STATISTIKA 2

Statistika a statistické zpracování dat

Blok 4-1

Jiří Fišer

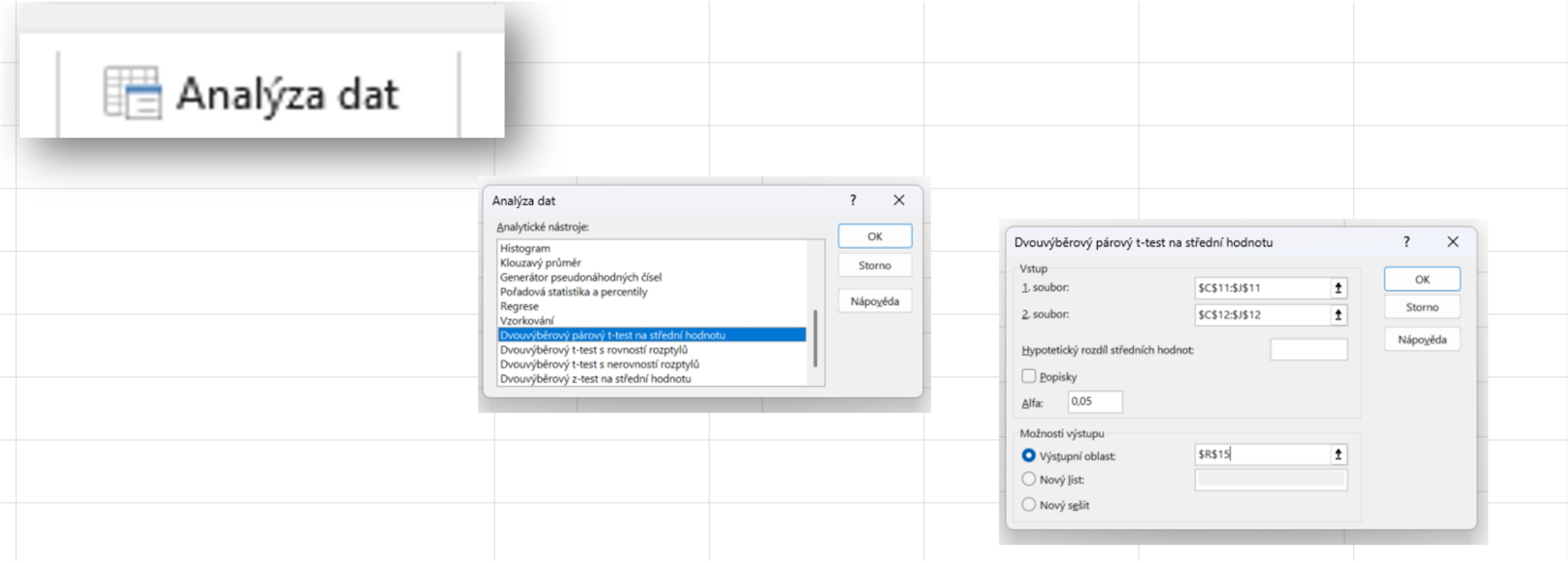
14. prosince 2024

MS Excel – Doplněk Analýza dat

8. V tabulce jsou uvedeny výsledky měření tlaku v pneumatikách před a po opravě. Použijte párový t-test k ověření, zda došlo ke statisticky významné změně tlaku po opravě na hladině významnosti $\alpha = 0{,}05$.

Před opravou	32,5	31,8	33,2	32,0	31,5	33,0	32,3	31,7
Po opravě	32,2	31,9	33,0	31,8	31,6	32,9	32,4	31,8

Data → Analýza dat → Dvouvýběrový párový t-test...



Dvouvýběrový párový t-test na střední hodnotu

	Soubor 1	*Soubor 2*
Stř. hodnota	32,25	32,2
Rozptyl	0,38	0,277143
Pozorování	8	8
Pears. korelace	0,968458	
Hyp. rozdíl stř. hodnot	0	
Rozdíl	7	
t Stat	0,83666	
P(T<=t) (1)	0,215208	
t krit (1)	1,894579	
P(T<=t) (2)	0,430416	
t krit (2)	2,364624	

Výběrové průměry

Výběrové rozptyly

$$T = \frac{\overline{d}\sqrt{n}}{s_d},$$

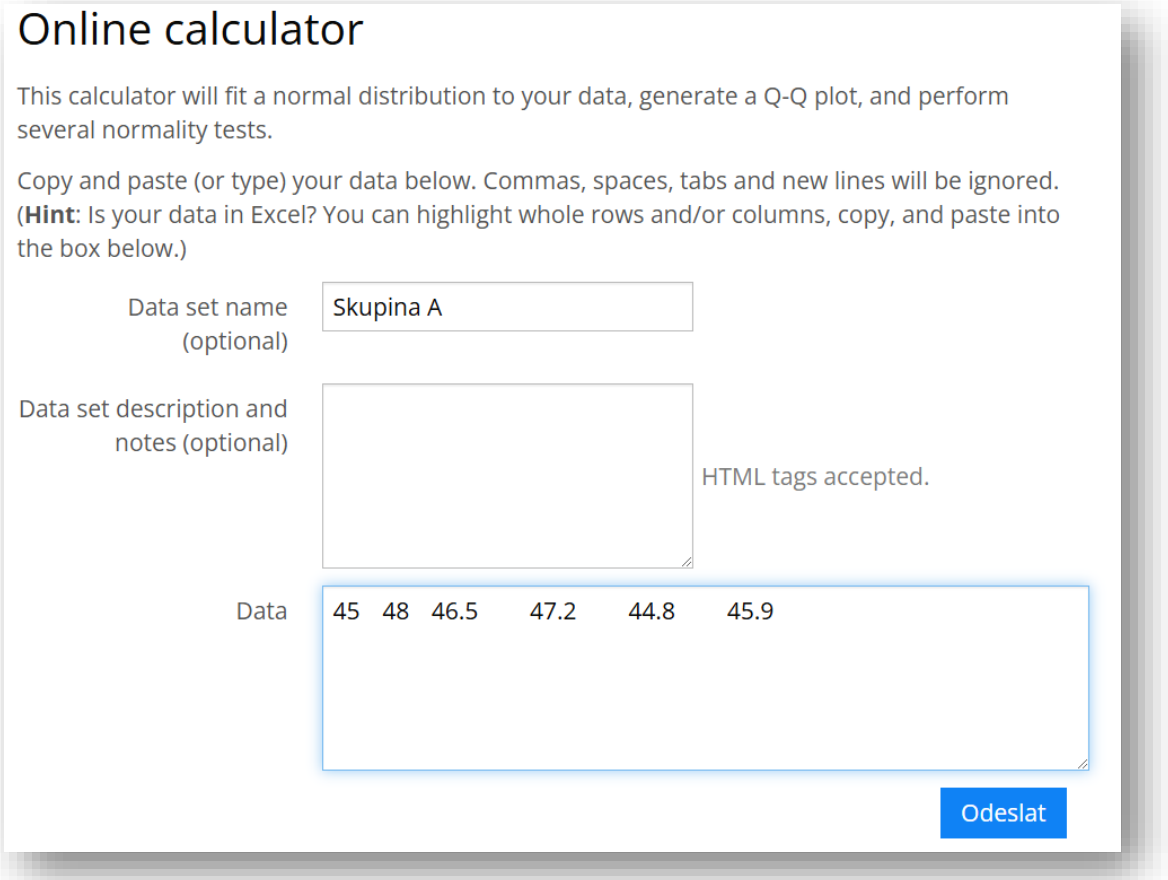
$t_{0,95}$ = T.INV(0,95;7)

$t_{0,975}$ = T.INV(0,975;7)

Test normality dat online

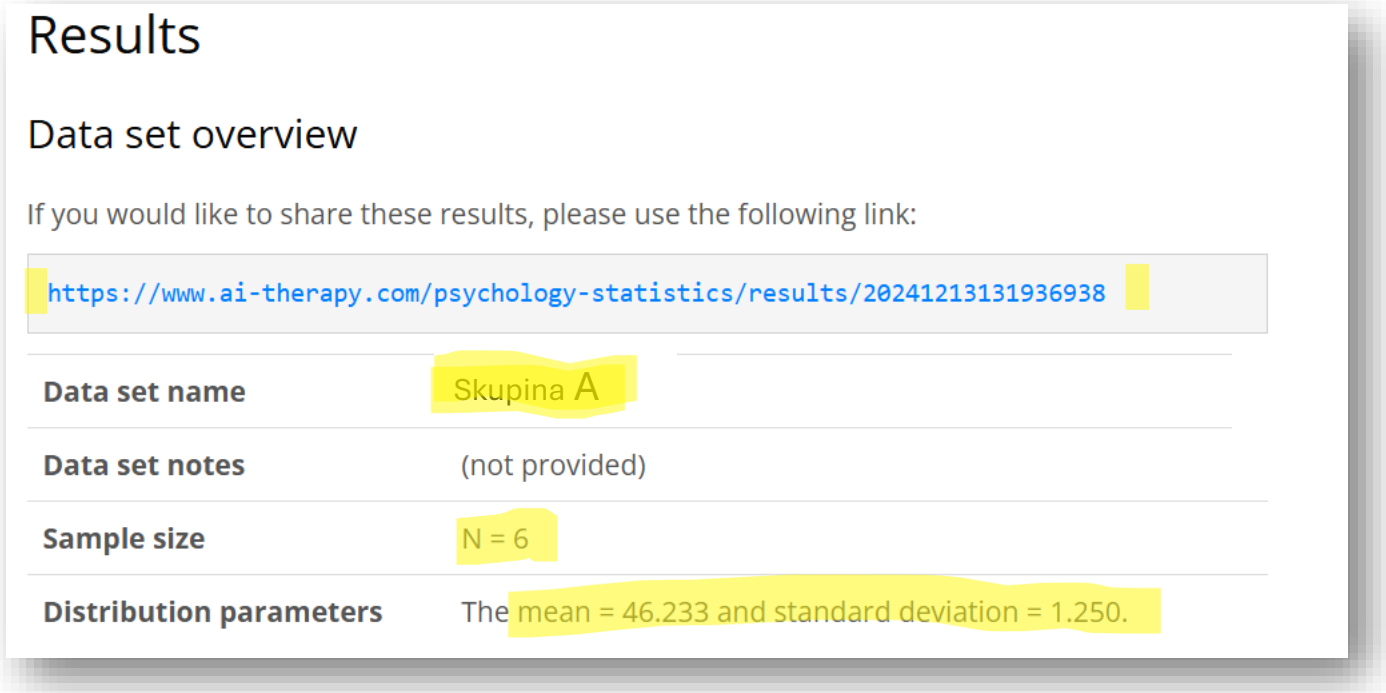
- https://www.ai-therapy.com/psychology-statistics/distributions/normal

- Otevřte stránku
- Zadejte název (volitelné)
- Zadejte data
- Pokud kopírujete z českého Excelu desetinná čísla, tak nejprve nahraďte desetinné čárky tečkami
- Odeslat



Test normality dat online

- Ve výsledcích najdete odkaz na tuto stránku s výsledky
- Základní údaje o datech



Results

Data set overview

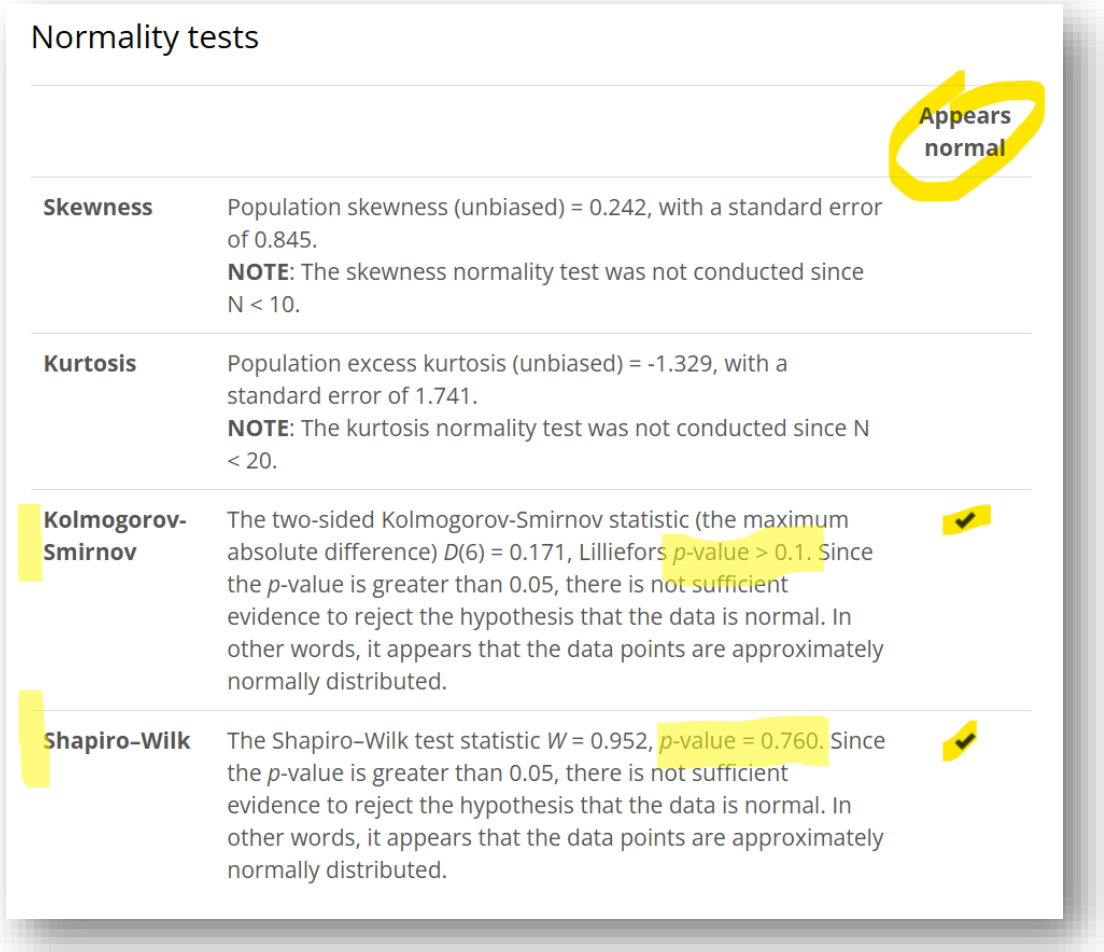
If you would like to share these results, please use the following link:

https://www.ai-therapy.com/psychology-statistics/results/20241213131936938

Data set name	Skupina A
Data set notes	(not provided)
Sample size	N = 6
Distribution parameters	The mean = 46.233 and standard deviation = 1.250.

Test normality dat online

- Vlastní test normality podložený údaji o šikmosti a špičatosti (pro dostatečný počet pozorování)
- Dva testy na normalitu dat (= nulová hypotéza)
- p-value > 0,05: nemůžeme zamítnout nulovou hypotézu (normalita „ověřena“)
- P-value <= 0.05: zamítáme (normalita vyvrácena)

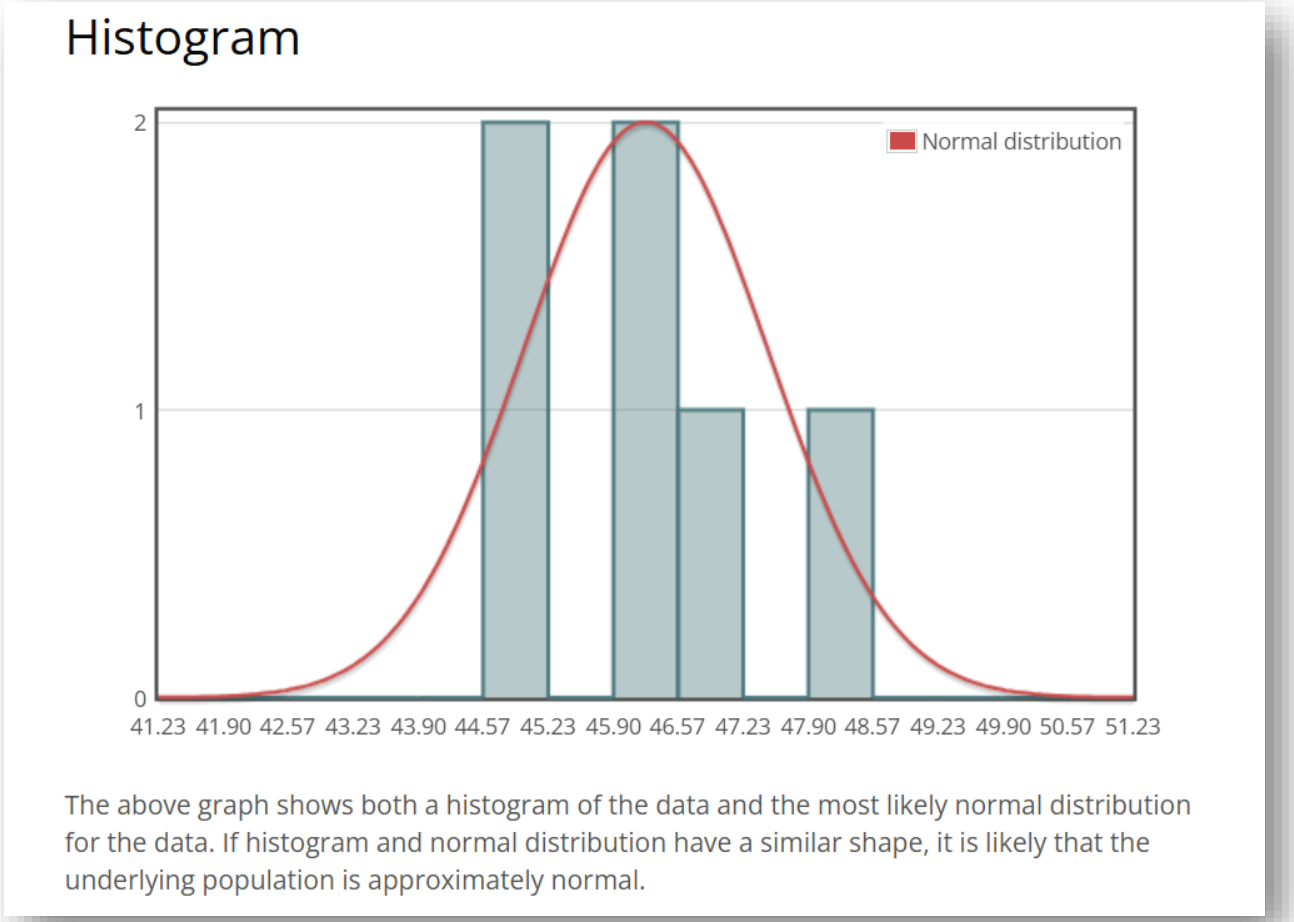


Normality tests

		Appears normal
Skewness	Population skewness (unbiased) = 0.242, with a standard error of 0.845. **NOTE**: The skewness normality test was not conducted since N < 10.	
Kurtosis	Population excess kurtosis (unbiased) = -1.329, with a standard error of 1.741. **NOTE**: The kurtosis normality test was not conducted since N < 20.	
Kolmogorov-Smirnov	The two-sided Kolmogorov-Smirnov statistic (the maximum absolute difference) $D(6) = 0.171$, Lilliefors p-value > 0.1. Since the p-value is greater than 0.05, there is not sufficient evidence to reject the hypothesis that the data is normal. In other words, it appears that the data points are approximately normally distributed.	✔
Shapiro–Wilk	The Shapiro–Wilk test statistic $W = 0.952$, p-value = 0.760. Since the p-value is greater than 0.05, there is not sufficient evidence to reject the hypothesis that the data is normal. In other words, it appears that the data points are approximately normally distributed.	✔

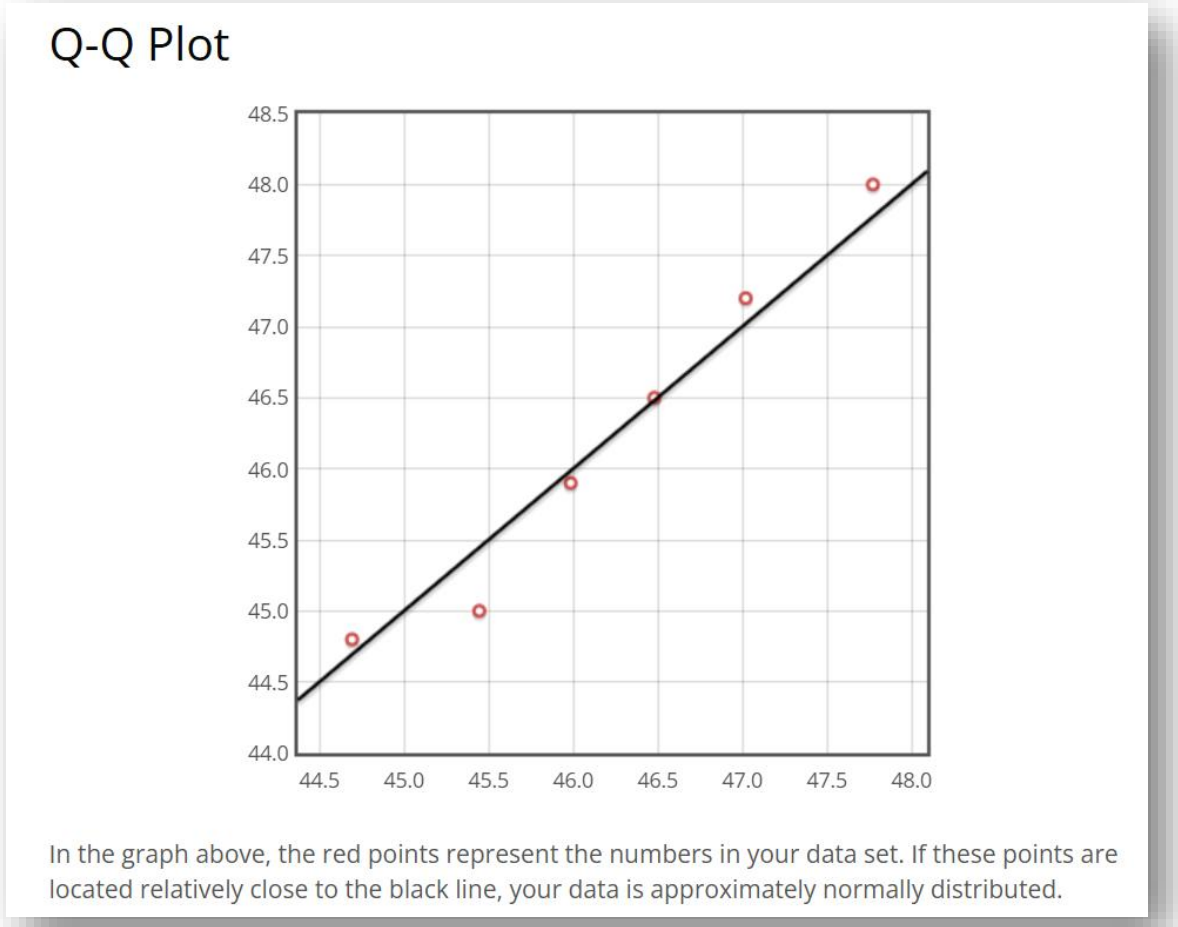
Test normality dat online

- Grafické znázornění dat pomocí histogramu
- Červená křivka znázorňuje hustotu normálního rozdělení s parametry odpovídajícími datům
- U malého rozsahu dat nelze očekávat, že sloupce histogramu budou dobře kopírovat zvonovitý tvar



Test normality dat online

- Kvantil-kvantil plot (Q-Q plot)
- Porovnávají se kvantily vypočítané z dat a kvantily odpovídajícího normálního rozdělení
- Při ideálním normálním rozložení dat by červené body ležely na černé úsečce (úhlopříčce)



Test normality dat online

- Test normality pro trochu rozsáhlejší data vygenerovaná pomocí doplňku Analýza dat v Excelu
- Pro jednodušší aplikaci zaokrouhlena na celé čísla
- Níže je odkaz na výsledky testu
- Všimněte si poznámky, že při větším počtu pozorování mohou být testy „přecitlivělé“ na i drobnou odchylku od normality (falešně negativní závěr testu)
- V takovém případě je doporučena vizuální kontrola histogramu a Q-Q plotu

https://www.ai-therapy.com/psychology-statistics/results/20241213215213897

Analýza rozptylu (ANOVA)

Úvodní příklad

Představte si, že pracujete v marketingovém oddělení firmy, která prodává tři různé druhy energetických nápojů: A, B a C. Chcete zjistit, zda existují rozdíly v průměrném prodeji těchto nápojů v různých regionech. Shromáždili jste data o týdenních prodejích v pěti regionech pro každý druh nápoje:

Nápoj	Region 1	Region 2	Region 3	Region 4	Region 5
A	100	110	95	105	102
B	98	85	88	90	92
C	120	115	130	125	118

Chcete zjistit, zda jsou rozdíly v průměrných prodejích mezi nápoji A, B a C statisticky významné, nebo zda jsou způsobeny náhodou.

Analýza rozptylu (ANOVA)

Formulace hypotéz

- **Nulová hypotéza** (H_0): Průměrné prodeje všech tří nápojů jsou stejné ($\mu_A = \mu_B = \mu_C$).
- **Alternativní hypotéza** (H_1): Existuje alespoň jeden pár nápojů, u kterého se průměrné prodeje liší.

Aplikace ANOVA

K prověření těchto hypotéz použijeme jednofaktorovou analýzu rozptylu (ANOVA), která nám umožní porovnat průměry více než dvou skupin současně.

Analýza rozptylu (ANOVA)

- **Co je to ANOVA?**
  - Statistická metoda pro testování rozdílů mezi průměry dvou nebo více skupin.
- **Co zkoumá:**
  - Zda je variabilita mezi skupinami větší než variabilita uvnitř skupin.
- **Závěr:**
  - Pokud je variabilita mezi skupinami výrazně větší, může to naznačovat, že skupiny pocházejí z různých populací.

Předpoklady ANOVA

- **Normalita**: Data v každé skupině jsou přibližně normálně rozložena.
- **Homogenita rozptylů**: Rozptyly v jednotlivých skupinách jsou stejné.
- **Nezávislost pozorování**: Data jsou nezávislá mezi a uvnitř skupin.

Nezávislost pozorování v ANOVA

- **Co to znamená?**
  - Výsledek jednoho měření nesmí ovlivnit jiná měření.

- **Mezi skupinami:**
  - Skupiny musí být nezávislé (např. studenti z různých škol spolu nesmí sdílet informace).

- **Uvnitř skupin:**
  - Jednotlivá měření (osoby) v jedné skupině nesmí vzájemně ovlivňovat výsledky (např. opisování).

- **Proč je důležitá?**
  - Porušení může vést k chybným nebo zkresleným závěrům.

- **Jak zajistit?**
  - Náhodné přiřazení účastníků do skupin.
  - Oddělený sběr dat bez interakce mezi skupinami.
  - Ověření nezávislosti už při návrhu experimentu.

Rozklad variability (ANOVA)

- **Variabilita mezi skupinami** (meziskupinová): Variabilita způsobená rozdíly mezi průměry skupin.
- **Variabilita uvnitř skupin** (vnitroskupinová): Variabilita způsobená rozdíly uvnitř jednotlivých skupin.

Variabilita se obvykle nějakým způsobem vyjadřuje pomocí součtu čtverců.

Matematicky lze **celkový součet čtverců** (SS) vyjádřit jako:

$$\mathrm{SS}_{\mathrm{celk}} = \mathrm{SS}_{\mathrm{mezi}} + \mathrm{SS}_{\mathrm{uvnitř}},$$

kde:

- $\mathrm{SS}_{\mathrm{celk}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{\mathrm{celk}})^2$,
- $\mathrm{SS}_{\mathrm{mezi}} = \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x}_{\mathrm{celk}})^2$,
- $\mathrm{SS}_{\mathrm{uvnitř}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2$,

kde k je počet skupin, n_i je počet pozorování v i-té skupině, x_{ij} je j-té pozorování v i-té skupině, $\overline{x}_i$ je průměr i-té skupiny a $\overline{x}_{\mathrm{celk}}$ je celkový průměr.

Jednofaktorová ANOVA

1. **Stanovení hypotéz**:
   - H_0: Všechny skupinové průměry jsou stejné ($\mu_1 = \mu_2 = \cdots = \mu_k$).
   - H_1: Alespoň jeden skupinový průměr se liší.

2. **Výpočet součtů čtverců** (SS):
   - Celkový součet čtverců (SS_{celk}).
   - Meziskupinový součet čtverců (SS_{mezi}).
   - Vnitroskupinový součet čtverců ($SS_{uvnitř}$).

Jednofaktorová ANOVA

3. **Výpočet stupňů volnosti** (df):
   - $\mathrm{df}_{\mathrm{mezi}} = k - 1$.
   - $\mathrm{df}_{\mathrm{uvnitř}} = N - k$, kde N je celkový počet pozorování.
   - $\mathrm{df}_{\mathrm{celk}} = N - 1$.
4. **Výpočet středních čtverců** (MS):
   - $\mathrm{MS}_{\mathrm{mezi}} = \frac{\mathrm{SS}_{\mathrm{mezi}}}{\mathrm{df}_{\mathrm{mezi}}}$.
   - $\mathrm{MS}_{\mathrm{uvnitř}} = \frac{\mathrm{SS}_{\mathrm{uvnitř}}}{\mathrm{df}_{\mathrm{uvnitř}}}$.
5. **Výpočet F-statistiky**:
$$F = \frac{\mathrm{MS}_{\mathrm{mezi}}}{\mathrm{MS}_{\mathrm{uvnitř}}}.$$
6. **Určení kritické hodnoty a rozhodnutí**:
   - Porovnáme vypočtenou hodnotu F s kritickou hodnotou z F-rozdělení pro zvolené hladiny významnosti α.
   - Pokud F překročí kritickou hodnotu, zamítáme H_0.

Jednofaktorová ANOVA – Řešený příklad

Příklad 10.1. Proveďte jednofaktorovou ANOVA na datech z úvodního příkladu a určete, zda existují statisticky významné rozdíly mezi průměrnými prodeji nápojů A, B a C na hladině významnosti $\alpha = 0{,}05$.

Nápoj	Region 1	Region 2	Region 3	Region 4	Region 5
A	100	110	95	105	102
B	98	85	88	90	92
C	120	115	130	125	118

Řešení: **Krok 1: Stanovení hypotéz**

- H_0: $\mu_A = \mu_B = \mu_C$.
- H_1: Ne všechny průměry jsou stejné.

ANOVA

Nápoj	Region 1	Region 2	Region 3	Region 4	Region 5
A	100	110	95	105	102
B	98	85	88	90	92
C	120	115	130	125	118

Krok 2: Výpočet průměrů

$$\overline{x}_A = \frac{100 + 110 + 95 + 105 + 102}{5} = 102{,}4,$$
$$\overline{x}_B = \frac{98 + 85 + 88 + 90 + 92}{5} = 90{,}6,$$
$$\overline{x}_C = \frac{120 + 115 + 130 + 125 + 118}{5} = 121{,}6.$$

Celkový průměr:

$$\overline{x}_{\text{celk}} = \frac{\sum_{i=1}^{3} \sum_{j=1}^{5} x_{ij}}{15} = \frac{512 + 453 + 608}{15} = \frac{1\,573}{15} = 104{,}867.$$

Nápoj	Region 1	Region 2	Region 3	Region 4	Region 5
A	100	110	95	105	102
B	98	85	88	90	92
C	120	115	130	125	118

Krok 2: Výpočet průměrů

$$\overline{x}_A = \frac{100+110+95+105+102}{5} = 102{,}4,$$
$$\overline{x}_B = \frac{98+85+88+90+92}{5} = 90{,}6,$$
$$\overline{x}_C = \frac{120+115+130+125+118}{5} = 121{,}6.$$

Celkový průměr:

$$\overline{x}_{\text{celk}} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij}}{15} = \frac{512+453+608}{15} = \frac{1\,573}{15} = 104{,}867.$$

Krok 3: Výpočet součtů čtverců

SS mezi:

$$\text{SS}_{\text{mezi}} = \sum_{i=1}^{3} n_i(\overline{x}_i - \overline{x}_{\text{celk}})^2 = 5(102{,}4-104{,}867)^2 + 5(90{,}6-104{,}867)^2 + 5(121{,}6-104{,}867)^2.$$

$$\text{SS}_{\text{mezi}} = 5 \times (6{,}083 + 203{,}577 + 280{,}005) = 5 \times 489{,}665 = 2\,448{,}325.$$

ANOVA

Nápoj	Region 1	Region 2	Region 3	Region 4	Region 5
A	100	110	95	105	102
B	98	85	88	90	92
C	120	115	130	125	118

Krok 2: Výpočet průměrů

$$\overline{x}_A = \frac{100+110+95+105+102}{5} = 102{,}4,$$
$$\overline{x}_B = \frac{98+85+88+90+92}{5} = 90{,}6,$$
$$\overline{x}_C = \frac{120+115+130+125+118}{5} = 121{,}6.$$

Celkový průměr:

$$\overline{x}_{\text{celk}} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij}}{15} = \frac{512+453+608}{15} = \frac{1\,573}{15} = 104{,}867.$$

SS uvnitř:

Pro každou skupinu spočítáme součet čtverců odchylek od skupinového průměru.

Pro nápoj A:

$$\text{SS}_A = (100-102{,}4)^2 + (110-102{,}4)^2 + (95-102{,}4)^2 + (105-102{,}4)^2 + (102-102{,}4)^2.$$

Pro nápoj B:

$$\text{SS}_B = (98-90{,}6)^2 + (85-90{,}6)^2 + (88-90{,}6)^2 + (90-90{,}6)^2 + (92-90{,}6)^2.$$

Pro nápoj C:

$$\text{SS}_C = (120-121{,}6)^2 + (115-121{,}6)^2 + (130-121{,}6)^2 + (125-121{,}6)^2 + (118-121{,}6)^2.$$

Celkový $\text{SS}_{\text{uvnitř}}$:

$$\text{SS}_{\text{uvnitř}} = \text{SS}_A + \text{SS}_B + \text{SS}_C = 125{,}2 + 95{,}2 + 141{,}2 = 361{,}6.$$

ANOVA

Nápoj	Region 1	Region 2	Region 3	Region 4	Region 5
A	100	110	95	105	102
B	98	85	88	90	92
C	120	115	130	125	118

Krok 2: Výpočet průměrů

$$\overline{x}_A = \frac{100+110+95+105+102}{5} = 102{,}4,$$
$$\overline{x}_B = \frac{98+85+88+90+92}{5} = 90{,}6,$$
$$\overline{x}_C = \frac{120+115+130+125+118}{5} = 121{,}6.$$

Celkový průměr:

$$\overline{x}_{\text{celk}} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij}}{15} = \frac{512+453+608}{15} = \frac{1\,573}{15} = 104{,}867.$$

SS uvnitř:

Pro každou skupinu spočítáme součet čtverců odchylek od skupinového průměru.

Pro nápoj A:

$$\text{SS}_A = (100-102{,}4)^2 + (110-102{,}4)^2 + (95-102{,}4)^2 + (105-102{,}4)^2 + (102-102{,}4)^2.$$

Pro nápoj B:

$$\text{SS}_B = (98-90{,}6)^2 + (85-90{,}6)^2 + (88-90{,}6)^2 + (90-90{,}6)^2 + (92-90{,}6)^2.$$

Pro nápoj C:

$$\text{SS}_C = (120-121{,}6)^2 + (115-121{,}6)^2 + (130-121{,}6)^2 + (125-121{,}6)^2 + (118-121{,}6)^2.$$

Celkový $\text{SS}_{\text{uvnitř}}$:

$$\text{SS}_{\text{uvnitř}} = \text{SS}_A + \text{SS}_B + \text{SS}_C = 125{,}2 + 95{,}2 + 141{,}2 = 361{,}6.$$

ANOVA

Krok 4: Výpočet stupňů volnosti

$$\mathrm{df}_{\mathrm{mezi}} = k - 1 = 3 - 1 = 2,$$
$$\mathrm{df}_{\mathrm{uvnitř}} = N - k = 15 - 3 = 12.$$

Krok 5: Výpočet středních čtverců

$$\mathrm{MS}_{\mathrm{mezi}} = \frac{\mathrm{SS}_{\mathrm{mezi}}}{\mathrm{df}_{\mathrm{mezi}}} = \frac{2\,448{,}325}{2} = 1\,224{,}1625,$$
$$\mathrm{MS}_{\mathrm{uvnitř}} = \frac{\mathrm{SS}_{\mathrm{uvnitř}}}{\mathrm{df}_{\mathrm{uvnitř}}} = \frac{361{,}6}{12} = 30{,}1333.$$

ANOVA

Krok 6: Výpočet F-statistiky

$$F = \frac{\mathrm{MS}_{\text{mezi}}}{\mathrm{MS}_{\text{uvnitř}}} = \frac{1\,224{,}1625}{30{,}1333} \approx 40{,}619.$$

Krok 7: Určení kritické hodnoty a rozhodnutí

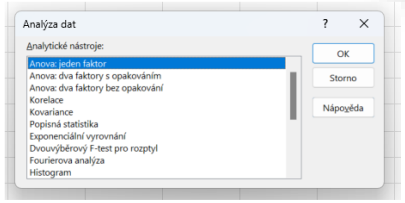
Kritická hodnota F_{krit} pro $\alpha = 0{,}05$, $\mathrm{df}_1 = 2$ a $\mathrm{df}_2 = 12$ je přibližně 3,8853 (lze zjistit z F-tabulek nebo pomocí Excelu (`F.INV(0,95;2;12)`).

Protože vypočtené $F \approx 40{,}619$ je větší než $F_{\text{krit}} = 3{,}8853$, zamítáme nulovou hypotézu H_0.

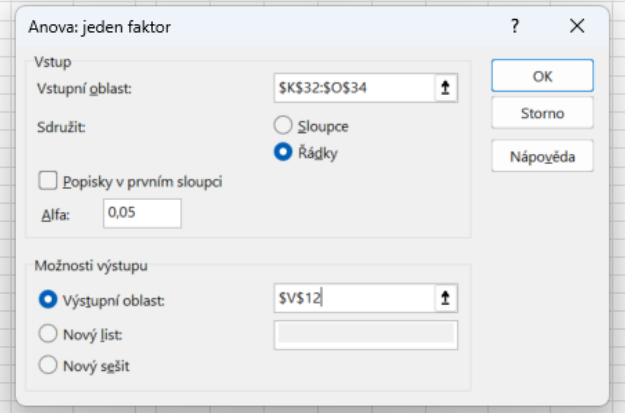
Závěr: Existují statisticky významné rozdíly mezi průměrnými prodeji nápojů A, B a C.

ANOVA v Excelu

Alternativně můžeme vyřešit tento příklad v Excelu. Můžeme použít postupné výpočty (jak jsou rozepsány výše), ale rychlejší je použít doplněk Analýza dat (pokud jej máme možnost nainstalovat). Spuštění, vložení dat a výstupy jsou na obrázcích



	Region 1	Region 2	Region 3	Region 4	Region 5
A	100	110	95	105	102
B	98	85	88	90	92
c	120	115	130	125	118



Anova: jeden faktor

Faktor

Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*
Řádek 1	5	512	102,4	31,3
Řádek 2	5	453	90,6	23,8
Řádek 3	5	608	121,6	35,3

ANOVA

Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
Mezi výběry	2448,133333	2	1224,066667	40,62168142	4,54339E-06	3,885293835
Všechny výběry	361,6	12	30,13333333			
Celkem	2809,733333	14				

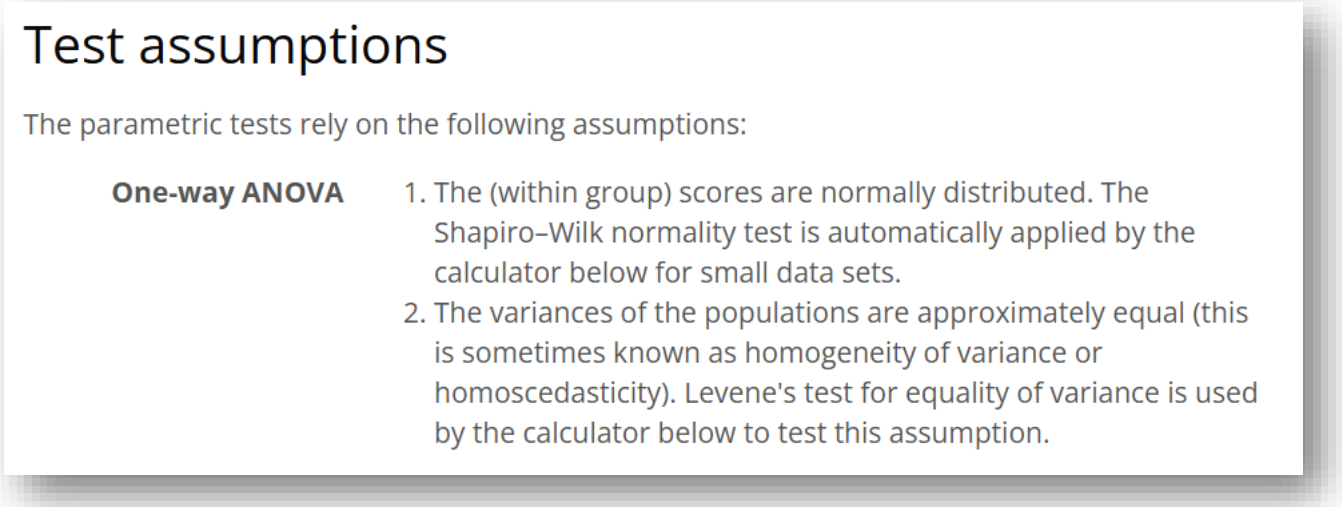
ANOVA online

https://www.ai-therapy.com/psychology-statistics/hypothesis-testing/anova

- Otevřte stránku
- Shrnuty předpoklady testu

(na dalším slajdu)

- Zadejte název (volitelné)
- Zadejte data
- Pokud kopírujete z českého Excelu desetinná čísla, tak nejprve nahraďte desetinné čárky tečkami
- Odeslat



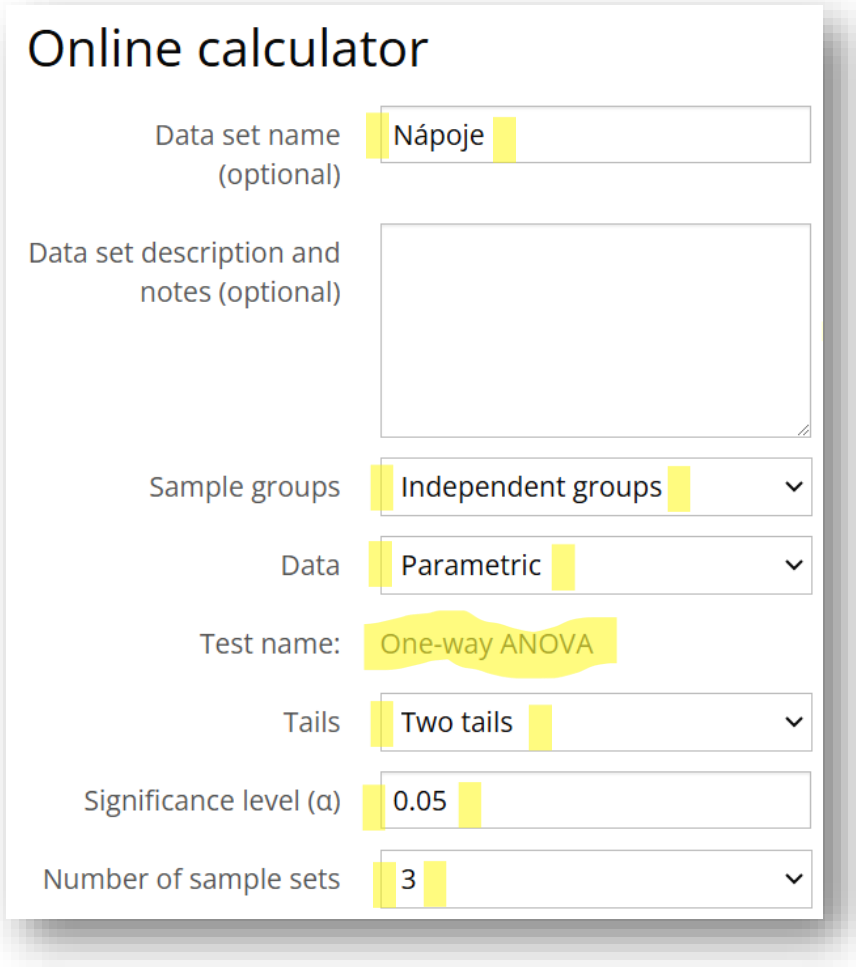
Test assumptions

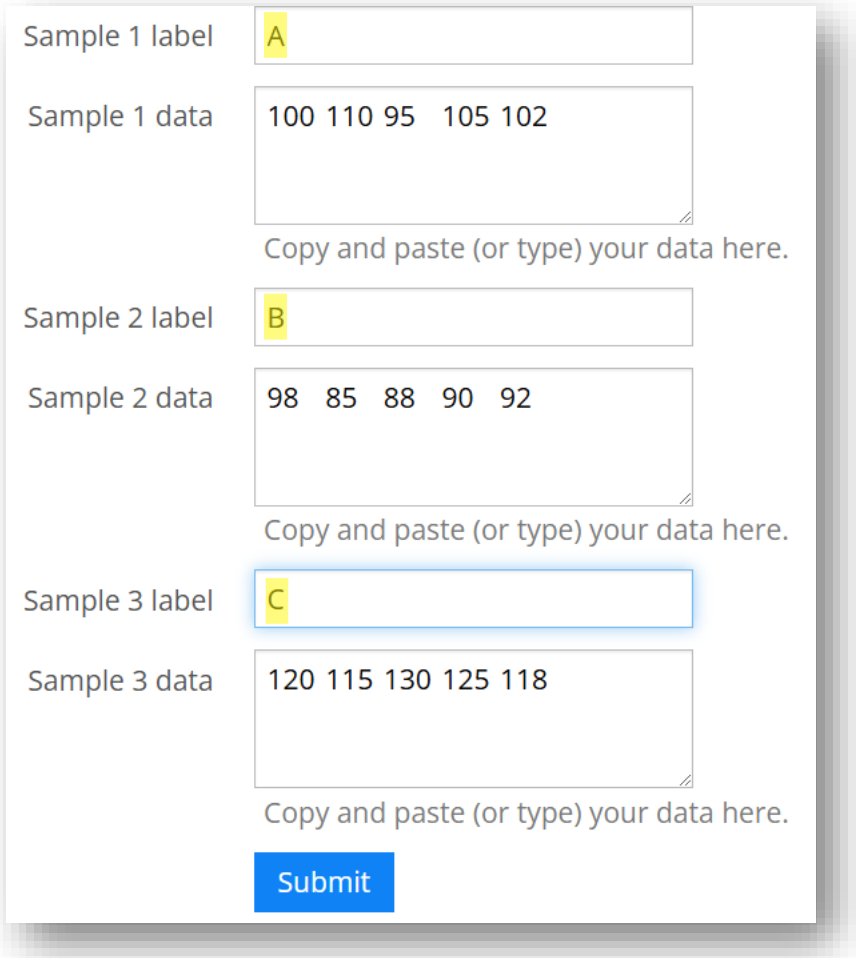
The parametric tests rely on the following assumptions:

One-way ANOVA

1. The (within group) scores are normally distributed. The Shapiro–Wilk normality test is automatically applied by the calculator below for small data sets.
2. The variances of the populations are approximately equal (this is sometimes known as homogeneity of variance or homoscedasticity). Levene's test for equality of variance is used by the calculator below to test this assumption.

ANOVA online





ANOVA online

Data set statistics

Sample name	Number of samples	Mean	Standard deviation	Standard error of the mean	Median
A	5	102.400	5.595	2.502	102.000
B	5	90.600	4.879	2.182	90.000
C	5	121.600	5.941	2.657	120.000

ANOVA online

Test summary

Test name	One way between groups ANOVA
Normality	The Shapiro–Wilk test was conducted on all data sets, and they appear to be approximately normal. Therefore, the assumption of normality is likely to be satisfied.
Levene's test for equality of variance	$F(2,12)=0.206$, $p= 0.817$
ANOVA test statistic	$F(2,12) = 40.622$
Statistical significance (Two-tailed)	$p < 0.001$
Conclusion	Based on a significance level of 0.05, there is a statistically siginificant difference between: {A,B,C}.

ANOVA online

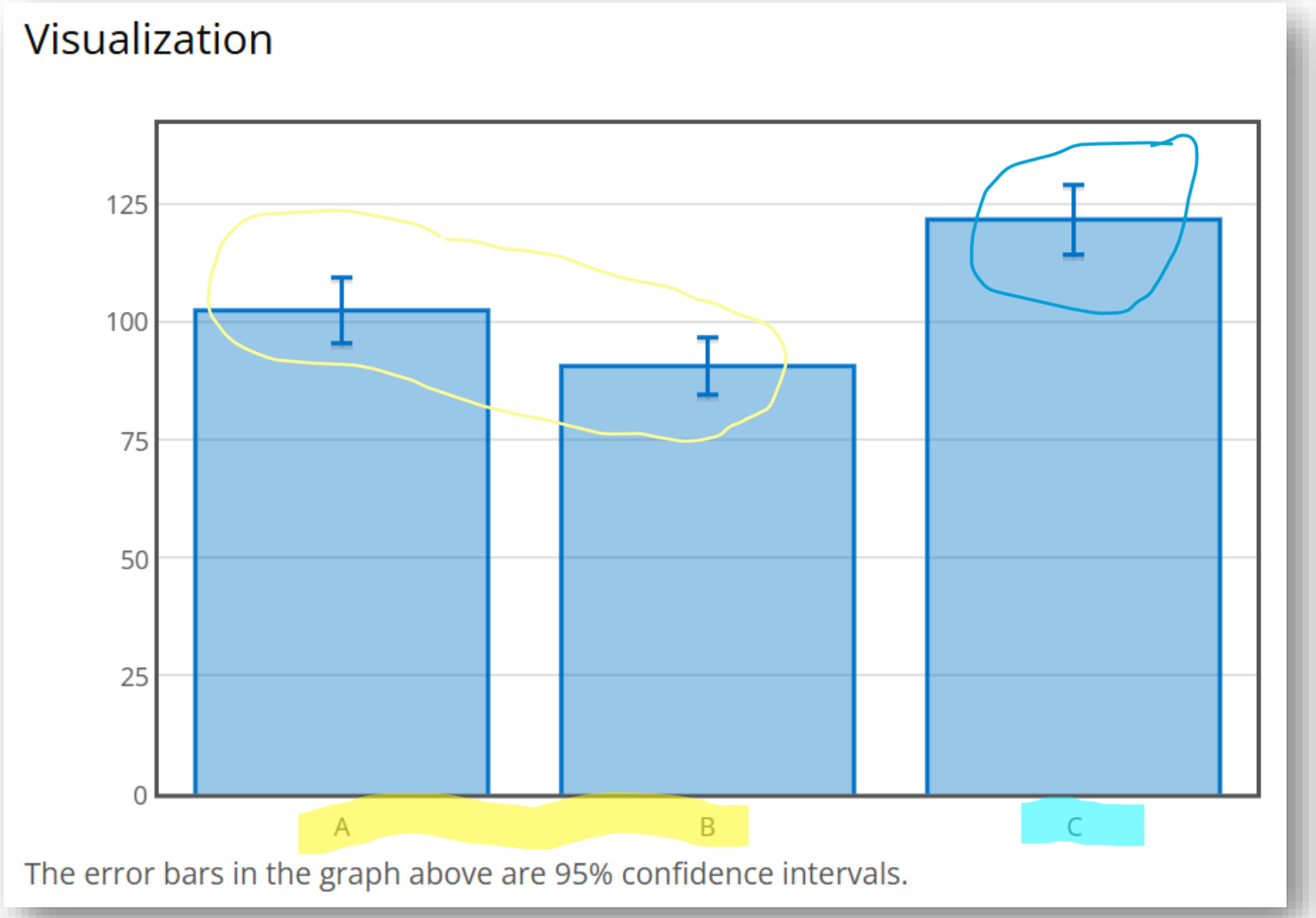
ANOVA test details

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean squares (MS)
Between Groups (Treatment/Model)	2448.133	2	1224.067
Within Groups (Error)	361.600	12	30.133
Total	2809.733	14	

Effect size

$\eta^2 = 0.871$	This is also known as r^2, and is the ratio of variance explained by the model.
$f^2 = 6.770$	Cohen's $f^2 = \eta^2 / (1 - \eta^2)$
$\omega^2 = 0.841$	This is an alternative measure of variance explained that has less bias than η^2.

ANOVA online



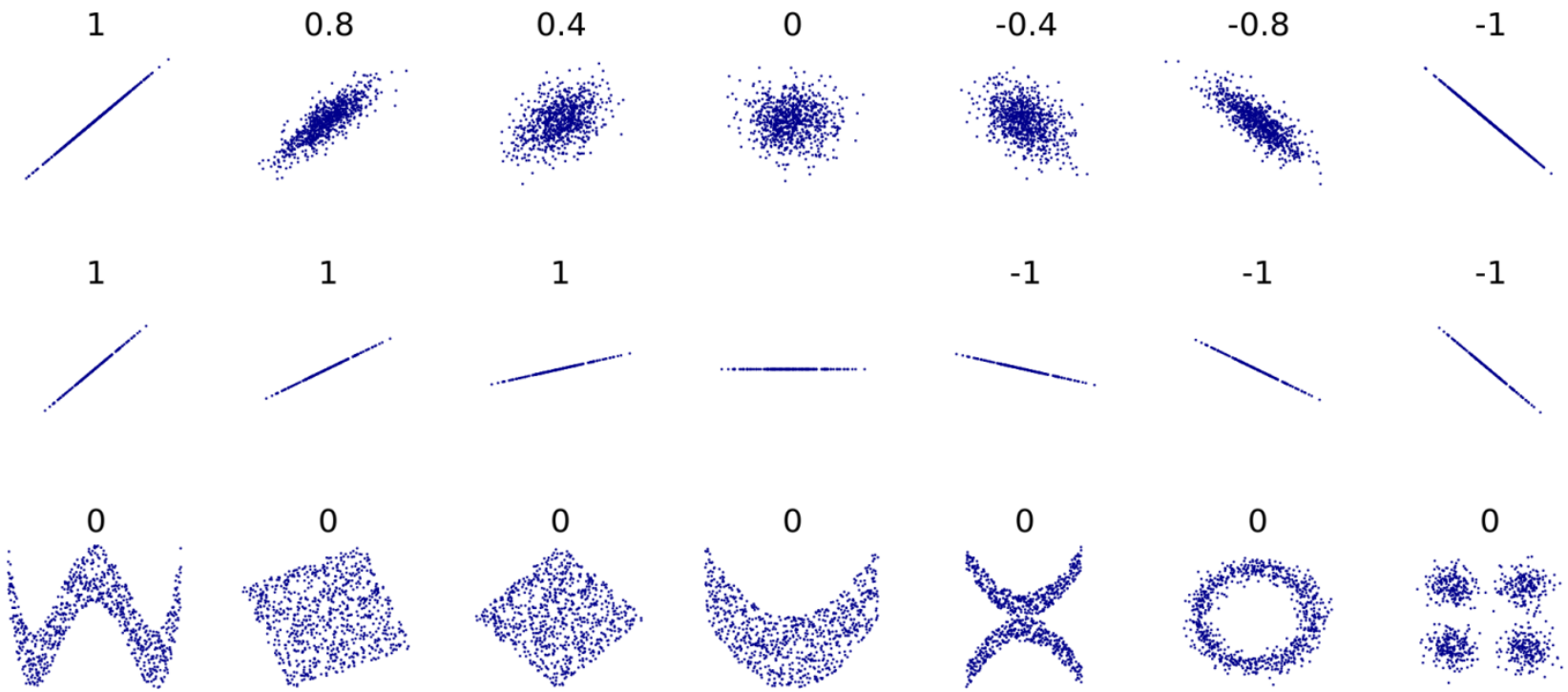
Korelační analýza

Co je to korelační koeficient?

Korelační koeficient je statistická míra, která určuje sílu a směr vztahu mezi dvěma proměnnými. Pearsonův korelační koeficient, označovaný jako r, měří lineární vztah mezi dvěma spojitými proměnnými a nabývá hodnot mezi -1 a 1. Pokud je $r = 1$, jedná se o perfektní pozitivní lineární vztah, pokud $r = -1$, jedná se o perfektní negativní lineární vztah, a pokud $r = 0$, neexistuje žádná lineární závislost mezi proměnnými.

Korelační analýza

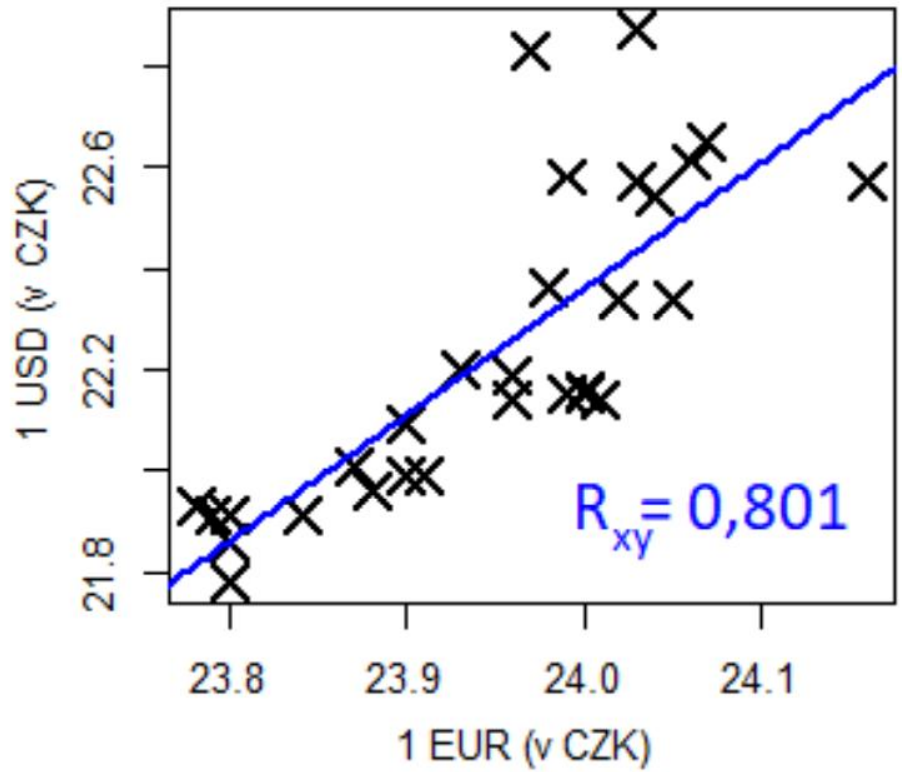
Vizualizace (ne)lineární závislosti



https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/1920px-Correlation_examples2.svg.png

Korelace - příklad

- Sledujeme kurzy české koruny k americkému dolaru a české koruny k euru během 30 dnů.
- Každý bod odpovídá kombinaci kurzů za daný den.
- Modrá přímka ukazuje lineární regresní vztah mezi oběma znaky.



Interpretace hodnot korelačního koeficientu

- 0-0,19: mezi znaky X a Y není lineární vztah
- 0,20-0,39: mezi X a Y je slabý pozitivní lineární vztah
- 0,40-0,59: mezi X a Y je středně silný pozitivní lineární vztah
- 0,60-0,79: mezi X a Y je silný pozitivní lineární vztah
- 0,80-1: mezi X a Y je velmi silný pozitivní lineární vztah
- analogicky pro záporné hodnoty

Korelační analýza

Výpočet korelačního koeficientu

Pearsonův korelační koeficient je definován vztahem:

$$r = \frac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \cdot \sum(y_i - \overline{y})^2}}$$

kde x_i a y_i jsou jednotlivé hodnoty obou proměnných, a $\overline{x}$ a $\overline{y}$ jsou jejich průměry.

Excel: Výpočet pomocí `CORREL(array1, array2)` v Excelu také ukazuje, že korelace je blízká nule, tedy nevýznamná.

Historie a varianty korelačních koeficientů

Historie korelačních koeficientů sahá až do 19. století, kdy Francis Galton poprvé navrhl metody pro kvantifikaci statistických vztahů mezi proměnnými. Na jeho práci navázal Karl Pearson, který formalizoval a popularizoval Pearsonův korelační koeficient.

V průběhu času byly vyvinuty další varianty korelačních koeficientů pro specifické účely:

- Spearmanův korelační koeficient (Spearman's rho): Používá se, pokud data nejsou normálně rozložena nebo vykazují monotónní, nikoli lineární vztah.
- Kendallův tau: Měří sílu vztahu mezi pořadím hodnot a používá se zejména u malých souborů dat.
- Point-biserial correlation: Využívá se pro měření korelace mezi spojitou a binární proměnnou.

Kdy je korelační koeficient vhodný?

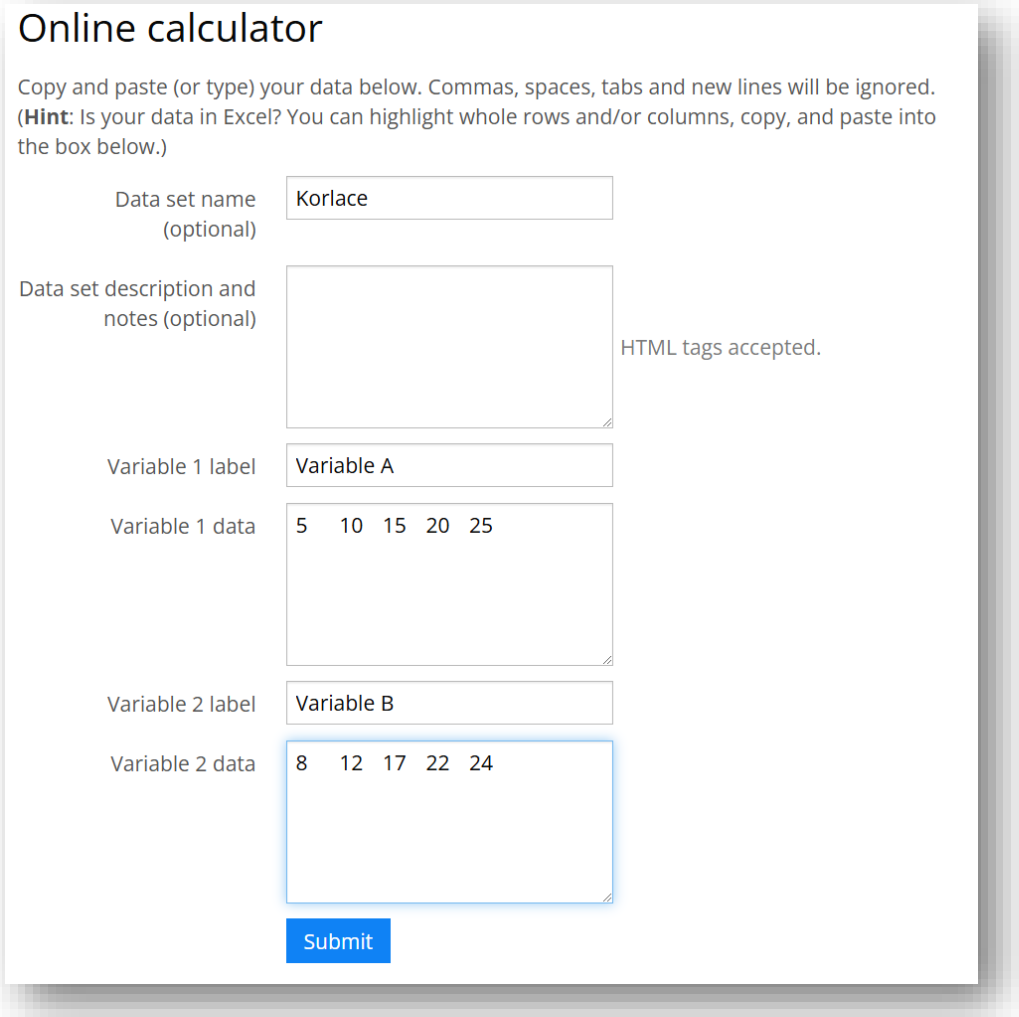
Korelační koeficient popisuje sílu a směr lineárního vztahu mezi dvěma spojitými proměnnými. Jeho použití je vhodné, pokud jsou splněny následující podmínky:

- Obě proměnné mají přibližně normální rozložení.
- Vztah mezi proměnnými je lineární.
- Nejsou přítomny výrazné odlehlé hodnoty, které by ovlivnily výsledek.

Korelační koeficient online

Odkaz pro zadávání:
https://www.ai-therapy.com/psychology-statistics/comparing-variables/correlation

Odkaz s výsledky:
https://www.ai-therapy.com/psychology-statistics/results/20241214004103808



Korelační koeficient online

Results

Data set overview

If you would like to share these results, please use the following link:

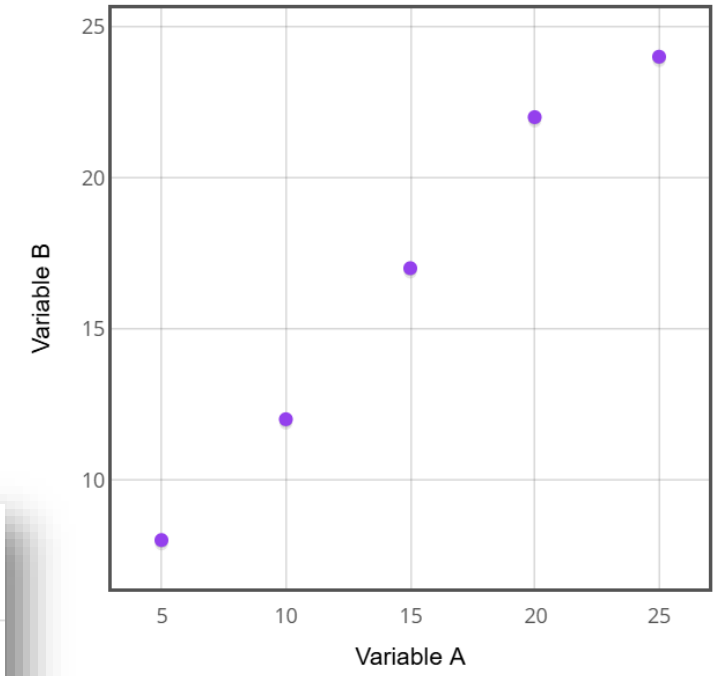
https://www.ai-therapy.com/psychology-statistics/results/20241214004103808

Data set name	Korlace
Data set notes	(not provided)

Data set statistics

Number of data points	N = 5
Covariance	σ(Variable A,Variable B) = 52.50
Pearson correlation coefficient	r(3) = 0.992, two-tailed *p*-value < 0.001
Spearman correlation	ρ(3) = 1.000, two-tailed *p*-value < 0.001
Kendall's correlation	τ(3) = 1.000, two-tailed *p*-value = 0.017

Scatter plot



Regresní analýza – Lineární regrese

Úvodní příklad

Představte si, že jste ekonomický analytik ve společnosti, která chce předpovědět tržby na základě výdajů na reklamu. Máte k dispozici následující data z posledních 10 měsíců:

Měsíc	1	2	3	4	5	6	7	8	9	10
Reklama (tis. Kč)	20	25	30	35	40	45	50	55	60	65
Tržby (tis. Kč)	200	220	250	280	310	330	360	390	420	450

Cílem je zjistit, jak silný je vztah mezi výdaji na reklamu a tržbami, a vytvořit model, který umožní předpovědět tržby při různých úrovních výdajů na reklamu.

Regresní analýza – Lineární regrese

Formulace problému

- **Závislá proměnná** (Y): Tržby (tis. Kč).
- **Nezávislá proměnná** (X): Výdaje na reklamu (tis. Kč).

Cíl analýzy

Pomocí lineární regrese odhadnout vztah mezi výdaji na reklamu a tržbami a posoudit, zda je tento vztah statisticky významný.

Co je to lineární regrese?

Lineární regrese je statistická metoda používaná k modelování vztahu mezi závislou proměnnou a jednou nebo více nezávislými proměnnými. V případě jednoduché lineární regrese se jedná o vztah mezi dvěma proměnnými, který je modelován pomocí přímky.

Lineární regresní model

Lineární regresní model lze vyjádřit rovnicí:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

kde:

- Y je závislá proměnná,
- X je nezávislá proměnná,
- β_0 je absolutní člen (intercept),
- β_1 je směrnice přímky (sklon),
- ε je náhodná chyba (reziduální složka).

Metoda nejmenších čtverců

Parametry β_0 a β_1 jsou odhadnuty pomocí *metody nejmenších čtverců*, která minimalizuje součet čtverců odchylek mezi skutečnými hodnotami Y a predikovanými hodnotami $\hat{Y}$:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Odhady parametrů

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

Odhady parametrů $\hat{\beta}_0$ a $\hat{\beta}_1$ lze vypočítat pomocí vzorců:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X},$$

kde $\overline{X}$ a $\overline{Y}$ jsou průměry X a Y.

Předpoklady lineární regrese

Aby byly odhady parametrů platné, musí být splněny následující předpoklady:

- **Linearita**: Vztah mezi X a Y je lineární.
- **Homoskedasticita**: Rozptyl náhodné složky ε je konstantní pro všechna X.
- **Nezávislost**: Hodnoty náhodné složky ε jsou nezávislé.
- **Normalita**: Náhodná složka ε je normálně rozložena.

Historické poznámky

Metoda lineární regrese byla poprvé formálně představena anglickým statistikem **Sir Francis Galtonem** v 19. století při studiu dědičnosti výšky mezi rodiči a dětmi. Termín *regrese* pochází z Galtonova pozorování, že extrémní hodnoty mají tendenci "regresovat" k průměru v následující generaci.

Později **Karl Pearson** a **Ronald A. Fisher** rozvinuli matematické základy regresní analýzy a metodu nejmenších čtverců, která je dnes standardním nástrojem v statistice a ekonometrice.

Odhad parametrů a interpretace

Výpočet odhadů

Pomocí výše uvedených vzorců lze spočítat odhady $\hat{\beta}_0$ a $\hat{\beta}_1$ na základě dostupných dat.

Interpretace parametrů

- **Směrnice přímky** ($\hat{\beta}_1$): Udává změnu v závislé proměnné Y při jednotkové změně nezávislé proměnné X.
- **Absolutní člen** ($\hat{\beta}_0$): Hodnota závislé proměnné Y, když nezávislá proměnná X je nulová.

Korelační koeficient a koeficient determinace

- **Pearsonův korelační koeficient** (r) měří sílu lineárního vztahu mezi X a Y.
- **Koeficient determinace** (R^2) udává, jaká část variability závislé proměnné Y je vysvětlena modelem.

Vzorec pro R^2 je:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

kde:

- SSR (regresní součet čtverců) $= \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$,
- SSE (reziduální součet čtverců) $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$,
- SST (celkový součet čtverců) $= \sum_{i=1}^{n} (Y_i - \overline{Y})^2$.

Testování významnosti regresních koeficientů

Testování směrnice přímky (β_1)

Cílem je zjistit, zda je vztah mezi X a Y statisticky významný.

Hypotézy:

- $H_0 : \beta_1 = 0$ (neexistuje lineární vztah mezi X a Y).
- $H_1 : \beta_1 \neq 0$ (existuje lineární vztah mezi X a Y).

Testová statistika:

$$t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)},$$

kde $\mathrm{SE}(\hat{\beta}_1)$ je směrodatná chyba odhadu $\hat{\beta}_1$.

Rozhodnutí: Porovnáme vypočtenou hodnotu t s kritickou hodnotou z t-rozdělení s $n - 2$ stupni volnosti.

Výpočet a testování v Excelu

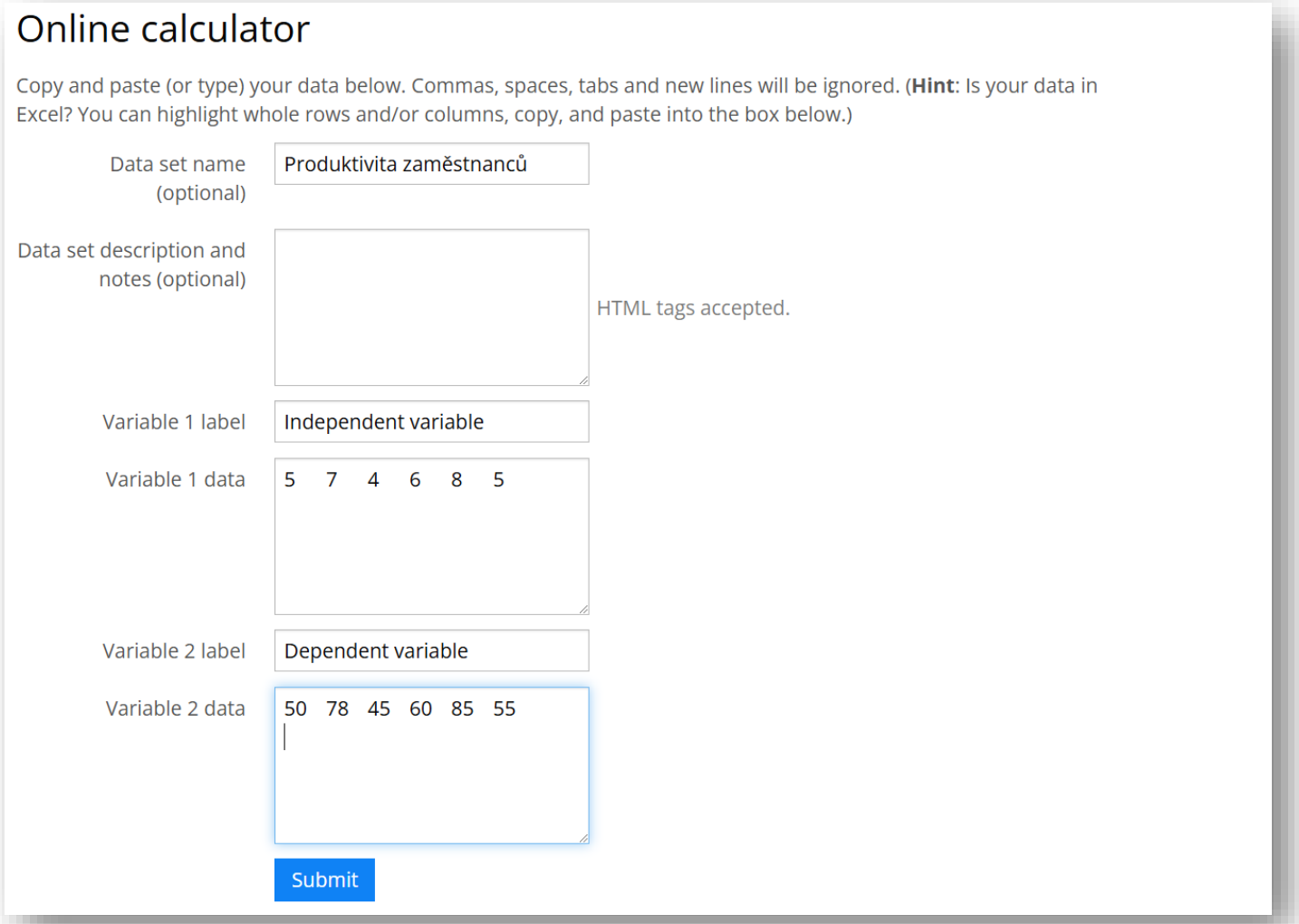
Parametr	Odhad	Směr. chyba	t	P-hodnota
$\hat{\beta}_0$	$-0{,}3333$	$5{,}0$	$-0{,}0667$	$0{,}9494$
$\hat{\beta}_1$	$10{,}7143$	$0{,}8660$	$12{,}3693$	$0{,}0001$

Rozhodnutí:

P-hodnota pro $\hat{\beta}_1$ je $0{,}0001$, což je menší než $\alpha = 0{,}05$. Zamítáme tedy nulovou hypotézu $H_0 : \beta_1 = 0$. Regresní koeficient $\hat{\beta}_1$ je statisticky významný.

Lineární regrese online

https://www.ai-therapy.com/psychology-statistics/comparing-variables/regression



Lineární regrese online

https://www.ai-therapy.com/psychology-statistics/results/20241214014106311

Regression results

Data set overview

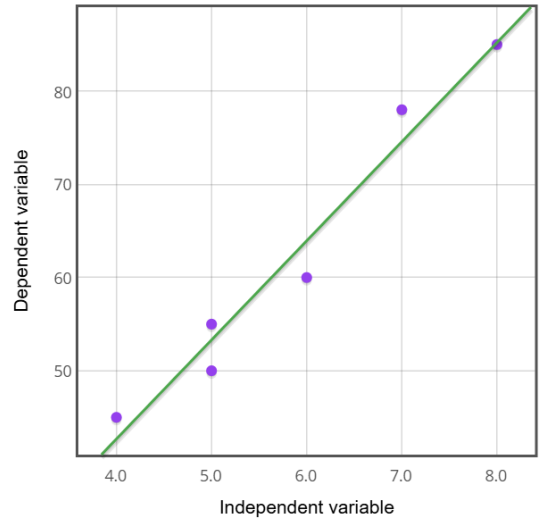
If you would like to share these results, please use the following link:

https://www.ai-therapy.com/psychology-statistics/results/20241214014106311

Data set name	Produktivita zaměstnanců
Data set notes	(not provided)

Data set statistics

Line equation	Dependent variable = (Independent variable × 10.631) + 0.154
Standard error of the slope	1.036
R-squared	0.963
Two-tailed p-value	< 0.001 (we can reject the null hypthesis that the slope is 0)



Hover the mouse over the plot to get the regression prediction.

Independent variable	3.94
Estimated Dependent variable	42.02