Pomůcka pro cvičení: 3. semestr Bc studia

Testy dobré shody - testování toho, zda daný výběr pochází z rozdělení určitého typu

Testy dobré shody

ChiSquareSuitableModelTest

balíček: Statistics

Pro účely, které se týkají našich výpočtů, vystačíme z balíčku **Statistics**s částí nazvanou **Tests**.
Při řešení problému, zda je náhodná proměnná vybrána z nějakého 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 chceme ověřit, že jde o výběr z normálního rozdělení, použijeme **ShapiroWilkWTest**.
Př. 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 \bar{x}, s, histogram četností a typ rozdělení doby mezi poruchami stroje.
Při řešení budeme postupovat následovně: sestrojíme nejprve histogram a graf hustoty pravděpodobnosti a pokusíme se odhadnout z jakého typu rozdělení je daný vzorek.

```
> with(Statistics):
>
X:=Array([63,278,4,323,415,100,529,272,46,188,156,15,35,275,140,1
89,310,117,236,124,184,561,73,176,17,176,22,169,387,385,786,229,2
03,427,293,672,69,466,92,174,822,8,178,196,991,134,130,286,177,23
]);
```

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

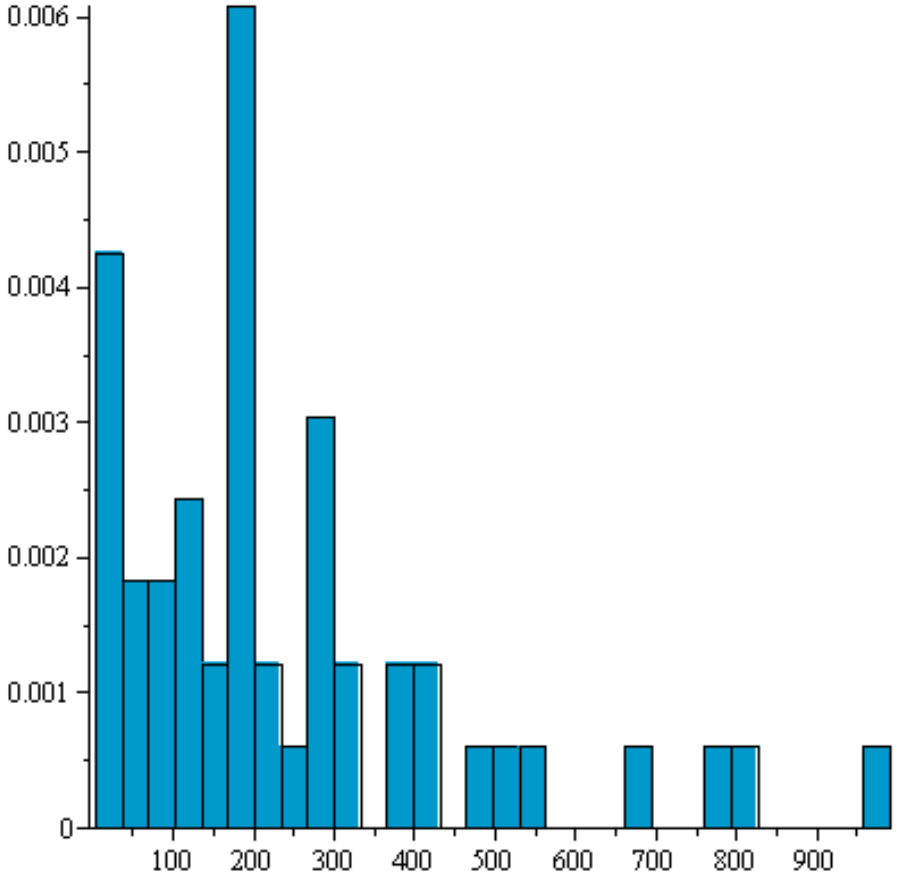
Při použití procedury **infolevel[Statistics]:=1** nám Maple poskytne i podrobný výpis informací vztahující se k danému výpočtu.

```
> infolevel[Statistics]:=1;
```

$$infolevel_{Statistics} := 1$$

```
> Histogram(X);
Histogram Type:  default
```

```
Data Range:       4. .. 991.
Bin Width:        32.90000000
Number of Bins:   30
Frequency Scale: relative
```



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

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

$$X1 := 246.4200000$$

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

$$S := 220.020879343163$$

```
> 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.

ShapiroWilkWTest

balíček: Statistics

Funkce **ShapiroWilkWTest** spočítá W-test Shapira a Wilka aplikovaný na datový soubor **X**. Tento test je založen na porovnání empirické distribuční funkce s teoretickou distribuční funkcí. Tento test se pokouší určit, jak blízko je daný vzorek normálnímu rozdělení. První parametr **X**je datový vzorek, který má být analyzován.

Př. Při přijímacích zkouškách z matematiky byl uchazečům předložen test. Maximální počet bodů v testu byl 60. Skupina 64 uchazečů
dosáhla těchto bodů:
46 24 33 17 42 32 37 26 37 43 36

37 42 42 41 27 15 28 24 15 22 34
33 29 19 33 26 35 27 32 31 16 24
20 40 24 23 30 40 15 12 37 36 30
31 39 35 35 27 24 24 23 30 42 39
27 23 34 36 37 41 42 26 33

Určete x, s a histogram četností pro dosažený počet bodů. Testujte na 5%-ní hladině významnosti hypotézu, že datový soubor je realizovanou hodnotou náhodného výběru z normálního rozdělení.

```
> restart;
> with(Statistics):
>
X:=Array([46,24,33,17,42,32,37,26,37,43,36,37,42,42,41,27,15,28,2
4,15,22,34,33,29,19,33,26,35,27,32,31,16,24,20,40,24,23,30,40,15,
12,37,36,30,31,39,35,35,27,24,24,23,30,42,39,27,23,34,36,37,41,42
,26,33]);
```

$$X := \begin{bmatrix} 1 .. 64\ Array \\ Data\ Type: anything \\ Storage: rectangular \\ Order: Fortran\_order \end{bmatrix}$$

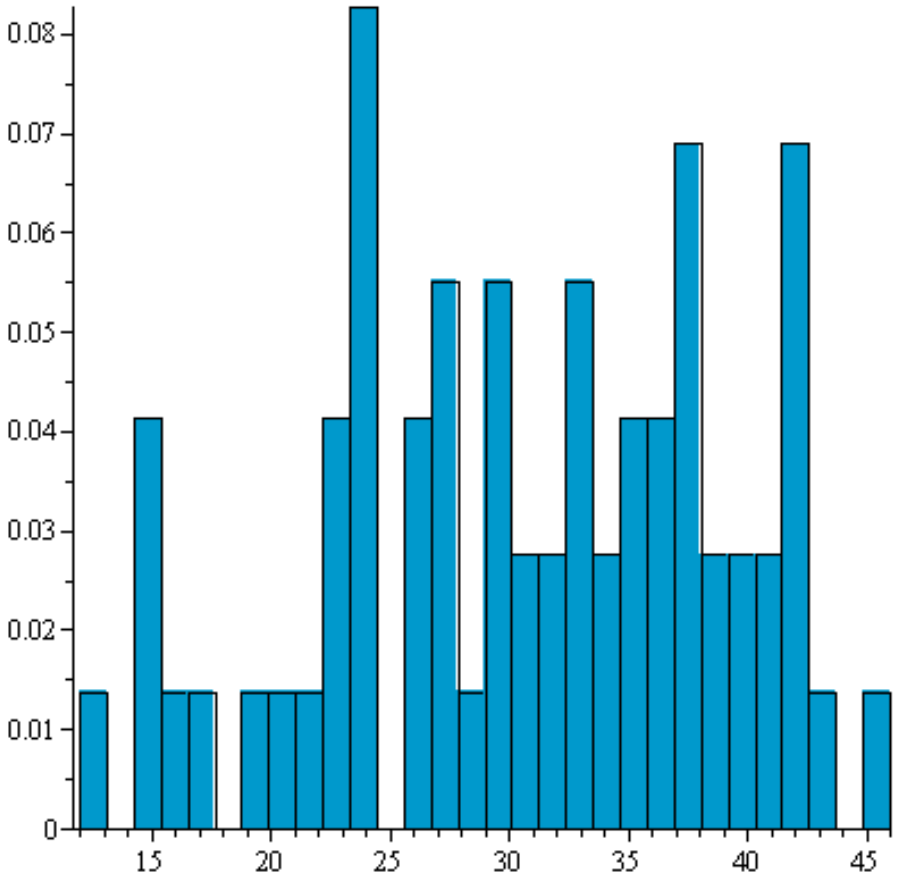
Protože je daný soubor příliš rozsáhlý, provedeme rozdělení do tříd. Pro rozdělení vzorku do tříd je možné použít příkaz **FrequencyTable**, kde si lze zvolit počet tříd.
Výsledná tabulka pak obsahuje počet prvků v jednotlivých třídách, procentulelní zastoupení prvků v jednotlivých třídách, kumulativní četnosti prvků v daných třídách a jejich procentuelní část.

```
> FrequencyTable(X,bins=8);
```

$$\begin{bmatrix} 12..16.25000000 & 5. & 7.812500000 & 5. & 7.812500000 \\ 16.25000000..20.50000000 & 3. & 4.687500000 & 8. & 12.50000000 \\ 20.50000000..24.75000000 & 10. & 15.62500000 & 18. & 28.12500000 \\ 24.75000000..29. & 8. & 12.50000000 & 26. & 40.62500000 \\ 29..33.25000000 & 12. & 18.75000000 & 38. & 59.37500000 \\ 33.25000000..37.50000000 & 13. & 20.31250000 & 51. & 79.68750000 \\ 37.50000000..41.75000000 & 6. & 9.375000000 & 57. & 89.06250000 \\ 41.75000000..46. & 7. & 10.93750000 & 64. & 100.0000000 \end{bmatrix}$$

Pro vykreslení histogramu užijeme příkaz **Histogram**. Pokud nezadáme další požadavky, získáme následující obrázek.

```
> Histogram(X);
```



Ze získaného histogramu není na první pohled zřejmé, že vzorek má být vybrán z normálního rozdělení.
Protože máme ověřit, že jde o výběr z normálního rozdělení, lze v programu Maple použít jeden z testů normality, a to **ShapiroWilkTest.** Při použití procedury **infolevel[Statistics]:=1** nám Maple kromě toho, zda byla nulová hypotéza přijata, tj. v našem případě, že šlo o výběr z normálního rozdělení, poskytne i podrobný výpis informací vztahující se k danému výpočtu.