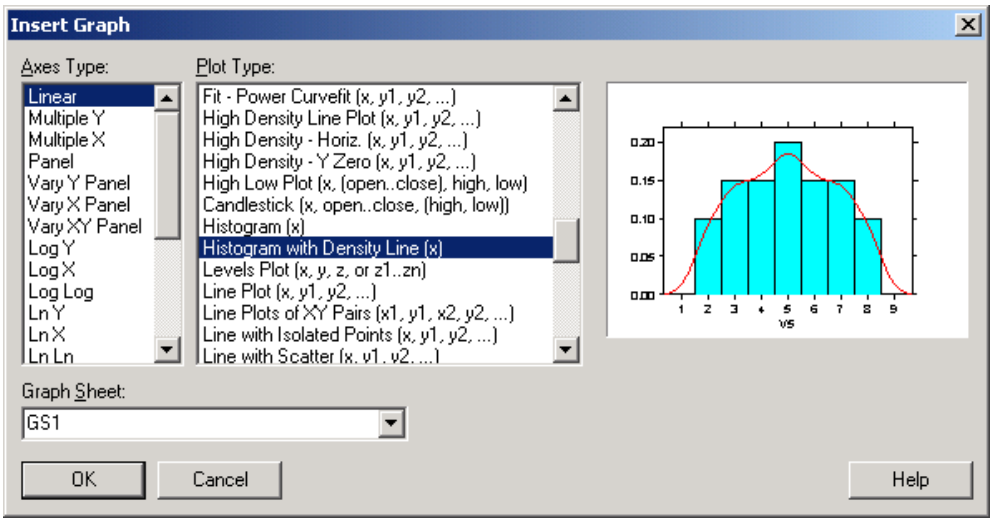
V případě, kdy chceme při jakémkoli výpočtu vyloučit nějakou hodnotu či jejich skupinu z použitého vektoru, S+ nabízí elegantní možnost.
Například spočteme průměr pro vektor X3 bez poslední hodnoty (tj. bez sedmnáctého čísla). Příkaz voláme: `mean (SDF1$X3 [-17])` kdy v hranaté závorce mínus značí, že hodnoty odpovídající podmínce (v tomto případě sedmnácté číslo souboru) nezahrneme do výpočtu. Příkaz `mean (SDF1$X3 [1:10])` naopak spočte průměr z prvních deseti hodnot.
Hodnoty

3. Vizualizace dat

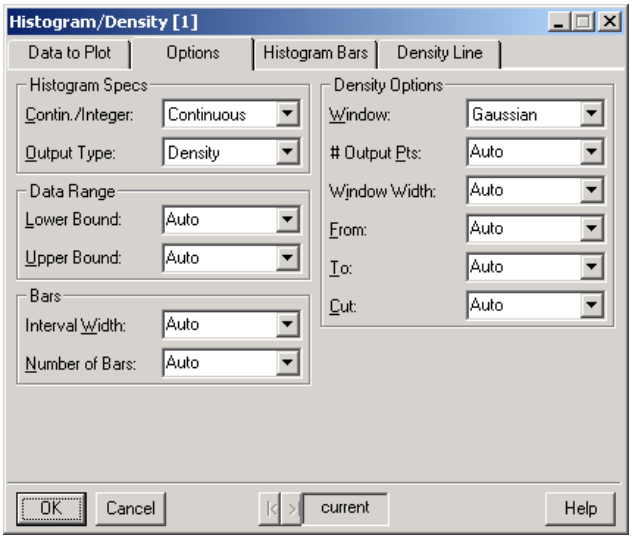
Pro získání základního přehledu o datech nám slouží grafické zobrazení datových souborů. Používáme histogramy četností, krabicové diagramy (box (and whisker) plots) a pravděpodobnostní grafy (Normal Probability Plots či Q-Q Plots).

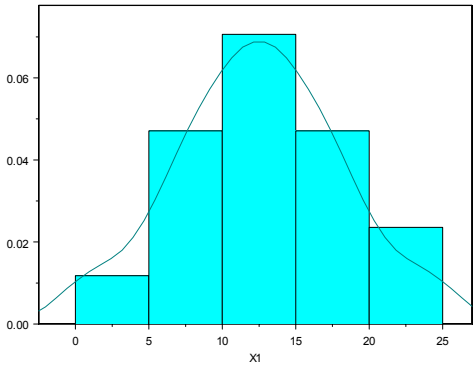
Grafy jsou v S+ k dispozici pod nabídkou Graph>2D Plot..., kde máme na výběr typy os použité v grafu (levé podokno) a typ grafu (pravé podokno).

Pro ***histogram*** četností máme k dispozici graf bez a s čárou denzity rozdělení.

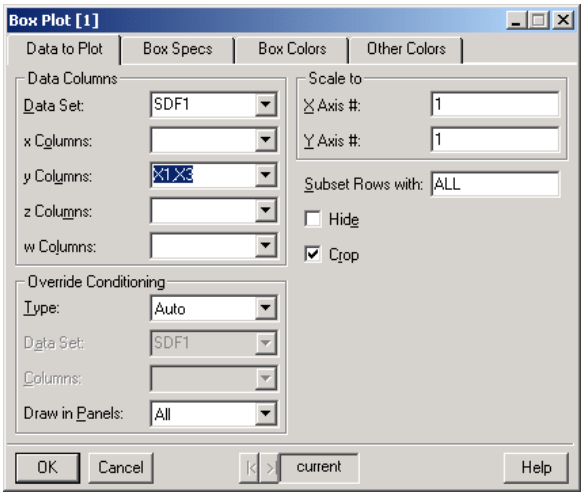


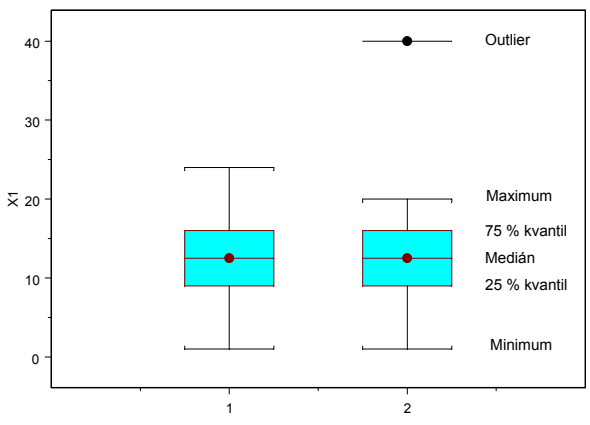
Proměnnou, pro kterou histogram zobrazujeme, zadáváme do pole "x Column". Záložka "Options" umožňuje navolit počet sloupců, šíři intervalů, rozsah histogramu.



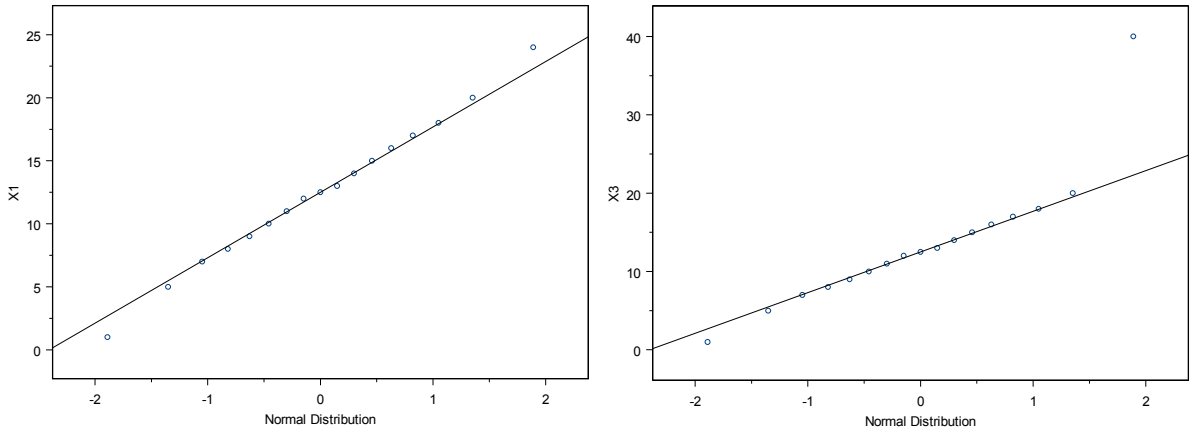


Krabicové diagramy jsou grafy, které zobrazují medián (středová čára), kvartily (hrany krabice), minimum a maximum a případně i odlehlé hodnoty. Odlehlé hodnoty jsou zde definovány jako hodnoty, které jsou mimo 1.5 × mezikvartilové rozmezí. V S+ jsou k dispozici v menu Graph>2D Plots... Plot Type: Box Plot (x, grouping-optional). Proměnné zadáváme do pole "y Column" (můžeme i více proměnných). Pokud máme proměnnou kódovanou pomocí jiné, zadáváme kódující proměnnou do pole "x Column".





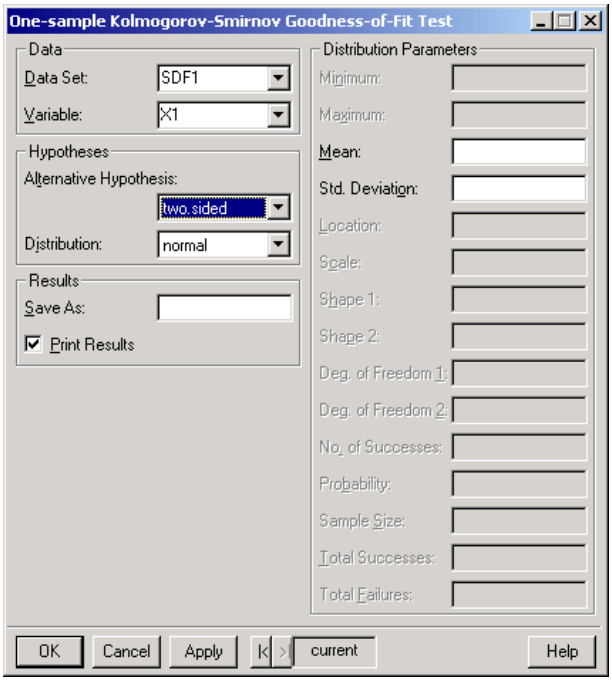
Grafy typu ***Probability Plots / Q-Q-Plots*** znázorňují, do jaké míry je rozdělení našeho datového souboru blízké normálnímu rozdělení. Čím je datový soubor bližší normálnímu rozdělení, tím více se hodnoty blíží přímce (ta odpovídá normálnímu rozdělení). Na osu x jsou vyneseny percentily normálního rozdělení a na y ose jsou sledované hodnoty. V S+ je tento graf pod nabídkou Graph>2D Plot...>QQ Normal with Line (x). Sledovanou proměnnou opět zadáváme do pole "y Column".



Q-Q graf pro proměnnou X1 ukazuje, že datový soubor odpovídá normálnímu rozdělení. Avšak pro proměnnou X3 je normalita výrazně narušena odlehlou hodnotou.

4. Test normality

Pro některé statistické analýzy je potřeba splnit určité předpoklady. Nejčastějším je podmínka, aby data odpovídala normálnímu rozdělení. Spolu s grafickým zobrazením dat (histogramy, krabicové diagramy a Q-Q diagramy) můžeme použít pro spojité proměnné Kolmogorovův-Smirnovův test. V S+ je dostupný z menu Statistics>Compare Samples>One Sample>Kolmogorov-Smirnov GOF... V následném dialogu vybereme sledovanou proměnnou. Průměr a SD nevyplňujeme, S+ automaticky doplní odhadnutý průměr testované proměnné.



Výsledek testu normality z okna "Results" ukazuje hodnotu testovací statistiky "ks" a výslednou dosaženou hladinu významnosti. Zde v tomto případě je p hodnota 0.5, což znamená, že dle K-S testu má náš testovaný soubor atributy normálního rozdělení.

```
        One sample Kolmogorov-Smirnov Test of Composite Normality

data:  X1 in SDF1
ks = 0.0533, p-value = 0.5
alternative hypothesis:
  True cdf is not the normal distn. with estimated parameters
sample estimates:
 mean of x standard deviation of x
      12.5                5.700877
```

Testy normality jsou však značně nespolehlivé pokud je počet měření malý (menší než cca 100). Proto je vždy vhodné ověřit si rozložení dat souboru vizuální kontrolou grafických zobrazení.

Tento materiál byl vytvořen na základě textů a příkladů P. Sklenáře pro kurz biostatistiky v NCSS.

Příloha: Vybrané funkce pro práci s vektory

`max(x)`	maximum vektoru x
`min (x)`	minimum vektoru x
`sum (x)`	součet všech hodnot vektoru x
`mean (x)`	aritmetický průměr vektoru x
`median (x)`	medián vektoru x
`range (x)`	vektor obsahující min a max
`stdev (x)`	směrodatná odchylka
`var (x)`	variance
`sort (x)`	setřídí vektor x
`rank(x)`	vytvoří vektor pořadí jednotlivých hodnot vektoru x
`quantile (x)`	vektor obsahující min, 25% kvantil, medián, 75% kvantil a maximum

`colMean (x)` sloupcové průměry pro datovou sestavu (x)
`colVar (x)` ...
`rowMean (x)` řádkové průměry pro datovou sestavu (x)
...

Mocniny a odmocniny:
2^2 píšeme `2^2`
odmocninu píšeme jako mocninu zlomku, tedy třetí odmocnina osmi:
`8^(1/3)`

Porovnávání hodnot:

≤	píšeme <=
≥	>=
=	==

Odpovědí programu je dvojice hodnot T (true, 1) a F (false, 0)

Některé další funkce:

`log (x, n)`	logaritmus x o základu n
`log (x)`	přirozený logaritmus
`log10(x)`	logaritmus o základu 10
`exp (x)`	e^x
`sqrt (x)`	druhá odmocnina x
`cos (x), sin (x) , tan(x)` `acos (x), asin (x), atan(x)`	trigonometrické fce
`abs (x)`	absolutní hodnota x
`floor (x)`	největší celé číslo menší než x
`ceiling (x)`	nejmenší celé číslo větší než x
`trunc (x)`	nejbližší celé číslo mezi x a 0
`round (x, digits=0)`	zaokrouhlí x na celé číslo

dělení se zbytkem:
```
21 % / % 4
5
```
zjištění zbytku (tzv. modulo)
```
21 % % 4
1
```

Příklady použití:
vytvoříme vektor x obsahující čísla 1 až 10
`x<-1:10`
součet všech hodnot: `sum(x)`
Pozor! Pokud chceme použít součet čísel splňujících určité kritérium (např. čísla menší než 5), pak voláme: `sum(x[x<5])`
Zadání `sum(x<5)` vyvolá počet hodnot, která splňují námi zadané kritérium.