SEMINÁŘ VÝPOČETNÍ STATISTIKY C4

2008-04-07

GRAFICKÉ PROCEDURY:

Možnosti:

- ✓ Procedura univariate – boxplot, stemplot…
- ✓ Grafické výstupy v Insight

Příklad:

- ✓ Vstupní údaje:

	Int			
40	Cena			
1	33.0	21	28.5	
2	34.0	22	28.0	
3	30.5	23	27.0	
4	28.5	24	33.0	
5	33.0	25	32.0	
6	32.0	26	30.5	
7	30.0	27	31.0	
8	29.0	28	28.5	
9	29.0	29	28.0	
10	30.0	30	29.0	
11	28.5	31	30.0	
12	29.0	32	33.0	
13	28.0	33	32.5	
14	28.0	34	30.0	
15	33.0	35	33.0	
16	32.0	36	30.0	
17	30.0	37	28.0	
18	30.5	38	29.5	
19	31.0	39	30.0	
20	32.0	40	31.0	

- ✓ Proměnná – kvantitativní, spojitá
- ✓ **Procedura freq:**

```
proc freq data=svs;
tables cena;
run;
```

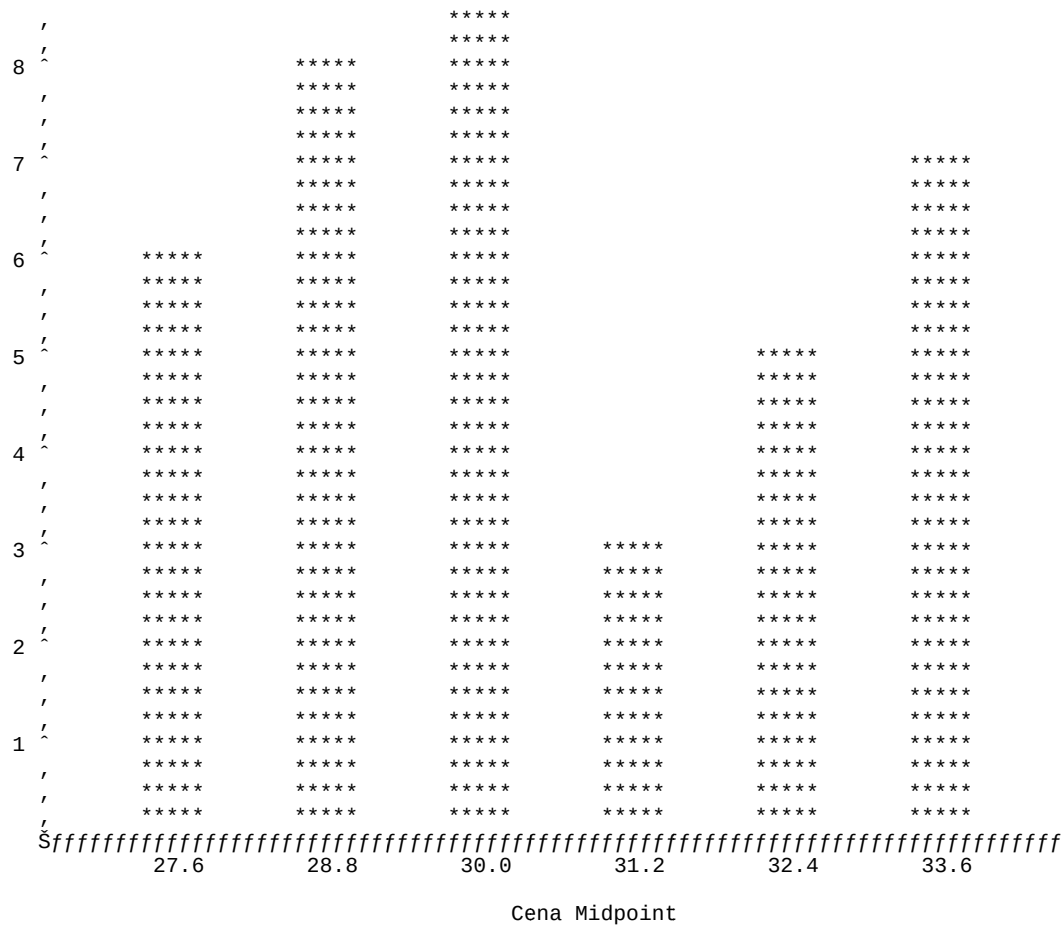
- ✓ Proložení Gaussovou křivkou u spojité veličiny – pouze při velkém množství měření a velkém rozptylu, za velký lze považovat soubor, který má více než 30 měření.
- ✓ Analyzování – procedura means nebo univariate, ale lze jiné grafické výstupy, např. procedura chart
- ✓ **Procedura chart:**
  - Lze histogram nebo stemplot
  - **Graf** může být vertikální i horizontální:
    - Vertikální:
      - *Procedura:*

```
proc chart data=svs;
vbar cena;
run;
```

      - *Výstup:*

```
Frequency

 11                                *****
                                   *****
   ,                               *****
   ,                               *****
   ,                               *****
 10 ^                              *****
   ,                               *****
   ,                               *****
   ,                               *****
  9 ^                              *****
   ,                               *****
```



- Horizontální:
  - *Procedura:*

```
proc chart data=svs;
hbar cena;
run;
```

  - *Výstup:*

```
 Cena                                                          Cum.                Cum.
Midpoint                                              Freq    Freq   Percent     Percent

  27.6   ******************************                6       6     15.00       15.00

  28.8   ****************************************      8      14     20.00       35.00

  30.0   ***************************************************  11      25     27.50       62.50

  31.2   ***************                               3      28      7.50       70.00

  32.4   ************************                      5      33     12.50       82.50

  33.6   ********************************              7      40     17.50      100.00

         1    2    3    4    5    6    7    8    9    10   11

                              Frequency
```

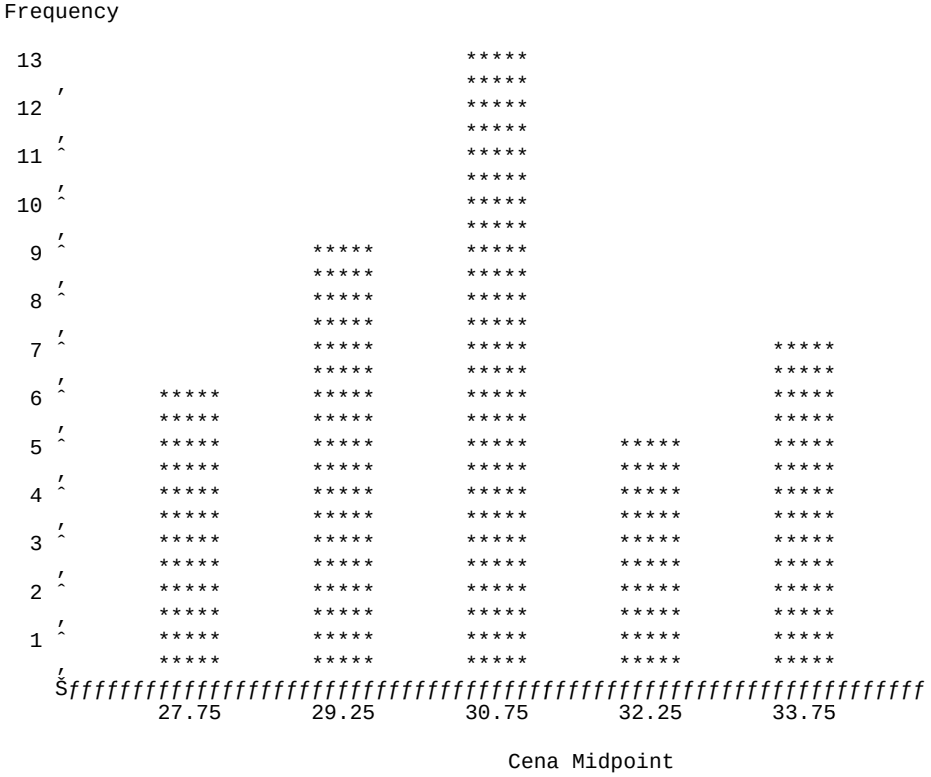
- **Modifikace:**
  - Zobrazení pouze grafu, bez statistik:

```
proc chart data=svs;
hbar cena/nostat;
run;
```

  - Pět intervalů – ne 6 podle Sturgesse:
    - *Procedura:*

```
proc chart data=svs;
vbar cena/levels=5;
run;
```

- *Výstup:*



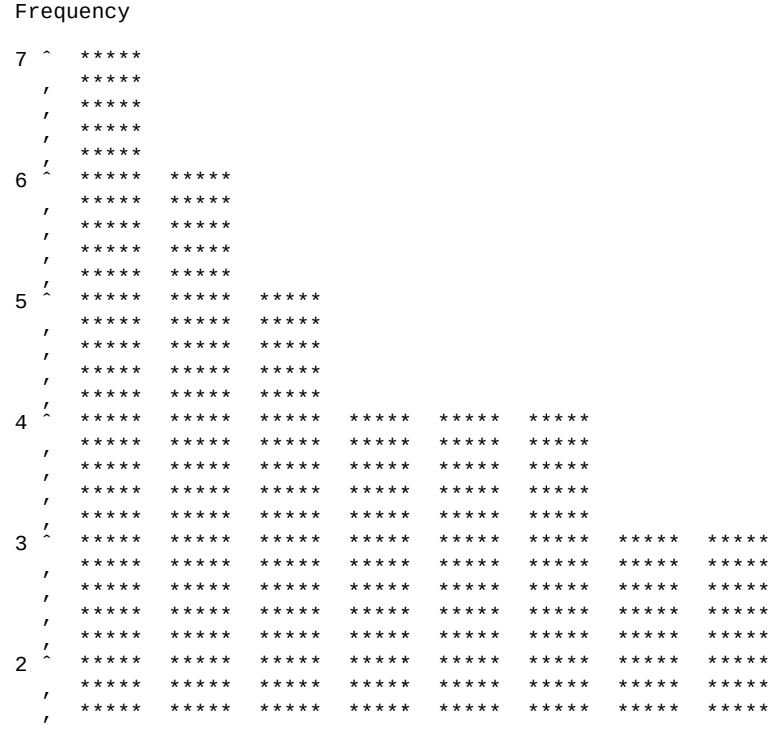
- Zobrazení skutečných četností ve 12 cenových relacích, tedy ve 12 intervalech

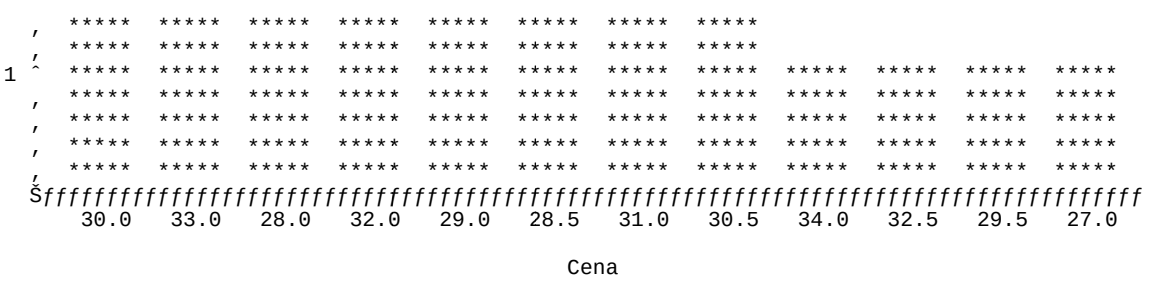
```
proc chart data=svs;
vbar cena/discrete;
run;
```

- Změna tendence:
  - *Příkazy* – tendence klesající = descending, stoupající = ascending
  - *Procedura:*

```
proc chart data=svs;
vbar cena/discrete descending;
run;
```

  - *Výstup:*



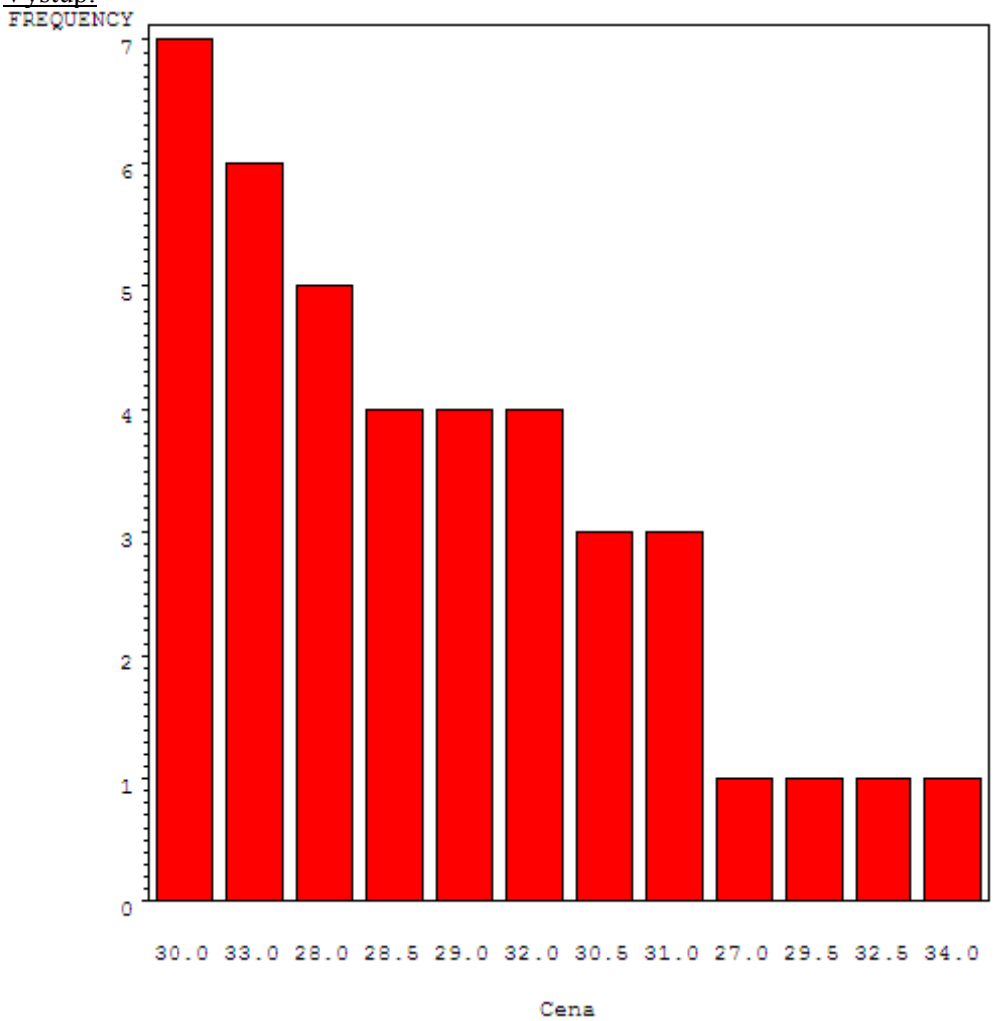


✓ Grafické – **procedura gchart:**

- **Obecně:**
  - Procedura:

```
proc gchart data=svs;
vbar cena/discrete descending;
run;
```

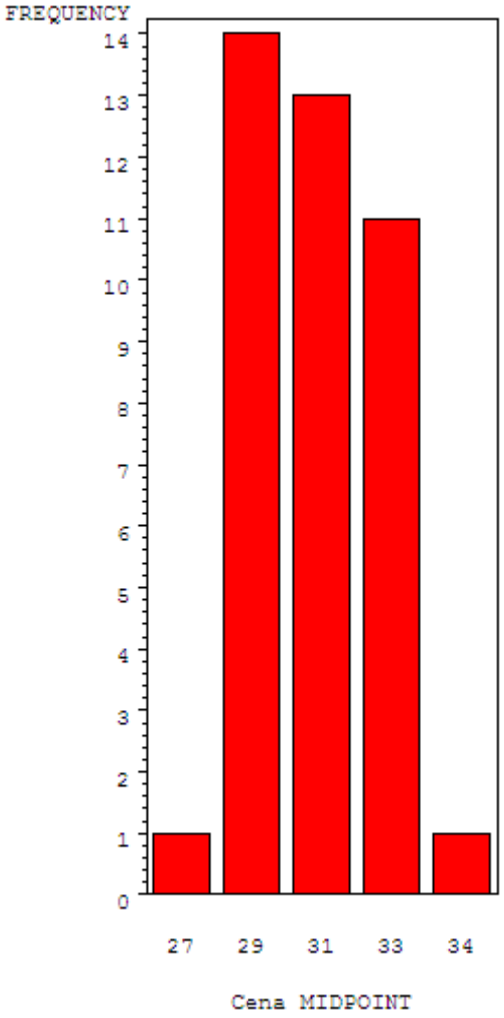
  - Výstup:



- **Zobrazení středů:**
  - Procedura:

```
proc gchart data=svs;
vbar cena/midpoints=27 29 31 33 34;
run;
```

  - Výstup:



TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ:

Příklad:

- ✓ Dva internetové obchody A a B:
- ✓ **Datový soubor** – píšeme do jednoho sloupce:
  - **Ne:**

WORK.A

12	A	B
1	500	485
2	498	490
3	485	500
4	500	510
5	505	520
6	510	475
7	485	.
8	495	.
9	500	.
10	498	.
11	495	.
12	505	.

  - **Ano:**

WORK.A

	Int	Nom
18	A	B
1	500	A
2	498	A
3	485	A
4	500	A
5	505	A
6	510	A
7	485	A
8	495	A
9	500	A
10	498	A
11	495	A
12	505	A
13	485	B
14	490	B
15	500	B
16	510	B
17	520	B
18	475	B

- ✓ Existuje mezi obchody statisticky významný rozdíl?
- ✓ Může být tvrzení výrobce, že máme prodávat za 500: $\mu_0 = 500$
- ✓ Budeme testovat, zda můžeme testovat soubory mezi sebou, zda existují významné rozdíly mezi obchody – **testování hypotéz:**
  - **Test:**
    - Párový test ne, jelikož předpokládá závislé výběry, data by musela být nezměněná, ale měřená v jiném čase, měření by musel být stejný počet
    - F-test – ne
    - Zajímají nás střední hodnoty, můžeme je popsat průměry => dvouvýběrový t-test
  - **Hypotéza:**
    - Neexistuje statisticky významný rozdíl mezi průměry souborů $H_0 : \mu_1 = \mu_2$, alternativní hypotéza: $H_0 : \mu_1 \neq \mu_2$
- ✓ **Procedura ttest:**
  - **Procedura:**

```
proc ttest data=svs;
class prodejna;
var cena;
run;
```

  - **Výstup:**

The TTEST Procedure

Statistics

Variable	prodejna	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
cena	A	12	493.27	498	502.73	5.2753	7.4468	12.644	2.1497
cena	B	6	479.21	496.67	514.12	10.383	16.633	40.795	6.7905
cena	Diff (1-2)		-10.5	1.3333	13.164	8.3129	11.162	16.987	5.5808

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
cena	Pooled	Equal	16	0.24	0.8142
cena	Satterthwaite	Unequal	6.02	0.19	0.8577

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
cena	Folded F	5	11	4.99	0.0250

- **Závěry:**
  - Předpokládáme hladinu významnosti 95%, Pr>F – hypotézu musíme zamítnout
  - Pokud $p<\alpha \Rightarrow H_1$, pokud $p>\alpha \Rightarrow H_0$
  - U t-testu je p hodnota větší než je hladina významnosti, a nulovou hypotézu musíme potvrdit a říkáme, že mezi průměry neexistuje statisticky významný rozdíl
  - Obchod B je sice průměrně levnější, směrodatná odchylka však vychází 16.7, takže v tomto obchodě je velká pravděpodobnost nakoupit za větší cenu. Chceme-li příznivou cenu, volíme A, chceme-li zariskovat, volíme B.
  - Toto funguje, pouze pokud by oba soubory měly normální rozdělení – <u>ověření normality</u>:
    - *Procedura:*

```
proc univariate data=svs normal;
class prodejna;
var cena;
run;
```

    - *Výstup:*

The UNIVARIATE Procedure
Variable: cena
prodejna = A

Moments

N	12	Sum Weights	12
Mean	498	Sum Observations	5976
Std Deviation	7.44678088	Variance	55.4545455
Skewness	-0.5309809	Kurtosis	0.21298396
Uncorrected SS	2976658	Corrected SS	610
Coeff Variation	1.49533753	Std Error Mean	2.14970047

Basic Statistical Measures

Location		Variability	
Mean	498.0000	Std Deviation	7.44678
Median	499.0000	Variance	55.45455
Mode	500.0000	Range	25.00000
		Interquartile Range	7.50000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	231.6602	Pr > |t|	<.0001
Sign	M	6	Pr >= |M|	0.0005
Signed Rank	S	39	Pr >= |S|	0.0005

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.922566	Pr < W	0.3079
Kolmogorov-Smirnov	D	0.176859	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.071122	Pr > W-Sq	0.2495
Anderson-Darling	A-Sq	0.450757	Pr > A-Sq	0.2324

Quantiles (Definition 5)

Quantile	Estimate
100% Max	510.0
99%	510.0

The UNIVARIATE Procedure
Variable: cena
prodejna = B

Moments

N	6	Sum Weights	6
Mean	496.666667	Sum Observations	2980
Std Deviation	16.6332999	Variance	276.666667
Skewness	0.19919397	Kurtosis	-1.0462331
Uncorrected SS	1481450	Corrected SS	1383.33333
Coeff Variation	3.34898656	Std Error Mean	6.79051626

Basic Statistical Measures

Location		Variability	
Mean	496.6667	Std Deviation	16.63330
Median	495.0000	Variance	276.66667
Mode	.	Range	45.00000
		Interquartile Range	25.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	73.14122	Pr > |t|	<.0001
Sign	M	3	Pr >= |M|	0.0313
Signed Rank	S	10.5	Pr >= |S|	0.0313

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.980781	Pr < W	0.9554
Kolmogorov-Smirnov	D	0.155717	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.020885	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.147194	Pr > A-Sq	>0.2500

Quantiles (Definition 5)

Quantile	Estimate
100% Max	520
99%	520
95%	520
90%	520
75% Q3	510
50% Median	495
25% Q1	485

- *Závěry:*
  - Shapiro-Wilk test říká soubor má normální rozdělení
  - Šikmost a Špičatost
  - Jeden test na normalitu nestačí, čím méně měření, tím jsou výsledky pochybnější
  - Nutno udělat test, který normalitu nevyžaduje => Wilcoxonův.

✓ **Wilcoxonův test** – neparametrická analogie parametrického t-testu:

- **Procedura:**

```
proc npar1way data=svs wilcoxon;
class prodejna;
var cena;
run;
```

- **Hypotézy:**
  - Test je neparametrický, nelze stanovit hypotézy týkající se průměrů
  - Neparametrický – soubor má velmi málo měření, o rozdělení statistického souboru nevíme nic, o rozdělení souboru víme, že není normální
  - Nulová říká, že výběry pocházejí z téhož rozdělení, máme tedy jeden základní soubor a v něm dva výběrové soubory s průměry a rozptyly.
  - Alternativa říká, že výběry se statisticky významně liší svou polohou, což znamená, že obecně máme soubor takový, že má průměr a rozptyl, druhý má jiný průměr a rozptyl, průměry jsou charakteristiky polohy, takže existuje i rozdíl mezi průměry.
- **Výstup:**

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable cena
Classified by Variable prodejna

prodejna	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
A	12	116.0	114.0	10.577445	9.666667
B	6	55.0	57.0	10.577445	9.166667

```
Average scores were used for ties.

    Wilcoxon Two-Sample Test

  Statistic             55.0000

  Normal Approximation
  Z                     -0.1418
  One-Sided Pr <  Z      0.4436
  Two-Sided Pr > |Z|     0.8872

  t Approximation
  One-Sided Pr <  Z      0.4444
  Two-Sided Pr > |Z|     0.8889

Z includes a continuity correction of 0.5.

        Kruskal-Wallis Test

  Chi-Square             0.0358
  DF                          1
  Pr > Chi-Square        0.8500
```

- **Závěry:**
  - U parametrických, když je testovací kritérium větší než tabulková hodnota, přijímáme nulovou hypotézu.
  - Stačí se podívat na p hodnotu
  - P hodnota 0,88, což je větší než 0,05 a přijímáme H_0, soubory tedy pocházejí ze stejného rozdělení
  - Kruskal-Wallisův test – zbytečně silný