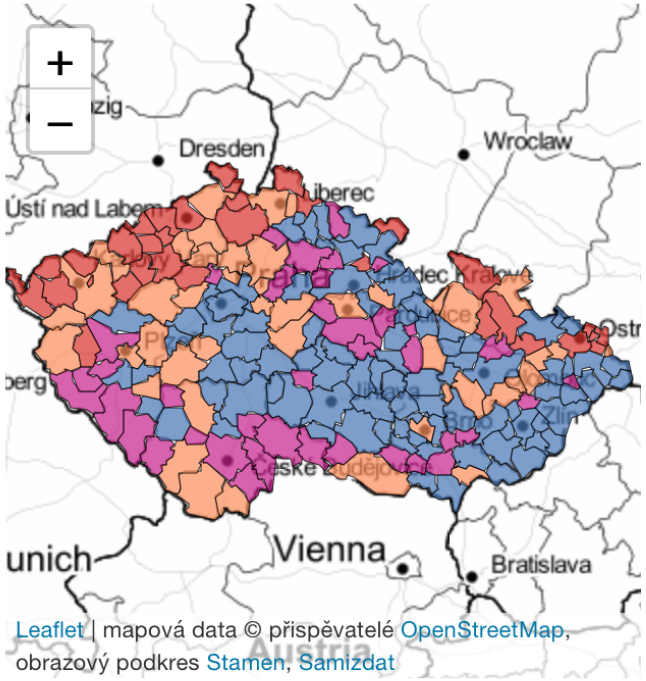
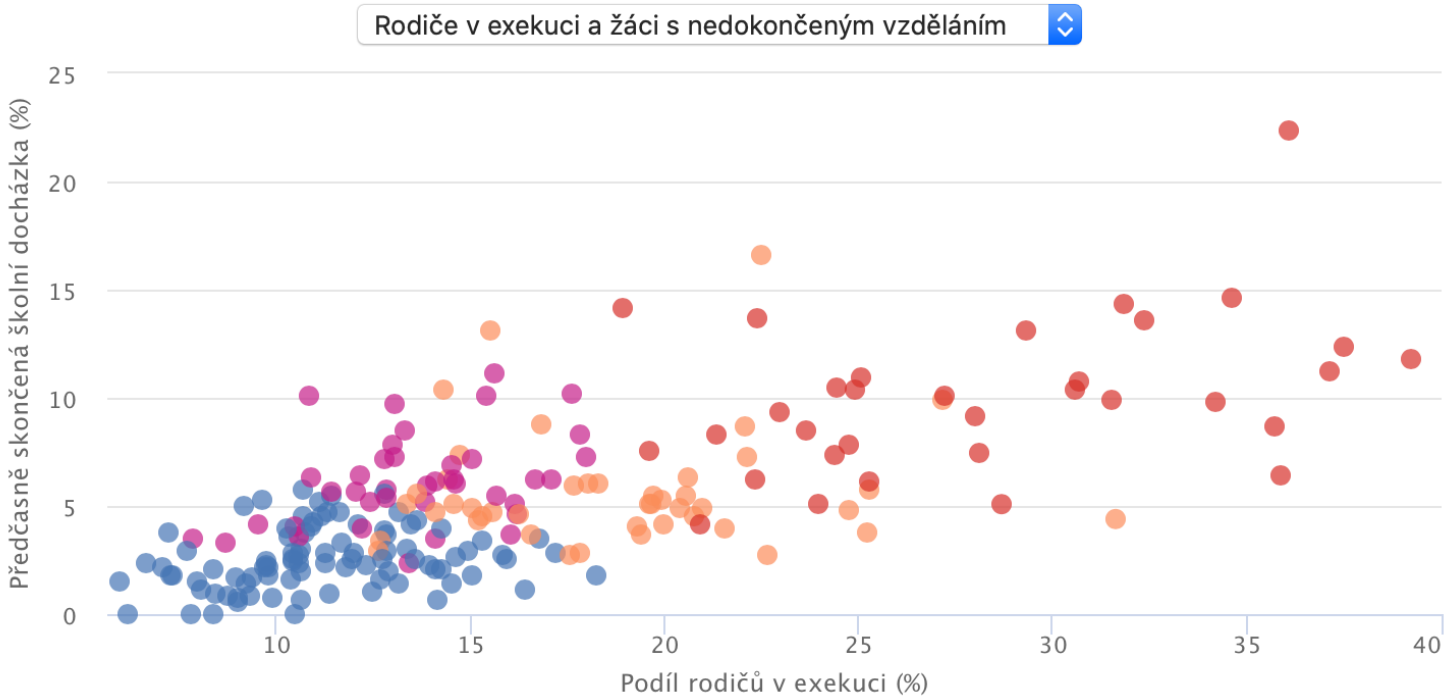
# Základní statistické nástroje

ANALÝZA DAT

25. BŘEZNA 2021

# Základní statistika v médiích



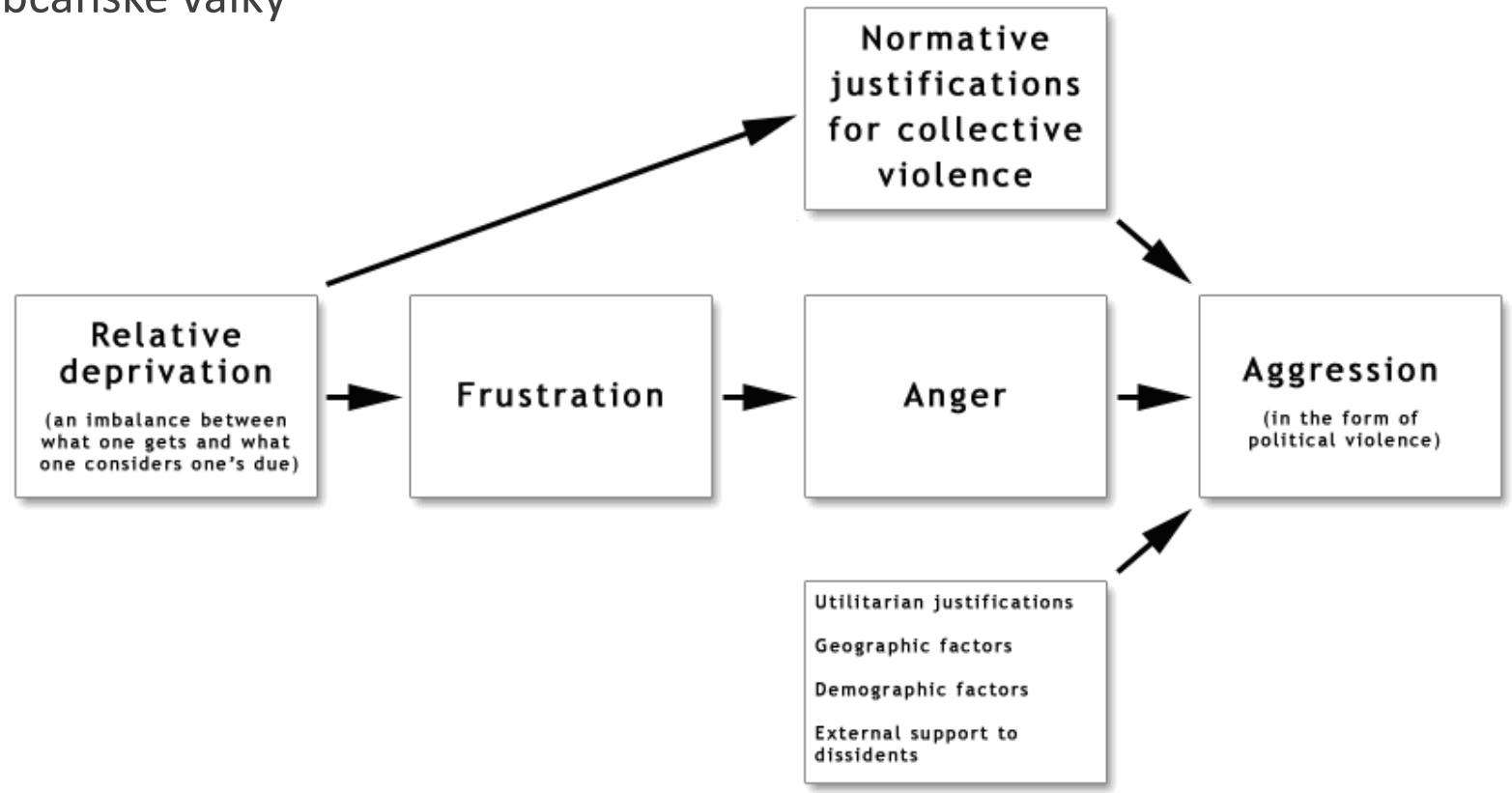

irozhlas.cz

# Vztahy jako objekt analýzy

- analýzy většinou odhalují vztahy mezi proměnnými
- na základě jejich popisu analýza formuluje hypotézy, teorie či dokonce zákony
- předobrazem pro výzkum jsou přírodní vědy, jejich hlavní výhodou je možnost vytvoření experimentálního prostředí
- výhodou experimentu je kontrola zkoumaného procesu (možnost modelovat)
  - pokud například v experimentálním prostředí udržujeme stabilní teplotu či vakuum a vztah mezi proměnnými se pořád projevuje, víme, že za něj nemůže ani teplota či přítomnost vzduchu
- inspirujeme se využitím metod přírodními vědami, ale jejich úplné a přímé využití není možné

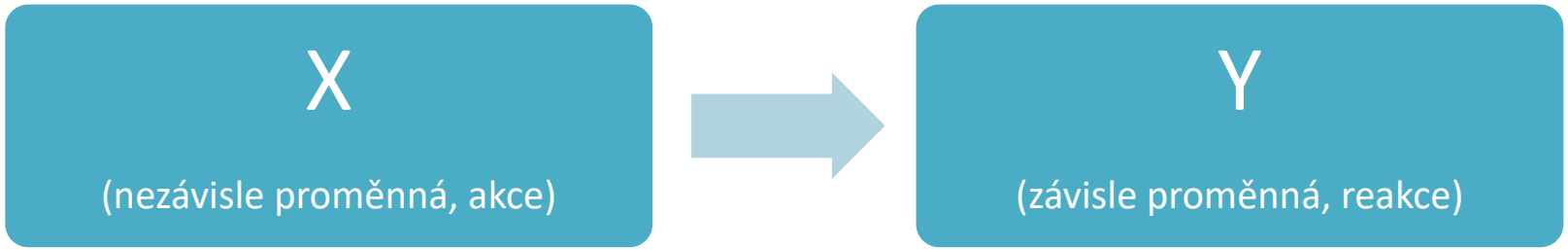
# Komplikované vztahy v sociálních vědách

- příčiny občanské války




popularsocialscience.com

# Kauzální hypotéza




- když X, tak (ne) Y
  - deterministická hypotéza
- když X, tak pravděpodobně (ne) Y
  - pravděpodobnostní hypotéza
- čím více X, tím více (méně) Y

# Testování hypotéz

- “In so far as a scientific statement speaks about reality, it must be falsifiable: and in so far as it is not falsifiable, it does not speak about reality.” (Karl R. Popper, The Logic of Scientific Discovery, 1934)

- tedy hypotéza, kterou nelze vyvrátit a testovat, není vědeckou
- příklady (ne)vědeckých hypotéz
- hypotézy lze testováním pouze vyvrátit (zamítnout), nikoliv potvrdit
- výsledek výzkumu tedy může být dvojí
  - hypotéza byla zamítnuta
  - hypotézu se nepodařilo zamítnout

# Testování hypotéz

- nulová hypotéza
  - jedná se o hypotézu, kterou testujeme (snažíme se ji vyvrátit)
  - popisuje výchozí stav poznání

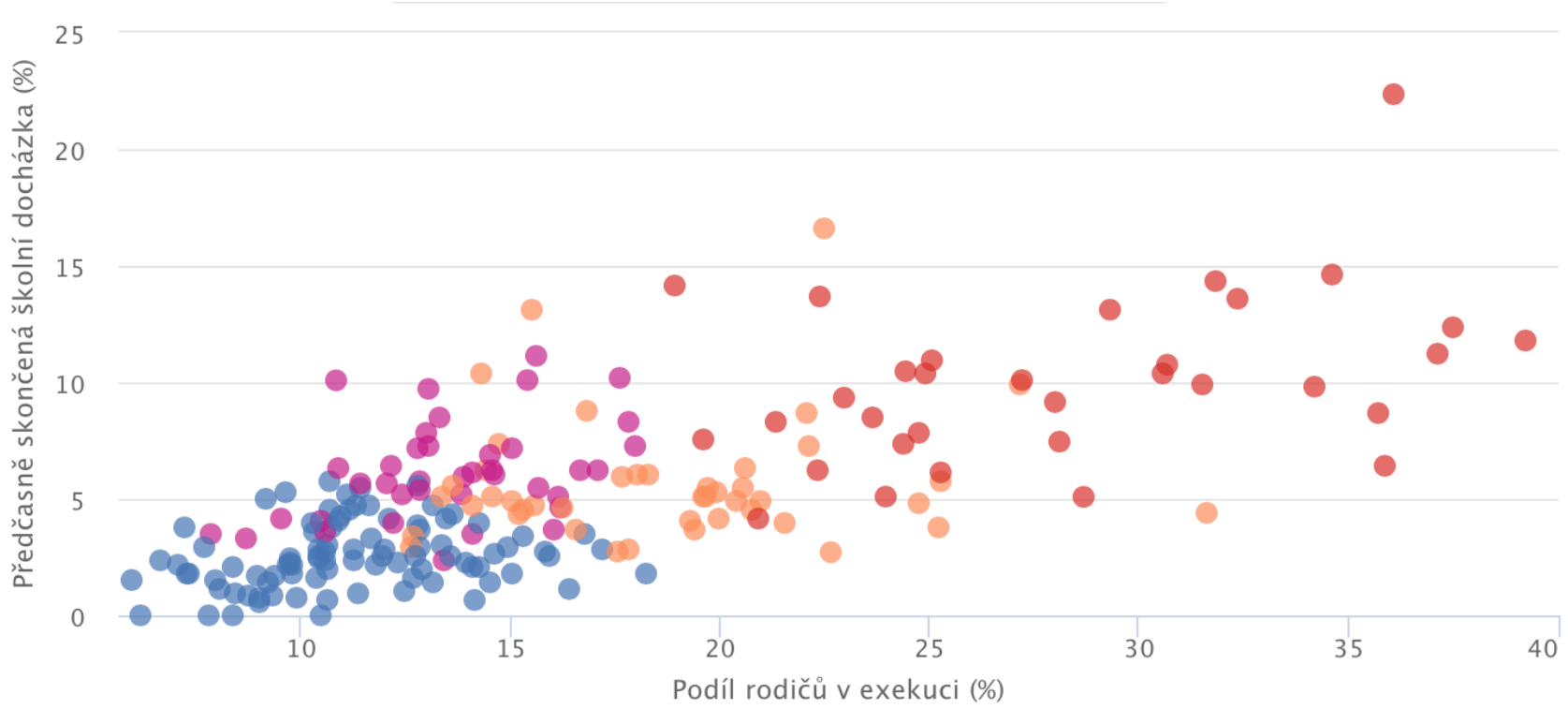
  - například chci odhalit, zda obce s vyšší nezaměstnaností volili spíše Miloše Zemana než Jiřího Drahoše
  - $H_0$: mezi mírou nezaměstnanosti a volbou Miloše Zemana není žádný vztah

- výzkumná (alternativní) hypotéza
  - jde o hypotézu popisující vztah, který předpokládáme

  - H: čím vyšší míra nezaměstnanosti, tím vyšší podíl hlasů pro Miloše Zemana

# Základem odhalování kauzality je korelace

- X na ose x
- Y na ose y

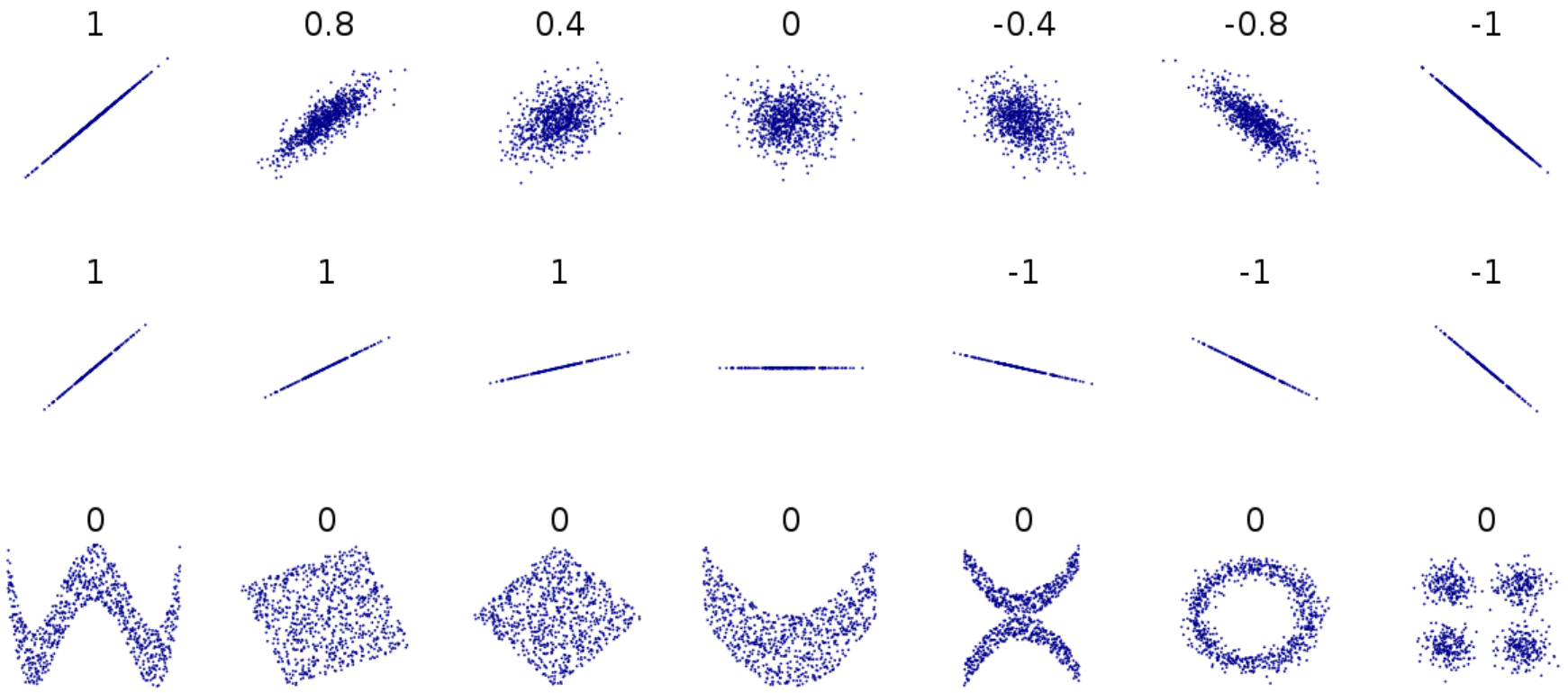



irozhlas.cz

# Korelace

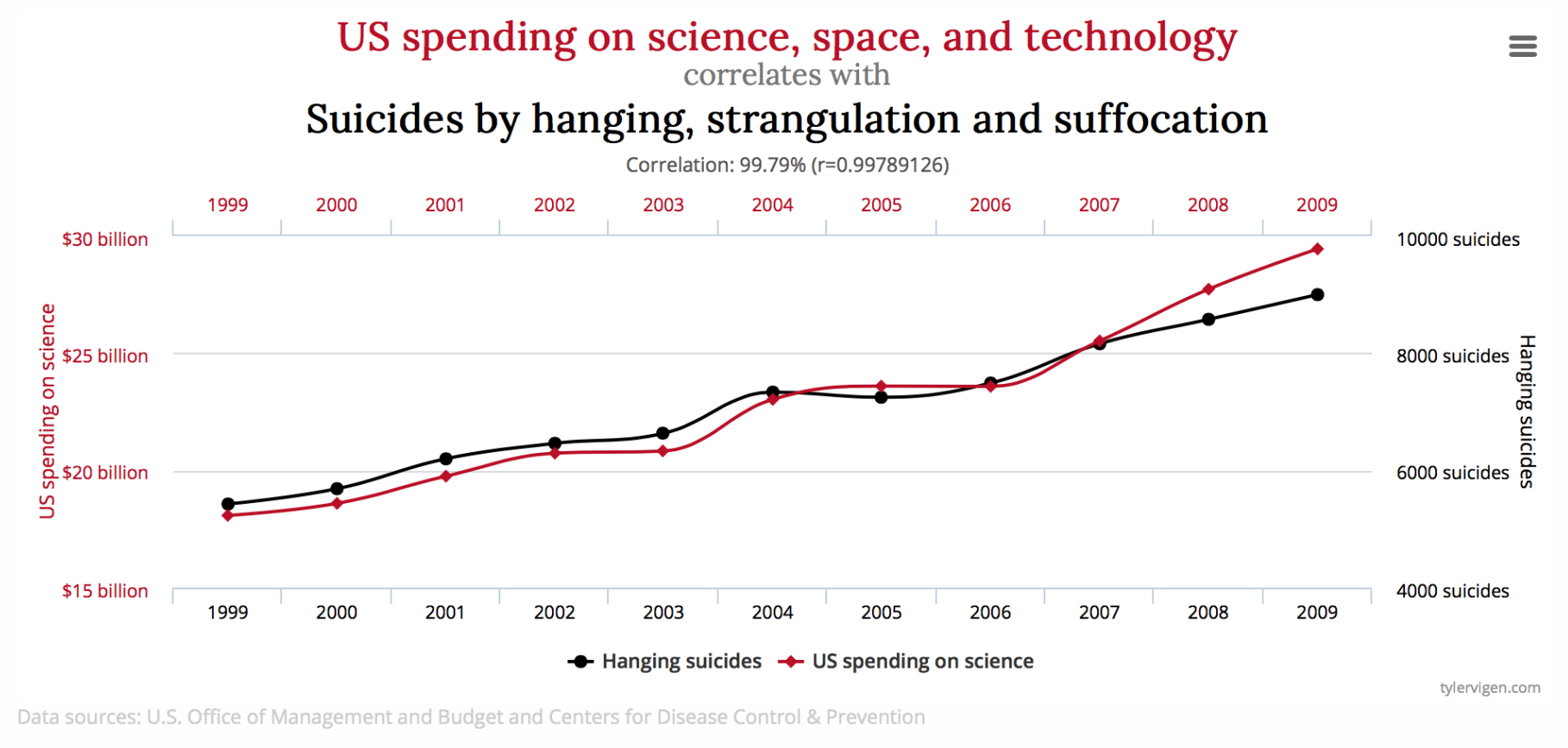
- míra vztahu mezi dvěma proměnnými
- skládá se ze dvou komponentů
  - směr – může být buď pozitivní (↑ X, ↑ Y), nebo negativní (↑ X, ↓ Y)
  - síla – nabývá hodnot od -1 (perfektní negativní korelace) přes 0 (žádný vztah) do 1 (perfektní pozitivní korelace)
- počítá se nejčastěji pomocí Pearsonova korelačního koeficientu
  - koeficient je získán pomocí kovariance, která je standardizována pomocí vydělení obou proměnných jejich směrodatnými odchylkami
- alternativou je například Spearmanův korelační koeficient (nepracuje s hodnotami proměnných, ale pořadím těchto hodnot)

# Korelace

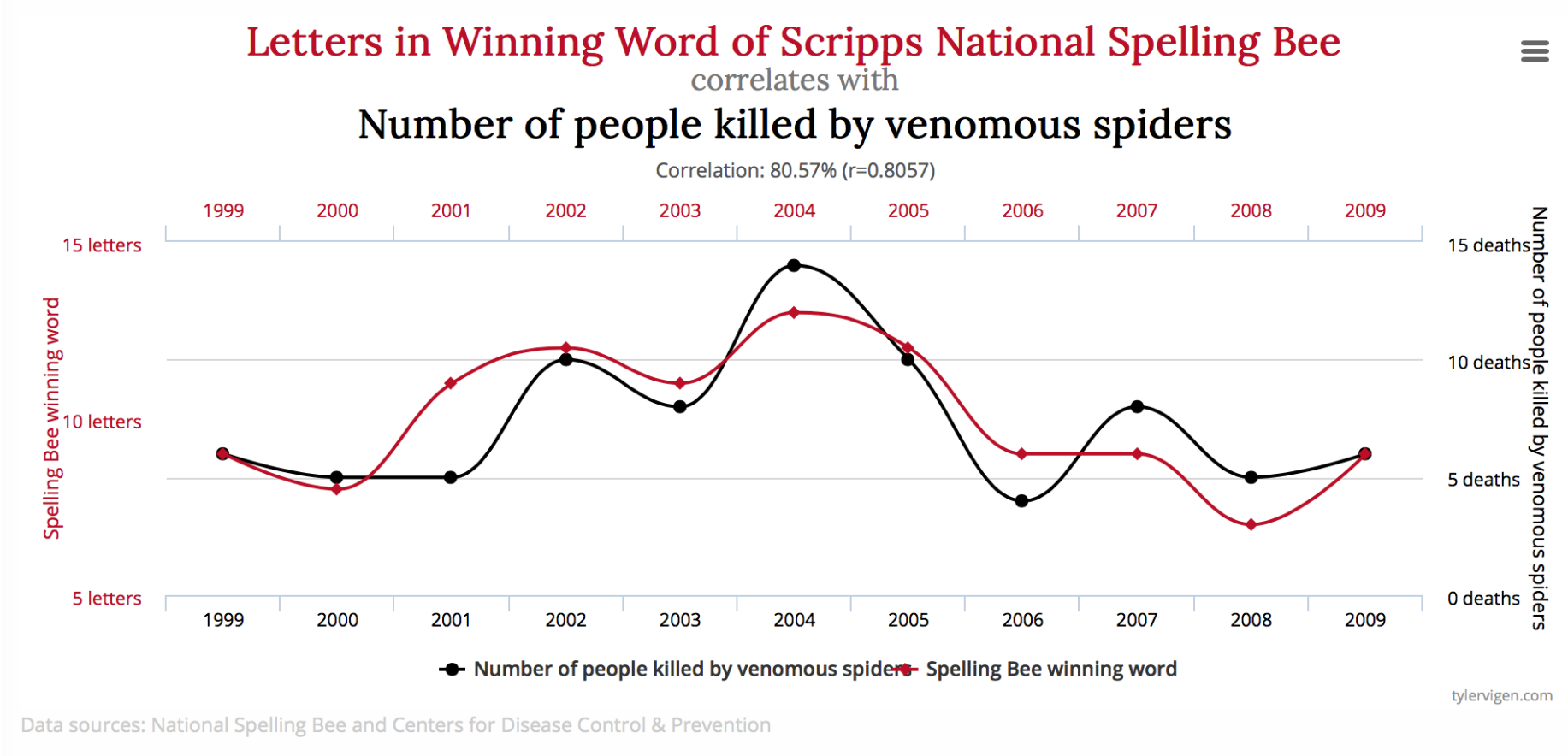



wikipedia.org

# Koreluje mnoho věcí

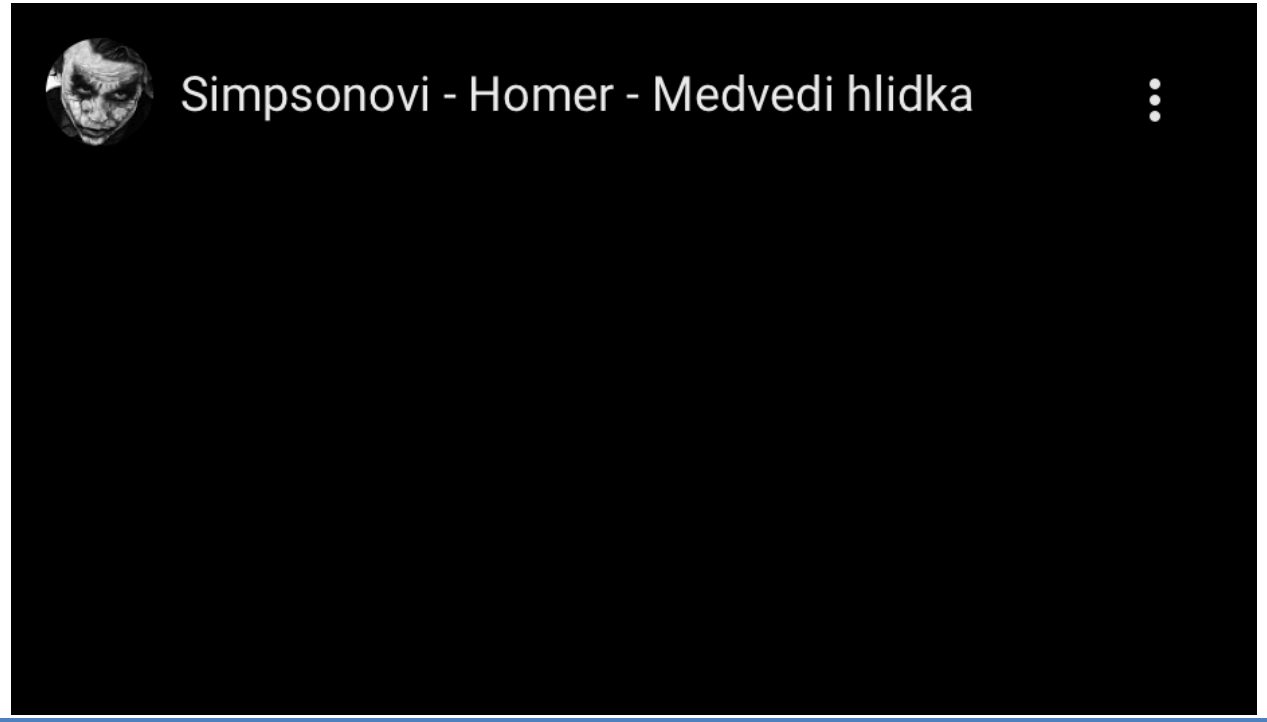


# Koreluje mnoho věcí



# Korelace není kauzalita!

- ne vždy je vztah dvou proměnných možné vysvětlit jako kauzalitu
- špatná interpretace vzájemného vztahu proměnných je jednou z nejčastějších a zároveň nejvážnějších chyb analýzy dat



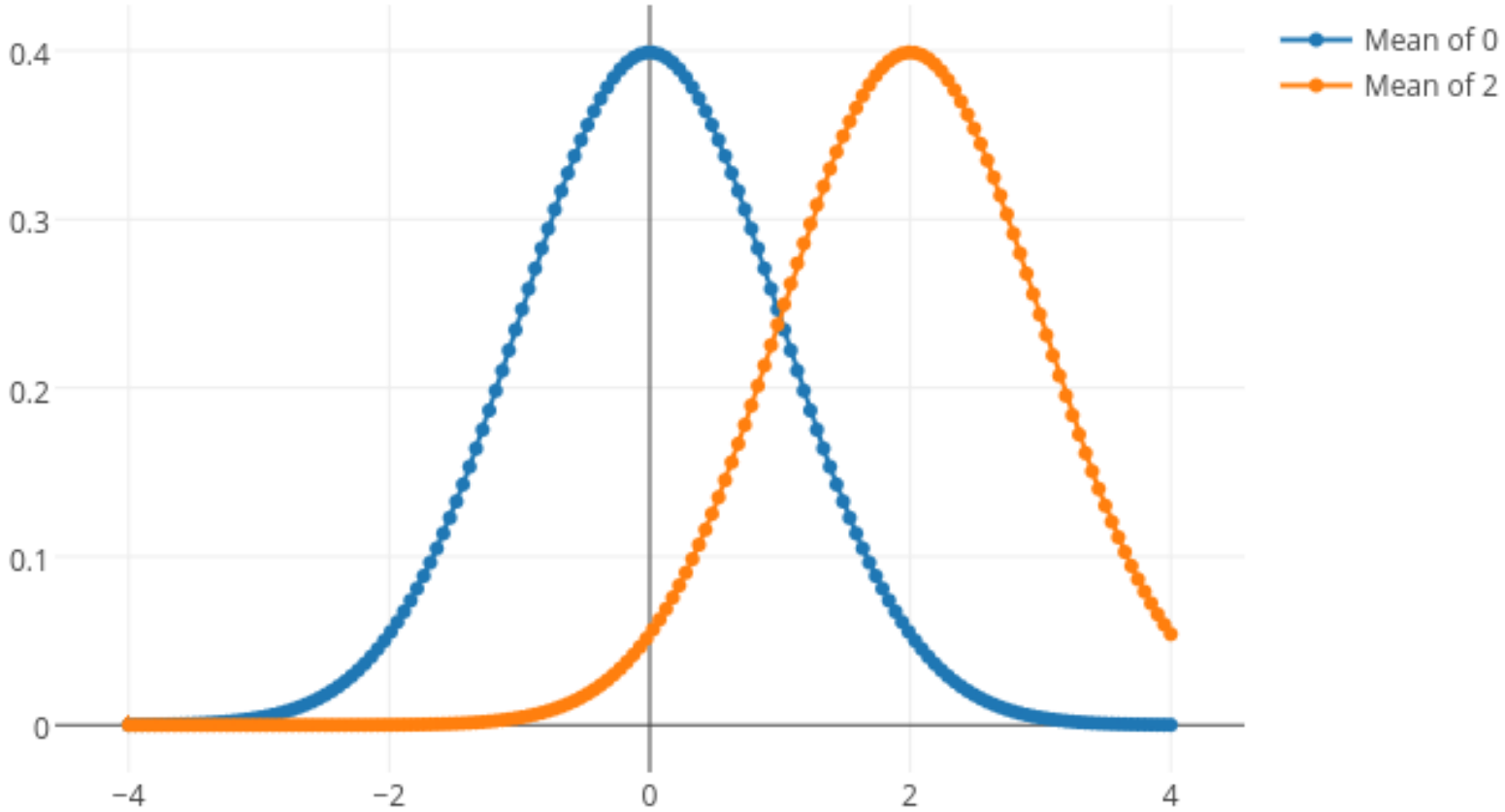
# Co korelace dále neodhalí

- nepravá korelace
  - X a Y jsou shodně ovlivňovány nepozorovanou proměnnou Z
  - „kdo je zdravější (X), volí pravici (Y)“ - ve skutečnosti oba faktory jsou ovlivněny vzděláním (Z)
- vývojová sekvence
  - X ovlivňuje Y, ale samo je ovlivňováno nepozorovanou proměnnou Z
  - „převážně psychicky labilní jedinci (X) vstupují do politiky (Y)“ – ve skutečnosti do politiky převážně vstupují ambiciózní lidé (Z), kteří mají větší sklon k labilitě
- chybějící střední člen
  - mezi X a Y je nepozorovaná proměnná Z
  - „pohlaví (X) ovlivňuje úspěch žen a mužů v médiích (Y)“ – ve skutečnosti je mezi tím starost o děti (Z)
- dvojí příčina
  - Y má více příčin, ale jen jedna je zahrnuta do výzkumu
  - „nezaměstnanost (X) zvyšuje podporu pro Zemana (Y)“ – ve skutečnosti i velikost obce ($Z_1$) či věk ($Z_2$)

# t-test

- proměnné, u kterých už máme možnost smysluplně spočítat například průměr (intervalové, poměrové), můžeme zkoumat dalšími, více pokročilými metodami
- jednou z nich je např. t-test
- umožňuje ověřit, zda dvě normální rozložení, z nichž pocházejí dva nezávislé výběry, mají stejné střední hodnoty (průměry)
- $H_0$: průměry dvou populací jsou stejné
  - $H_0$: $\mu_1 = \mu_2$
- alternativní hypotézou je, že průměry dvou populací se signifikantně liší (liší se tedy tyto dvě populace v určité hodnotě)

# t-test




plot.ly

# Co to prakticky znamená?

## V listopadu si pohoršilo hnutí ANO i Piráti. ODS by byla druhá

8. 12. 2019

Horšího výsledku dosáhlo v listopadovém volebním modelu agentury Kantar CZ hnutí ANO, které ztratilo tři procentní body, stejně tak Piráti, kteří přišli o 3,5 procentního bodu. Zatímco vládní hnutí ale nadále zůstává na prvním místě, druhý nejlepší výsledek by tentokrát měla ODS. Poprvé se pak dostala nad pětiprocentní hranici Trikolóra. Výzkum agentura zpracovala pro Českou televizi.

ct24.ceskatelevize.cz

# t-test

- postup při výpočtu je následující (POZOR! – platný pouze v případě nepárového testu a různých rozptylů obou vzorků, což je nejvíce konzervativní test)
  1. vypočítat průměry a směrodatné odchylky obou vzorků
  2. vypočítat jednotlivé standardní chyby obou vzorků
  3. vypočítat celkovou standardní chybu vzorků

$$SE_d = \sqrt{SE_1^2 + SE_2^2}$$

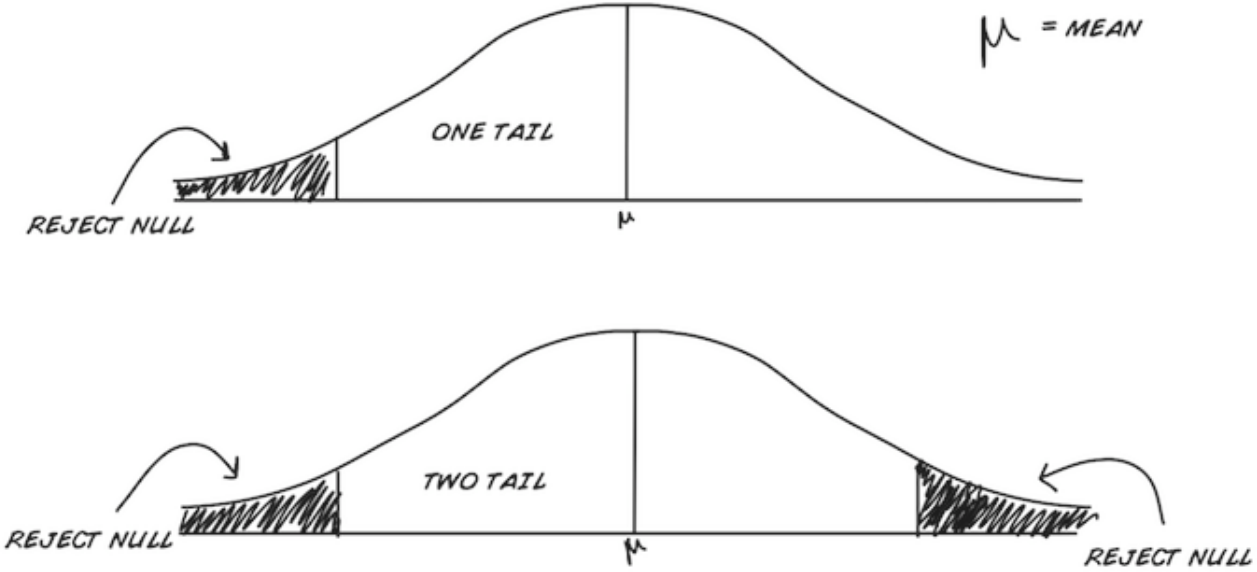
  4. vypočítat t-skóre

$$t = \frac{\overline{X_1} - \overline{X_2}}{celková\ standardní\ chyba}$$

  5. porovnat t-skóre s t-tabulkou za účelem odhalení pravděpodobnosti platnosti nulové hypotézy
- při výpočtu t-testu například v Excelu získáme přímo hodnotu pravděpodobnosti platnosti nulové hypotézy (viz dále)
- pozor na různé t-testy vzhledem k vzájemné závislosti vzorků
  - v případě nejistoty vždy používat nejvíce konzervativní nepárový test pro vzorky s rozdílnými rozptyly

# Test jednostranný nebo oboustranný?

- proti nulové hypotéze stavíme alternativní hypotézu
- ta může být buď jednostranná nebo oboustranná
- pokud je alternativní hypotéza $H_A$: $\mu_1 \neq \mu_2$, je možné, že $\mu_1 > \mu_2$, nebo $\mu_1 < \mu_2$ a musíme proto použít dvoustranný test
  - například ANO má v lednu 2020 jinou volební podporu než v lednu 2019
- pokud je ale alternativní hypotéza například jen $H_A$: $\mu_1 > \mu_2$, použijeme jednostranný test
  - například ANO má v lednu 2020 vyšší volební podporu než v lednu 2019




backyardbrains.com

# t-tabulka

- při výpočtu t-skóre se podíváme do t-tabulky, co hodnota značí
- najdeme si příslušný řádek podle stupňů volnosti (degrees of freedom)
  - pokud řádek s příslušnými stupni volnosti chybí, použijeme nejbližší konzervativnější hodnotu (např. pokud chybí df = 36, použijeme df = 30)
- určíme dva sloupce, mezi kterými se hodnota t-skóre nachází
- sloupec s nižší hodnotou jistoty značí naši minimální pravděpodobnost odlišnosti srovnávaných dat
- vzhledem ke konvenci nás obecně zajímá, zda je pravděpodobnost vyšší či nižší než 95 %

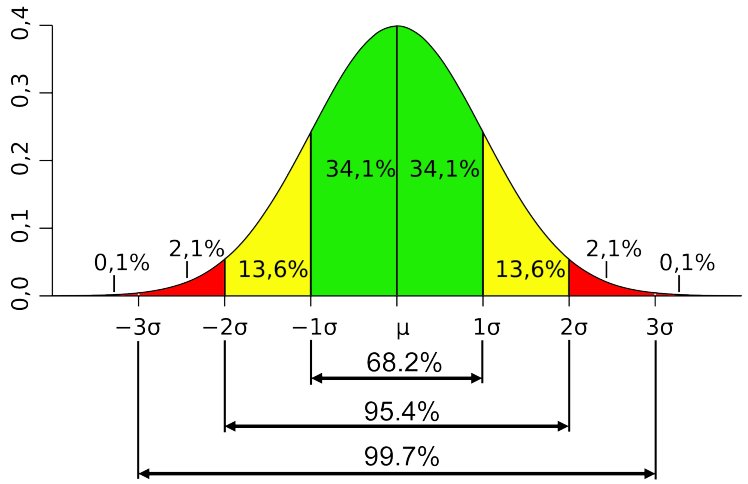
| cum. prob | $t_{.50}$ | $t_{.75}$ | $t_{.80}$ | $t_{.85}$ | $t_{.90}$ | $t_{.95}$ | $t_{.975}$ | $t_{.99}$ | $t_{.995}$ | $t_{.999}$ | $t_{.9995}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| one-tail | 0.50 | 0.25 | 0.20 | 0.15 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | 0.0005 |
| two-tails | 1.00 | 0.50 | 0.40 | 0.30 | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 | 0.002 | 0.001 |
| df | | | | | | | | | | | |
| 1 | 0.000 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 318.31 | 636.62 |
| 2 | 0.000 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 | 31.599 |
| 3 | 0.000 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 | 12.924 |
| 4 | 0.000 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 0.000 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 0.000 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 0.000 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 0.000 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 0.000 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 0.000 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 0.000 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 0.000 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 0.000 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 0.000 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 0.000 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 0.000 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 0.000 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 0.000 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| 19 | 0.000 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 0.000 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | 0.000 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 0.000 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 0.000 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.768 |
| 24 | 0.000 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | 0.000 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | 0.000 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | 0.000 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | 0.000 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | 0.000 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | 0.000 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |
| 40 | 0.000 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 | 3.551 |
| 60 | 0.000 | 0.679 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 | 3.460 |
| 80 | 0.000 | 0.678 | 0.846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 | 3.416 |
| 100 | 0.000 | 0.677 | 0.845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 | 3.390 |
| 1000 | 0.000 | 0.675 | 0.842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 | 3.098 | 3.300 |
| z | 0.000 | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 | 3.291 |
| | 0% | 50% | 60% | 70% | 80% | 90% | 95% | 98% | 99% | 99.8% | 99.9% |
| | Confidence Level | | | | | | | | | | |

ttable.org

# z-skóre

- se zvyšujícím se počtem df se rozložení blíží normálnímu
- v normálním rozložení z-skóre

$$z = \frac{x - \mu}{\sigma}$$




kanbanize.com

| cum. prob | $t_{.50}$ | $t_{.75}$ | $t_{.80}$ | $t_{.85}$ | $t_{.90}$ | $t_{.95}$ | $t_{.975}$ | $t_{.99}$ | $t_{.995}$ | $t_{.999}$ | $t_{.9995}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| one-tail | 0.50 | 0.25 | 0.20 | 0.15 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | 0.0005 |
| two-tails | 1.00 | 0.50 | 0.40 | 0.30 | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 | 0.002 | 0.001 |
| df | | | | | | | | | | | |
| 1 | 0.000 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 318.31 | 636.62 |
| 2 | 0.000 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 | 31.599 |
| 3 | 0.000 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 | 12.924 |
| 4 | 0.000 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 0.000 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 0.000 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 0.000 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 0.000 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 0.000 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 0.000 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 0.000 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 0.000 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 0.000 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 0.000 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 0.000 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 0.000 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 0.000 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 0.000 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| 19 | 0.000 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 0.000 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | 0.000 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 0.000 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 0.000 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.768 |
| 24 | 0.000 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | 0.000 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | 0.000 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | 0.000 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | 0.000 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | 0.000 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | 0.000 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |
| 40 | 0.000 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 | 3.551 |
| 60 | 0.000 | 0.679 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 | 3.460 |
| 80 | 0.000 | 0.678 | 0.846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 | 3.416 |
| 100 | 0.000 | 0.677 | 0.845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 | 3.390 |
| 1000 | 0.000 | 0.675 | 0.842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 | 3.098 | 3.300 |
| z | 0.000 | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 | 3.291 |
| | 0% | 50% | 60% | 70% | 80% | 90% | 95% | 98% | 99% | 99.8% | 99.9% |
| | Confidence Level | | | | | | | | | | |

ttable.org

# p-hodnota

- klíčová informace, která určuje pravděpodobnost platnosti nulové hypotézy (v t-testu je to situace, kdy oba vzorky nejsou odlišné – pocházejí ze stejné populace)
- pokud nás zajímá odhalení rozdílu obou vzorků (tedy alternativní hypotéza), pravidlo je, že čím je p-hodnota nižší, tím lépe (tím je totiž méně pravděpodobná platnost nulové hypotézy a komplementárně roste pravděpodobnost zamítnutí nulové hypotézy)
- pokud je naopak naším cílem platnost nulové hypotézy, vyšší p-hodnota je lepší
- tradičně jde o to, aby p-hodnota překročila konvenční hranici (95 % nebo 99 %)
- pomocí p-hodnoty ale můžeme odhalit i přesné procento jistoty možnosti zamítnutí nulové hypotézy
- v případě korelační analýzy p-hodnota analogicky odhaluje procento platnosti nulové hypotézy, tedy neexistujícího vztahu
- p-hodnotu získáme díky výpočtu například v Excelu

# Shrnutí

- mezi základní statistické nástroje patří korelační analýza a t-test
- jde o velmi jednoduché testy - při jejich interpretaci je třeba být o to více obezřetný
- odhalují totiž pouze vztah, nikoliv kauzalitu
- stejně jako v případě aplikace jiných testů je třeba naplnit jejich předpoklady