Laboratorní cvičení - Statistika

Základní pojmy

balíček: Statistics

Pro veškeré výpočty je třeba načíst balíček **Statistic**. Při řešení můžeme použít proceduru **infolevel[Statistics]:=1,** která nám poskytne podrobný výpis informací vztahující se k danému výpočtu.
Dále budou uvedeny a na příkladech ukázány jen některé základní pojmy z přednášek. Balíček Statistic je výrazně bohatší, případné zájemce o další zpracování statistického materiálu odkazuji na Help.

Příklad.

```
> with(Statistics):
> infolevel[Statistics]:=1;
```

$$infolevel_{Statistics} := 1$$

Nastavíme počáteční hodnotu generátoru náhodných čísel, aby byl výsledek gererování pokaždé stejný.

```
> with(RandomTools[MersenneTwister]):
> SetState(state=3141592653589);
>
data:=[6.5,6.2,5.5,5.25,4.8,4.75,4.2,3.5,1.5,1.4,0.75,0.575,0.5,0
.46,0.35,0.315,0.29,0.1425,0.1375,0.135,0.125,0.1115,0.1115,0.109
,0.109,0.109];
```

$$data := [6.5, 6.2, 5.5, 5.25, 4.8, 4.75, 4.2, 3.5, 1.5, 1.4, 0.75, 0.575, 0.5, 0.46, 0.35, 0.315, 0.29, 0.1425, 0.1375, 0.135, 0.125, 0.1115, 0.1115, 0.109, 0.109, 0.109]$$

Zadaná data setřiďte do neklesající posloupnosti.
Příkaz **sort** setřídí zadaná data do neklesající posloupnosti.

```
> sort(data);
```

$$[0.109, 0.109, 0.109, 0.1115, 0.1115, 0.125, 0.135, 0.1375, 0.1425, 0.29, 0.315, 0.35, 0.46, 0.5, 0.575, 0.75, 1.4, 1.5, 3.5, 4.2, 4.75, 4.8, 5.25, 5.5, 6.2, 6.5]$$

Nejmenší $x_{\min}$ a největší $x_{\max}$ hodnota zadaných dat.
Příkaz **min** najde nejmenší $x_{\min}$ a příkaz **max** největší $x_{\max}$ hodnotu v zadaných datech.

```
> min(data);
```

$$0.109$$

```
> max(data);
```

$$6.5$$

Variační rozpětí $R = x_{\max} - x_{\min}$.
Variační rozpětí $R = x_{\max} - x_{\min}$ nalezeme pomocí příkazu **Range.**

```
> Range(data);
```

$$6.391000000$$

Výběrový průměr $x_p = \frac{1}{n}\sum_{i=1}^{n} x_i$, nalezneme příkazem **Mean**.

```
> Mean(data);
```

$$1.843461538$$

Rozptyl $s^2 = \frac{1}{(n-1)}\left[\sum_{i=1}^{n} x_i^2 - n \cdot x_p^2\right]$,

stanovíme příkazem **Variance.**

```
> Variance(data);
```

$$5.26568033846154$$

Směrodatná odchylka $s = \sqrt{s^2}$.

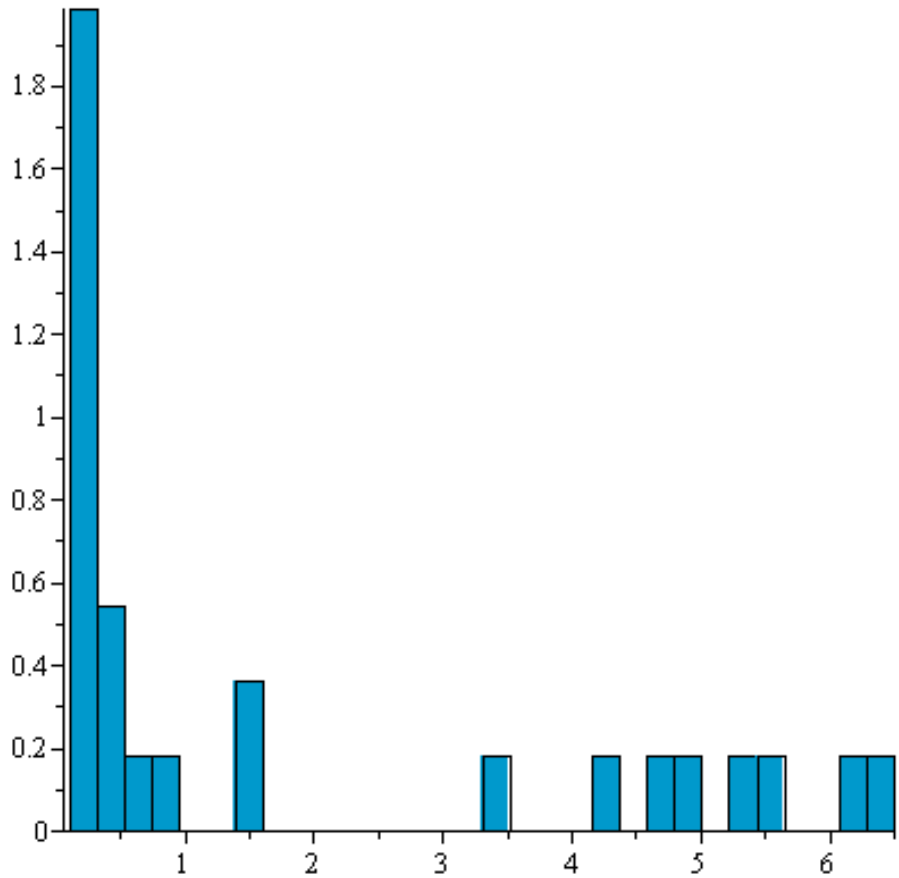
```
> StandardDeviation(data);
```

$$2.29470702671638$$

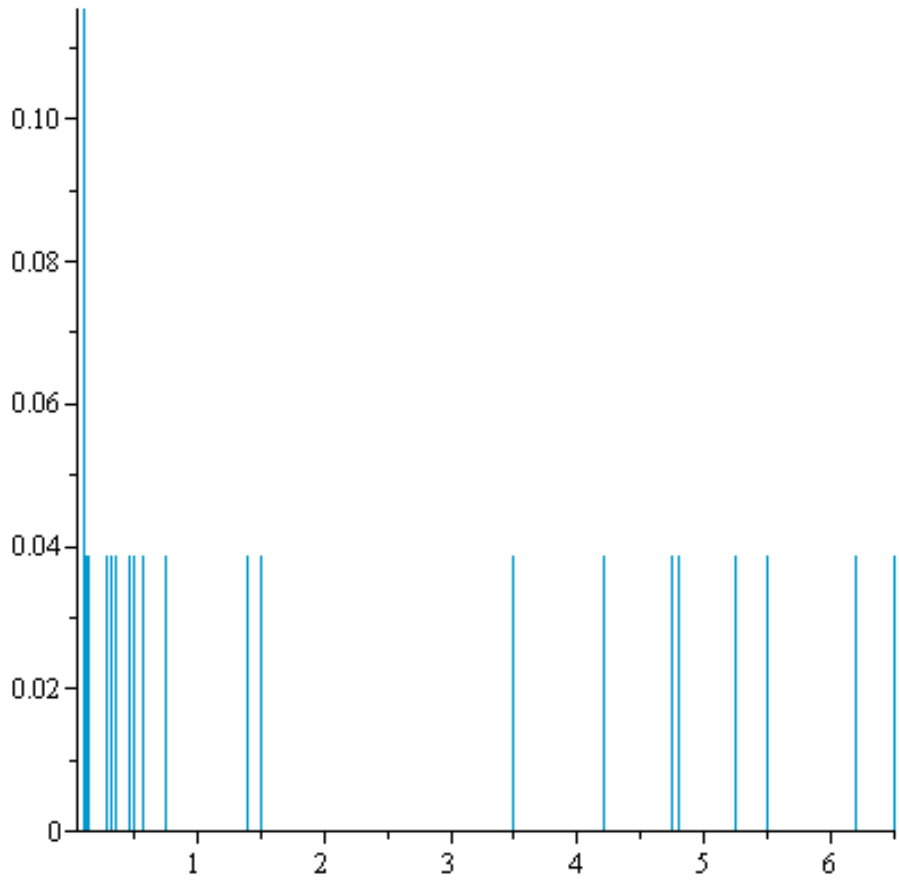
Histogram

Příkaz **Histogram** vytvoří histogram pro zadaná data. Je možné zadat, že náhodná proměnná je diskrétní (pak je výsledkem tyčkový diagram), říci, zda se použijí relativní nebo absolutní četnosti, nebo zadat počet tříd.

```
> Histogram(data);
Histogram Type:  default
Data Range:      .1090000000 .. 6.500000000
Bin Width:       .2130333333
Number of Bins:  30
Frequency Scale: relative
```

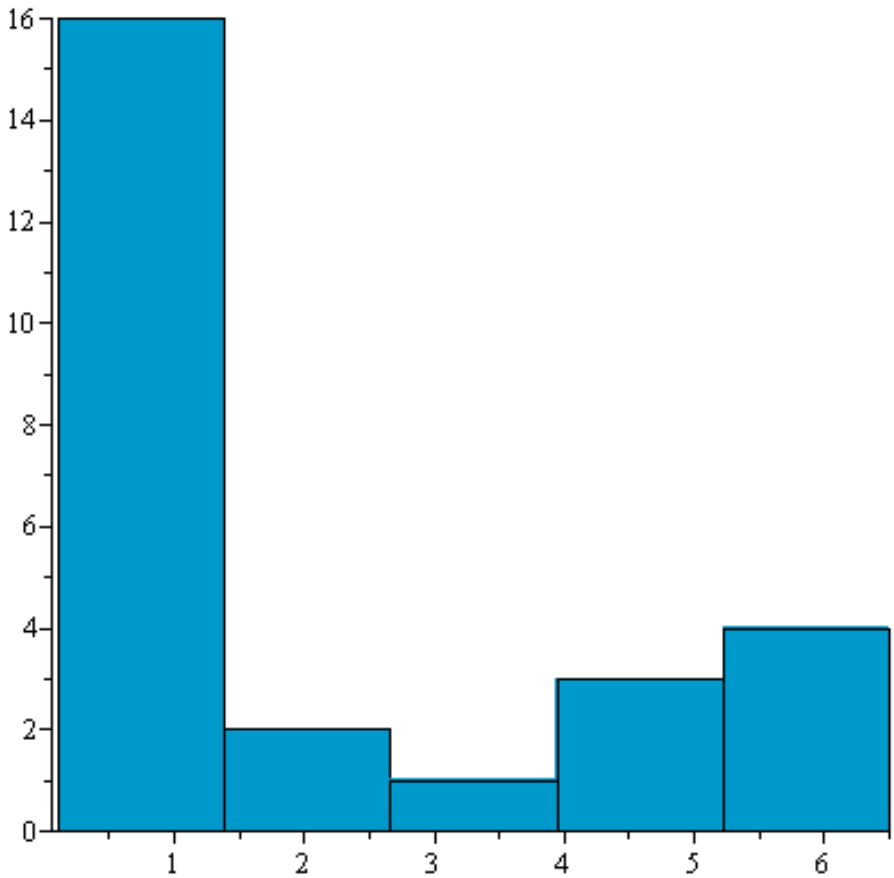



```
> Histogram(data,discrete=true);
Histogram Type:  discrete
Data Range:      .1090000000 .. 6.500000000
Number of Bins:  23
Frequency Scale: relative
```



V případě, že chceme změnit počet intervalů v histogramu, je potřeba přidat parametr **bincount=n,**kde n je celé číslo.

```
> Histogram(data, bincount=5,frequencyscale = absolute);
Histogram Type:  default
Data Range:      .1090000000 .. 6.500000000
Bin Width:       1.278200000
Number of Bins:  5
Frequency Scale: absolute
```



Rozdělení datového souboru do tříd. Počet tříd je interně nastaven na 10.
FrequencyTableobsahuje 5 sloupců. V prvním je variační rozpětí, ve druhém the absolute frequency, ve třetím the percentage, ve čtvrtém the cumulative frequency a v pátém the cumulative percentage of the data.

```
> FrequencyTable(data);
```

0.1090000000	..0.7481000000	15.	57.69230769	15.	57.69230769
0.7481000000	..1.387200000	1.	3.846153846	16.	61.53846154
1.387200000	..2.026300000	2.	7.692307692	18.	69.23076923
2.026300000	..2.665400000	0.	0.	18.	69.23076923
2.665400000	..3.304500000	0.	0.	18.	69.23076923
3.304500000	..3.943600000	1.	3.846153846	19.	73.07692308
3.943600000	..4.582700000	1.	3.846153846	20.	76.92307692
4.582700000	..5.221800000	2.	7.692307692	22.	84.61538462
5.221800000	..5.860900000	2.	7.692307692	24.	92.30769231
5.860900000	..6.500000000	2.	7.692307692	26.	100.0000000

Pokud chceme změnit počet tříd, je potřeba použít příkaz **bins=n,**kde n je celé čílo.

```
> FrequencyTable(data,bins=5);
```

$$\begin{bmatrix} 0.1090000000..1.387200000 & 16. & 61.53846154 & 16. & 61.53846154 \\ 1.387200000..2.665400000 & 2. & 7.692307692 & 18. & 69.23076923 \\ 2.665400000..3.943600000 & 1. & 3.846153846 & 19. & 73.07692308 \\ 3.943600000..5.221800000 & 3. & 11.53846154 & 22. & 84.61538462 \\ 5.221800000..6.500000000 & 4. & 15.38461538 & 26. & 100.0000000 \end{bmatrix}$$

Medián je 50%-ní kvantil. Je používán v těch případech, kdy náhodná veličina nemá definovánu střední hodnotu. Obecně není určen jednoznačně.

```
> Median(data);
```

$$0.4800000000$$

```
> Skewness(data);
```

$$0.910065788271881$$

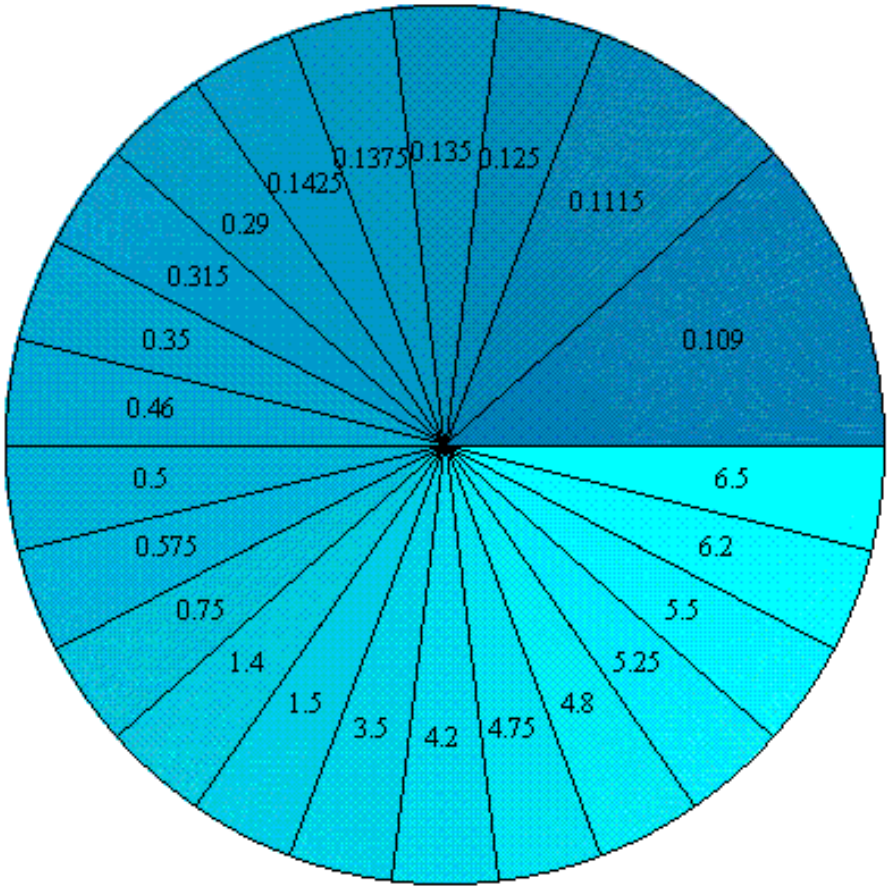
The Quantile function computes the quantile corresponding to the given probability p for the specified random variable or data set.

```
> Quantile(data,1/3);
```

$$0.1588888889$$

Data lze znázornit různými diagramy, jedním z nich je např. kruhový (koláčový) diagram.

```
> PieChart(data);
```



Definování náhodné proměnné a generování výběrů

Náhodnou proměnnou definujeme pomocí příkazu **RandomVariable(typ rozdělení)**, kde

napíšeme o jaké rozdělení a s jakými parametry by se mělo jednat. Pokud chceme vidět vygenerované hodnoty, je potřeba použít příkazu**Sample(název náh. prom., počet vygenerovaných prvků).**
Následně pak můžeme počítat veškeré charakteristiky, které potřebujeme.
Př. 1 Obchodní cestující prodává pračky. Na obchodní cesty jezdí se čtyřmi pračkami. Statisticky má zjištěno, že průměrně dva z devíti zákazníků, kterým pračku nabídne, si ji koupí. Chce odhadnout pravděpodobnosti pro počet prodaných praček a střední hodnotu tohoto počtu po čtyřech nabídkách.
Zřejmě jde o binomické rozdělení s parametry n=4 a p=2/9.

```
> R := RandomVariable(Binomial(4, 2/9));
```

$$R := \_R$$

Pokud bycho chtěli vidět jednu z možnosí prodeje obchodního cestujícího během 10 cest, je možné si je nechat vygenerovat.

```
> S:=Sample(R,10);
```

$$S := \left[\ 0.\ \ 0.\ \ 0.\ \ 1.\ \ 0.\ \ 0.\ \ 2.\ \ 0.\ \ 1.\ \ 2.\ \right]$$

Pokud chceme zjistit s jakou pravděpodobnosí prodá na svých cestách jednu pračku, využijeme příkazu **ProbabilityFunction.**Zde do závorky zadáme jméno náhodné proměnné, uvedeme v jakém čísle chceme pravděpodobnost určit a jako nepovinný parametr zadáme numeric, tj. výsledek bude zapsán desetinným číslem.

```
> f:=ProbabilityFunction(R, 1,numeric);
```

$$f := 0.4182289285$$

POZN: Pro zajímavost si můžete nechat vygenerovat několikrát úspěšný prodej obchodního cestujícího, zjistíte, že číslo 4, tj. prodej čtyř praček, se v něm téměř nevyskytne. Určete si pravděpodobnost prodeje čtyř praček.
Střední hodnotu a rozptyl počtu prodaných praček určíme následovně:

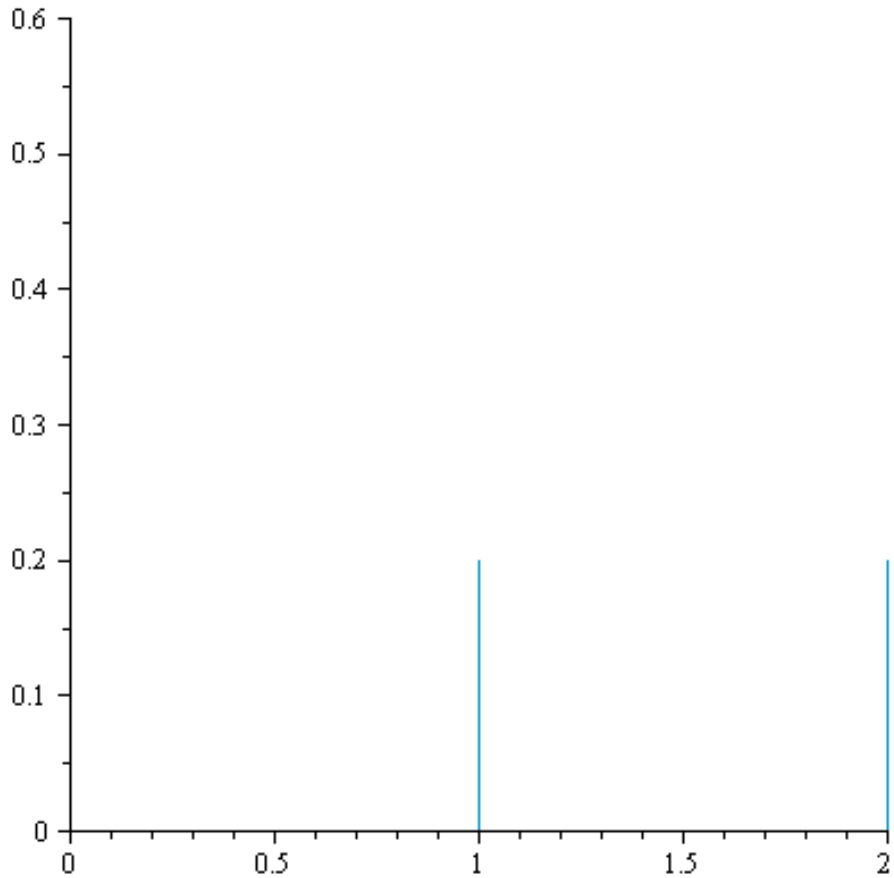
```
> Mean(R);
```

$$\frac{8}{9}$$

```
> Variance(R);
```

$$\frac{56}{81}$$

```
> Histogram(S,discrete=true);
Histogram Type:  discrete
Data Range:      0. .. 2.
Number of Bins:  3
Frequency Scale: relative
```



Př. 2 Stroj, který vyrábí součásty, je seřízen tak, aby střední hodnota jejich délek byla 42 mm. Přesnost je taková, že směrodatná odchylka délky součástek je 1,2 mm. Vygenerujte sérii deseti součástek a spočítejte s jakou pravděpodobností bude vyrobena součástka délky 44,3 mm.

```
> N:=RandomVariable(Normal(42,1.2));
```

$$N := \_R0$$

```
> S1:=Sample(N,10);
```

$$S1 := [42.5938118480705, 40.5414220478957, 40.6288291380228, 40.6204677304127, 43.3259055936066, 42.9136661418140, 43.7208199152479, 42.0070429979413, 41.8001278070596, 42.1465348252823]$$

```
> ProbabilityDensityFunction(Normal(a,b),t);
```

$$\frac{1}{2}\,\frac{\sqrt{2}\,e^{-\frac{1}{2}\,\frac{(t-a)^2}{b^2}}}{\sqrt{\pi}\,b}$$

```
> ProbabilityDensityFunction(N,t);
```

$$\frac{0.4166666667\,\sqrt{2}\,e^{-0.3472222222\,(t-42)^2}}{\sqrt{\pi}}$$

```
> ProbabilityDensityFunction(N,44.3);
```

$$0.05296808873$$

```
> ProbabilityDensityFunction(N,42,numeric);
```

$$0.3324519002$$

Intervaly spolehlivosti

Intervaly spolehlivosti určíme podle vzorců ze skript. Při výpočtu použijeme příkaz **Quantile(typ rozdělení,hladina)**.

Příklad. Určete 95% interval spolehlivosti pro střední hodnotu a směrodatnou odchylku z daného datového souboru.

```
> NPX1:=RandomVariable(Normal(3,2.2));
```

$$NPX1 := \_R2$$

```
> Data1:=Sample(NPX1,300);
```

$$Data1 := \begin{bmatrix} 1 .. 300\ Vector_{row} \\ Data\ Type: float_8 \\ Storage: rectangular \\ Order: Fortran\_order \end{bmatrix}$$

```
> prum1:=Mean(Data1); sm_od1:=StandardDeviation(Data1);
```

$$prum1 := 3.030304024$$

$$sm\_od1 := 2.11908483073986$$

Intervalový odhad střední hodnoty.

```
>
d1:=evalf(prum1-sm_od1/sqrt(300)*Quantile(StudentT(299),0.975));
```

$$d1 := 2.78953684255501$$

```
>
h1:=evalf(prum1+sm_od1/sqrt(300)*Quantile(StudentT(299),0.975));
```

$$h1 := 3.27107120544499$$

Intervalový odhad směrodatné odchylky.

```
> dd1:=sqrt(299*sm_od1^2/Quantile(ChiSquare(299),0.975));
```

$$dd1 := 1.96200099896890$$

```
> hh1:=sqrt(299*sm_od1^2/Quantile(ChiSquare(299),0.025));
```

$$hh1 := 2.30372246787967$$

Testy dobré shody a další testy statistických hypotéz

Pro účely, které se týkají našich výpočtů, vystačíme z balíčku **Statistics**s částí nazvanou **Tests**. Při řešení
problému, zda je náhodná proměnná vybrána z daného rozdělení, využijeme test **ChiSquareSuitableModelTest**.
Funkce **ChiSquareSuitableModelTest(X, F, options)**testuje shodu vhodného modelu odpovídajícího pozorovaným datům a známé
náhodné proměnné nebo rozdělení pravděpodobnosti. Test se pokouší po setřídění užitím testu dobré shody
určit, zda lze daný vzorek považovat za vybraný z dané náhodné proměnné nebo rozdělení

pravděpodobnosti.
První parametr **X**je jednorozměrná **r**-tabulka pozorovaných dat, která mají být analyzována.
Druhý parametr **F**je náhodná proměnná nebo rozdělení pravděpodobnosti, které je srovnáváno se souborem pozorovaných dat.
Pokud nezadáme jinak, proběhne test na hladině významnosti 0,05.
Příklad. Byla měřena doba (v minutách) mezi poruchami stroje, získané hodnoty jsou:

63	278	4	323	415	100	529	272	46	188
156	15	35	275	140	189	310	117	236	124
184	561	73	176	17	176	22	169	387	385
786	229	203	427	293	672	69	466	92	174
822	8	178	196	991	134	130	286	177	23

Určete výběrový průměr x, směrodatnou odchylku s, histogram četností a typ rozdělení doby mezi poruchami stroje.
Při řešení využijeme proceduru **infolevel[Statistics]:=1.** Při jejím použití nám Maple poskytne podrobný výpis informací vztahující se k danému výpočtu.
Při řešení budeme postupovat následovně: sestrojíme nejprve histogram a pokusíme se odhadnout z jakého typu rozdělení je daný vzorek.

```
> with(Statistics):
> infolevel[Statistics]:=1;
```

$$infolevel_{Statistics} := 1$$

```
> X:=Array([63,278,4,323,415,100,529,272,46,188,156,15,35,275,140,189,310,117,236,124,184,561,73,176,17,176,22,169,387,385,786,229,203,427,293,672,69,466,92,174,822,8,178,196,991,134,130,286,177,23]);
```

$$X := \begin{bmatrix} 1 .. 50\ Array \\ Data\ Type:\ anything \\ Storage:\ rectangular \\ Order:\ Fortran\_order \end{bmatrix}$$

Nejprve vypočítáme výběrový průměr náhodné veličiny X a výběrovou směrodatnou odchylku.

```
> X1:=Mean(X);
```

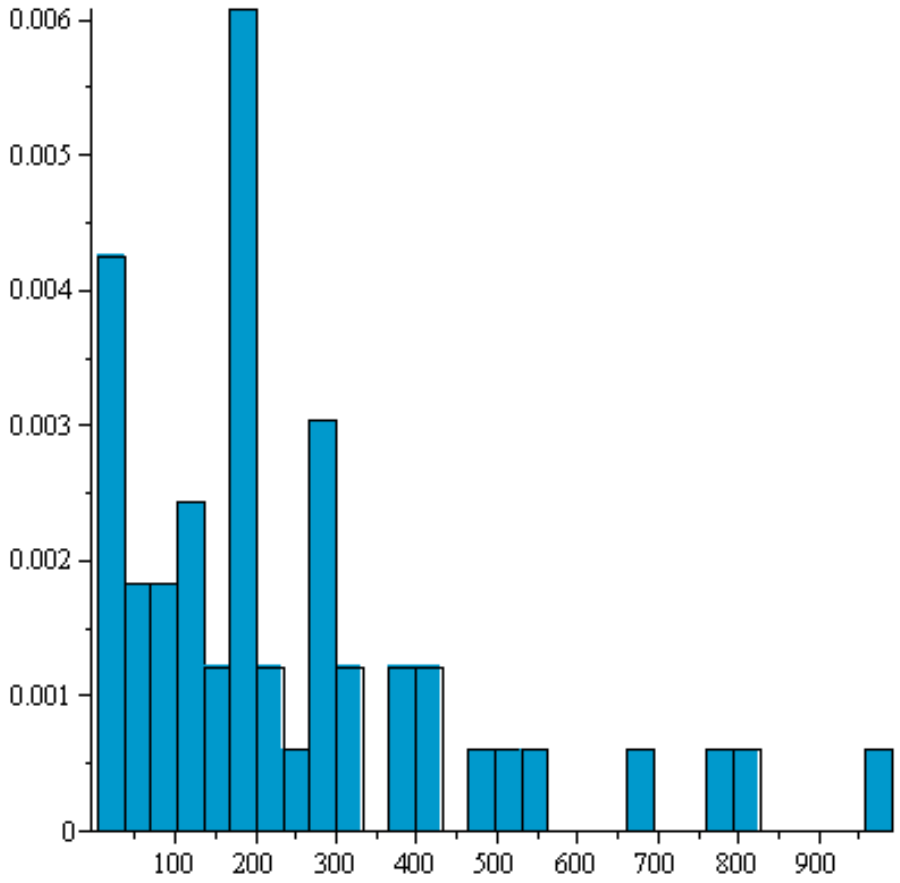
$$X1 := 246.4200000$$

```
> S:=StandardDeviation(X);
```

$$S := 220.020879343163$$

Dále sestrojíme nejprve histogram a pokusíme se odhadnout z jakého typu rozdělení je daný vzorek.

```
> Histogram(X);
Histogram Type:   default
Data Range:       4. .. 991.
Bin Width:        32.90000000
Number of Bins:   30
Frequency Scale:  relative
```



Podle tvaru histogramu by v úvahu mohlo připadat buď normální nebo exponenciální rozdělení. K otestování hypotézy o daném typu rozdělení použijeme **ChiSquareSuitableModelTest**. Pro test je potřeba znát příslušné empirické charaktristiky (výběrový průměr a výběrovou směrodatnou odchylku), které pak slouží jako bodové odhady parametrů testovaných rozdělení.

Nejprve otestujeme, zda je daný vzorek z normálního rozdělení.

```
> ChiSquareSuitableModelTest(X, Normal(X1,S));
Chi-Square Test for Suitable Probability Model
----------------------------------------------
Null Hypothesis:
Sample was drawn from specified probability distribution
Alt. Hypothesis:
Sample was not drawn from specified probability distribution

Bins:                    8
Distribution:            ChiSquare(7)
Computed statistic:      16.8465
Computed pvalue:         0.0184134
Critical value:          14.06714058

Result: [Rejected]
There exists statistical evidence against the null hypothesis
```

$$hypothesis = false, criticalvalue = 14.06714058, distribution = ChiSquare(7), pvalue = 0.0184134120, statistic = 16.84647588$$

Protože byla zamítnuta nulová hypotéza ve prospěch alternativní, provedeme další test, a to zda je daný vzorek vybrán z exponenciálního rozdělení.

```
> ChiSquareSuitableModelTest(X, Exponential(S));
Chi-Square Test for Suitable Probability Model
----------------------------------------------
Null Hypothesis:
Sample was drawn from specified probability distribution
Alt. Hypothesis:
Sample was not drawn from specified probability distribution

Bins:                    8
Distribution:            ChiSquare(7)
Computed statistic:      3.94721
Computed pvalue:         0.785838
Critical value:          14.06714058

Result: [Accepted]
There is no statistical evidence against the null hypothesis
```

$$hypothesis = true, criticalvalue = 14.06714058, distribution = ChiSquare(7), pvalue = 0.7858377360, statistic = 3.947207125$$

V tomto případě byla nulová hypotéza přijata. A můžeme tedy konstatovat, že na 5% hladině významnosti vzorek pochází z exponenciálního rozdělení.

Testování, zda je střední hodnota rovna danému číslu (výběr z normálního dělení).

```
> NPX2:=RandomVariable(Normal(2.3,3.8));
```

$$NPX2 := \_R9$$

```
> Data2:=Sample(NPX2,500);
```

$$Data2 := \begin{bmatrix} 1 .. 500\ Vector_{row} \\ Data\ Type: float_8 \\ Storage: rectangular \\ Order: Fortran\_order \end{bmatrix}$$

```
> Mean(Data2); StandardDeviation(Data2);
```

$$2.064108376$$

$$3.95607663699109$$

```
> OneSampleTTest(Data2,2.5);
Standard T-Test on One Sample
-----------------------------
Null Hypothesis:
Sample drawn from population with mean 2.5
Alt. Hypothesis:
Sample drawn from population with mean not equal to 2.5
```

```
Sample size:              500
Sample mean:              2.06411
Sample standard dev.:     3.95608
Distribution:             StudentT(499)
Computed statistic:       -2.46376
Computed pvalue:          0.0140849
Confidence interval:      1.71650624038993 .. 2.41171051161007
                          (population mean)

Result: [Rejected]
There exists statistical evidence against the null hypothesis
```

$hypothesis = false, confidenceinterval = 1.71650624038993$
$..2.41171051161007, distribution = StudentT(499), pvalue$
$= 0.01408494466, statistic = -2.46376243765148$

Totéž pro směrodatnou odchylku.

```
> OneSampleChiSquareTest(Data2,3.5);
Chi-Square Test on One Sample
-----------------------------
Null Hypothesis:
Sample drawn from population with standard deviation equal to 3.5
Alt. Hypothesis:
Sample drawn from population with standard deviation not equal to 3.5

Sample size:              500
Sample standard dev.:     3.95608
Distribution:             ChiSquare(499)
Computed statistic:       637.52
Computed pvalue:          5.0154e-05
Confidence interval:      3.72513466617362 .. 4.21777770711276
                          (population standard deviation)

Result: [Rejected]
There exists statistical evidence against the null hypothesis
```

$hypothesis = false, confidenceinterval = 3.72513466617362$
$..4.21777770711276, distribution = ChiSquare(499), pvalue$
$= 0.000050154, statistic = 637.520051960458$

Test rozdílu dvou středních hodnot z náhodných výběrů z normálních rozdělení.

```
> NPX3:=RandomVariable(Normal(-4,3));
```

$$NPX3 := \_R10$$

```
> NPX4:=RandomVariable(Normal(1,2.5));
```

$$NPX4 := \_R11$$

```
> Data3:=Sample(NPX3,120);
```

$$Data3 := \begin{bmatrix} 1 .. 120\ Vector_{row} \\ Data\ Type: float_8 \\ Storage: rectangular \\ Order: Fortran\_order \end{bmatrix}$$

```
> Data4:=Sample(NPX4,180);
```

$$Data4 := \begin{bmatrix} 1 .. 180\ Vector_{row} \\ Data\ Type: float_8 \\ Storage: rectangular \\ Order: Fortran\_order \end{bmatrix}$$

Nejprve testujeme hypotézu na rovnost rozptylů:

```
> TwoSampleFTest(Data3,Data4,1,confidence=0.95);
F-Ratio Test on Two Samples
---------------------------
Null Hypothesis:
Sample drawn from populations with ratio of variances equal to 1
Alt. Hypothesis:
Sample drawn from population with ratio of variances not equal to 1

Sample sizes:            120, 180
Sample variances:        8.65697, 6.36674
Ratio of variances:      1.35972
Distribution:            FRatio(119,179)
Computed statistic:      1.35972
Computed pvalue:         0.0627989
Confidence interval:     .983736918003741 .. 1.90027107050859
                         (ratio of population variances)

Result: [Accepted]
There is no statistical evidence against the null hypothesis
```

$$hypothesis = true, confidenceinterval = 0.983736918003741 ..1.90027107050859, distribution = FRatio(119, 179), pvalue = 0.062798886, statistic = 1.35971833686992$$

Nyní testujeme hypotézu, že rozdíl středních hodnot je -4,5.

```
> TwoSampleTTest(Data3,Data4,-4.5,confidence=0.95,equalvariances=false);
Standard T-Test on Two Samples (Unequal Variances)
--------------------------------------------------
Null Hypothesis:
Sample drawn from populations with difference of means equal to -4.5
Alt. Hypothesis:
Sample drawn from population with difference of means not equal to
-4.5
```

```
Sample sizes:            120, 180
Sample means:            -3.47294, 0.737471
Sample standard devs.:   2.94227, 2.52324
Difference in means:     -4.21041
Distribution:            StudentT(227.879266553893)
Computed statistic:      0.883202
Computed pvalue:         0.378058
Confidence interval:     -4.85649132519817 .. -3.56432183080183
                         (difference of population means)

Result: [Accepted]
There is no statistical evidence against the null hypothesis
```

$$hypothesis = true, confidenceinterval = -4.85649132519817 .. -3.56432183080183, distribution = StudentT(227.879266553893), pvalue = 0.3780580304, statistic = 0.883201795643943$$

```
>
```