Mathtrace – príručka

V. Štatistika

Obsah

V. Štatistika

41. Štatistika

41. Štatistika

Dáta zapíšete ako data(…): hodnoty oddeľte bodkočiarkou, medzerou alebo novým riadkom, desatinná čiarka platí ako inde (data(1,5 2,5 3)). Tabuľku – riadok na jedno pozorovanie, stĺpec na premennú – napíšete po riadkoch oddelených bodkočiarkou, prvý riadok môže niesť názvy premenných. Pohodlnejšie: kláves s mriežkou na záložke f(x) otvorí tabuľku, kam dáta vyplníte, vložíte zo schránky (bunky z Excelu či Google Tabuliek) alebo načítate zo súboru CSV, a tlačidlom „Ako dáta“ ich vložíte pre štatistiku – aj tabuľku väčšiu, ako je mriežka.

Pri jednej premennej vyberiete čipom, čo vypočítať. Popisná štatistika: aritmetický priemer, medián, modus, kvartily (mediány dolnej a hornej polovice), variačné a medzikvartilové rozpätie, rozptyl a smerodajnú odchýlku výberovú (n − 1) aj populačnú (n), variačný koeficient a odľahlé hodnoty (mimo Q_1 − 1,5·IQR a Q_3 + 1,5·IQR); graf je krabicový. Početnosti a histogram: tabuľka absolútnych, relatívnych a kumulatívnych početností – pri celých číslach s málo hodnotami po hodnotách, inak po triedach (počet tried Sturgesovým pravidlom). z-skóre: každá hodnota štandardizovaná priemerom a smerodajnou odchýlkou dát, s T-skóre.

Dve premenné (dva stĺpce) aplikácia vypočíta ako koreláciu a regresiu: súčty, kovarianciu, Pearsonov korelačný koeficient r so slovným hodnotením sily závislosti, regresnú priamku metódou najmenších štvorcov a koeficient determinácie r^2; graf je bodový s priamkou. Pri troch a viacerých premenných vyjde popisná štatistika každej z nich vedľa seba s krabicovými grafmi.

Štatistická indukcia z dát: interval spoľahlivosti priemeru (čip „Interval spoľahlivosti priemeru“, spoľahlivosť 90, 95 alebo 99 %) cez t-rozdelenie; t-test (čip „t-test“) – jedna premenná proti testovanej hodnote μ_0, dve premenné ako nezávislé výbery (Welchov test, alebo s rovnakými rozptylmi) alebo párové hodnoty (rozdiely dvojíc); test χ^2 (čip „Test χ^2“) – jeden stĺpec početností ako test dobrej zhody s rovnomerným rozdelením, tabuľka ako test nezávislosti s Cramérovým V. Pod čipmi zadáte hladinu významnosti α, alternatívu (obojstranná, ľavostranná, pravostranná) a μ_0. Postup: hypotézy, testové kritérium, kritická hodnota a kritický obor, p-hodnota a rozhodnutie; graf ukáže rozdelenie s vyfarbeným kritickým oborom. Dva nezávislé výbery rôznej dĺžky zapíšte ako pomenované stĺpce – data(A = 23 25 28; B = 20 22 21 25).

Porovnanie skupín: čip „Poradový test (Mann–Whitney, Wilcoxon, Kruskal–Wallis)“ porovná dve skupiny Mannovým–Whitneyho testom (s voľbou „Párové“ párové hodnoty Wilcoxonovým testom) a tri a viac skupín Kruskalovým–Wallisovým testom – podľa poradia hodnôt, bez predpokladu normálneho rozdelenia; pri malých vzorkách bez zhôd je p-hodnota presná. Čip „ANOVA (analýza rozptylu)“ porovná priemery skupín jednofaktorovou analýzou rozptylu s tabuľkou ANOVA, F-testom a veľkosťou účinku η^2. Skupiny zapíšte do stĺpcov, môžu mať rôznu dĺžku (pomenované stĺpce); graf ukáže krabicové grafy skupín a rozdelenie testového kritéria. Pri teste χ^2 jedného stĺpca početností zadáte do poľa „Očakávané podiely“ vlastné očakávané podiely – 1:2:1, 0,25 0,5 0,25 alebo v percentách.

Regresia dvoch premenných má pod čipom voľbu modelu: lineárna (s testom významnosti korelácie), kvadratická (normálne rovnice metódy najmenších štvorcov), exponenciálna y = a·e^(bx), mocninová y = a·x^b a logaritmická y = a + b·ln x (linearizáciou cez logaritmy) – s koeficientom determinácie R^2 a grafom preloženej krivky. Vo formulári „Štatistika a psychometria“ sú navyše interval spoľahlivosti a t-testy zo súhrnných hodnôt (rozsah, priemer, smerodajná odchýlka) – aj z-test pri známom σ a t-test dvoch výberov.

Normálne rozdelenie nájdete vo formulári „Štatistika a psychometria“ (sekcia „Aplikácie“ na úvodnej obrazovke, odbor Štatistika): pravdepodobnosť P(X < x), P(X > x), P(a < X < b) a kvantil – hodnota, pod ktorou leží zadaný podiel. Postup vedie cez štandardizáciu z = (x − μ)/σ a distribučnú funkciu Φ, záporné z ako 1 − Φ(|z|) podľa tabuliek; graf ukáže krivku hustoty s vyfarbenou plochou.

Ako zadať

`data(3; 5; 7; 7; 9)`	Jedna premenná – hodnoty bodkočiarkou, medzerou alebo pod sebou.
`data(x y; 1 52; 2 58; 3 61)`	Tabuľka po riadkoch, prvý riadok s názvami premenných (nepovinný).
`data(x = 1 2 3 4; y = 2 4 5 4)`	Pomenované stĺpce, každý so svojimi hodnotami.

Ukážky

Popisná štatistika ôsmich hodnôt.

Zadanie: `data(2; 4; 4; 4; 5; 5; 7; 9)`

Výsledok

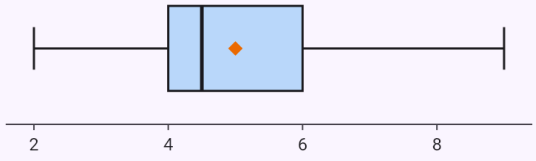
$$\bar{x} = 5$$
$$\tilde{x} = 4{,}5$$
$$s \approx 2{,}138$$

Postup

- Počet hodnôt $n = 8$, zoradené dáta:
  $2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9$
- Aritmetický priemer: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{40}{8} = 5$
- Medián je prostredná hodnota zoradených dát:
  $\tilde{x} = \frac{x_{(4)} + x_{(5)}}{2} = \frac{4+5}{2} = 4{,}5$
- Modus (najčastejšia hodnota, 3×): $\hat{x} = 4$
- Kvartily sú mediány dolnej a hornej polovice dát:
  $Q_1 = 4, \quad Q_2 = \tilde{x} = 4{,}5, \quad Q_3 = 6$
- Variačné rozpätie a medzikvartilové rozpätie
  $R = x_{max} - x_{min} = 9 - 2 = 7$
  $IQR = Q_3 - Q_1 = 6 - 4 = 2$
- Súčet štvorcov odchýlok od priemeru:
  $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = 32$
- Rozptyl – výberový s² a populačný σ²
  $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{32}{7} \approx 4{,}571$
  $\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{32}{8} = 4$
- Smerodajná odchýlka
  $s = \sqrt{s^2} = \sqrt{4{,}571} \approx 2{,}138$
  $\sigma = \sqrt{4} = 2$
- Variačný koeficient: $v = \frac{s}{\bar{x}} = \frac{2{,}138}{5} \approx 0{,}4276 = 42{,}8\,\%$
- Hradby pre odľahlé hodnoty
  $Q_1 - 1{,}5\,IQR = 1, \quad Q_3 + 1{,}5\,IQR = 9$ – žiadna hodnota mimo nich neleží.
- Výberový rozptyl s² delí n − 1 a odhaduje rozptyl celej populácie z výberu – obvyklá voľba. Populačný σ² delí n a použijeme ho, len keď dáta sú celá populácia.
- Pozor: dát je málo, odhady (hlavne rozptylu a kvartilov) sú neisté.

Graf

Krabicový graf (medián, kvartily, odľahlé hodnoty; ◆ priemer)



Odľahlá hodnota 31 v krabicovom grafe.

Zadanie: `data(12 15 14 13 16 15 14 31)`

Výsledok

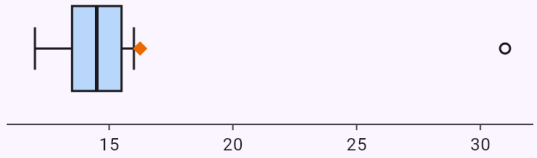
$$\bar{x} = 16{,}25$$
$$\tilde{x} = 14{,}5$$
$$s \approx 6{,}089$$

Postup

- Počet hodnôt $n = 8$, zoradené dáta:
  $12,\ 13,\ 14,\ 14,\ 15,\ 15,\ 16,\ 31$
- Aritmetický priemer: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{130}{8} = 16{,}25$
- Medián je prostredná hodnota zoradených dát:
  $\tilde{x} = \frac{x_{(4)} + x_{(5)}}{2} = \frac{14 + 15}{2} = 14{,}5$
- Modus (najčastejšia hodnota, 2×): $\hat{x} = 14,\ 15$
- Kvartily sú mediány dolnej a hornej polovice dát:
  $Q_1 = 13{,}5, \quad Q_2 = \tilde{x} = 14{,}5, \quad Q_3 = 15{,}5$
- Variačné rozpätie a medzikvartilové rozpätie
  $R = x_{max} - x_{min} = 31 - 12 = 19$
  $IQR = Q_3 - Q_1 = 15{,}5 - 13{,}5 = 2$
- Súčet štvorcov odchýlok od priemeru:
  $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = 259{,}5$
- Rozptyl – výberový s² a populačný σ²
  $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{259{,}5}{7} \approx 37{,}07$
  $\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{259{,}5}{8} \approx 32{,}44$
- Smerodajná odchýlka
  $s = \sqrt{s^2} = \sqrt{37{,}07} \approx 6{,}089$
  $\sigma = \sqrt{32{,}44} \approx 5{,}695$
- Variačný koeficient: $v = \frac{s}{\bar{x}} = \frac{6{,}089}{16{,}25} \approx 0{,}3747 = 37{,}5\,\%$
- Odľahlé hodnoty ležia mimo hradieb
  $Q_1 - 1{,}5\,IQR = 10{,}5, \quad Q_3 + 1{,}5\,IQR = 18{,}5$: 31
- Výberový rozptyl s² delí n − 1 a odhaduje rozptyl celej populácie z výberu – obvyklá voľba. Populačný σ² delí n a použijeme ho, len keď dáta sú celá populácia.
- Pozor: dát je málo, odhady (hlavne rozptylu a kvartilov) sú neisté.

Graf

Krabicový graf (medián, kvartily, odľahlé hodnoty; ◆ priemer)



Tabuľka početností známok a histogram.

Zadanie: `data(1 2 2 3 3 3 3 4 4 5 1 3 2 4 3)` Početnosti a histogram

Výsledok

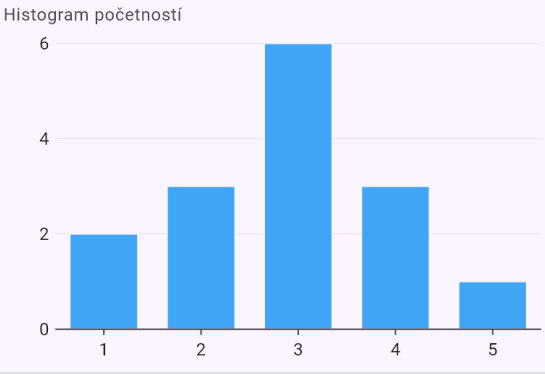
$$n = 15$$
$$\hat{x} = 3$$

Postup

- Počet hodnôt $n = 15$, zoradené dáta:
  $1,\ 1,\ 2,\ 2,\ 2,\ 3,\ 3,\ 3,\ 3,\ 3,\ 3,\ 4,\ 4,\ 4,\ 5$
- Tabuľka početností – absolútna n, relatívna f, kumulatívna N a F

x_i	n_i	f_i	N_i	F_i
1	2	0,133	2	0,133
2	3	0,200	5	0,333
3	6	0,400	11	0,733
4	3	0,200	14	0,933
5	1	0,067	15	1,000
$\sum$	15	1		

Graf



Spojité dáta rozdelené do tried.

Zadanie: `data(12.5 14.1 15.8 17.2 18.9 21.3 22.0 23.7 25.1 26.4 28.8 31.2)`
Početnosti a histogram

Výsledok

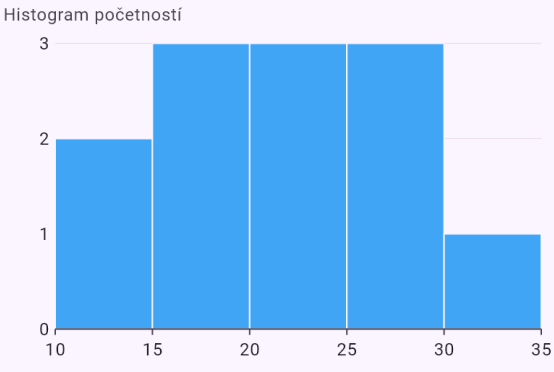
$$n = 12$$
$$k = 5, \quad h = 5$$

Postup

- Počet hodnôt $n = 12$, zoradené dáta
  $12{,}5,\ 14{,}1,\ 15{,}8,\ 17{,}2,\ 18{,}9,\ 21{,}3,\ 22,\ 23{,}7,\ 25{,}1,\ 26,$
- Hodnôt je veľa, rozdelíme ich do 5 tried (Sturgesovo pravidlo)
  $k \approx 1 + \log_2 n = 1 + \log_2 12 \approx 4{,}585 \Rightarrow k = 5$
  $h \approx \frac{R}{k} = \frac{18{,}7}{5} \approx 3{,}74 \Rightarrow h = 5$
- Tabuľka početností tried (x je stred triedy)

I_j	x_j	n_j	f_j	N_j	F_j
$\langle 10;\ 15)$	12,5	2	0,167	2	0,167
$\langle 15;\ 20)$	17,5	3	0,250	5	0,417
$\langle 20;\ 25)$	22,5	3	0,250	8	0,667
$\langle 25;\ 30)$	27,5	3	0,250	11	0,917
$\langle 30;\ 35\rangle$	32,5	1	0,083	12	1,000
$\sum$		12	1		

Graf



z-skóre a T-skóre hodnôt.

Zadanie: `data(2 4 4 4 5 5 7 9)` z-skóre

Výsledok

$$\bar{x} = 5$$
$$s \approx 2{,}138$$

Postup

- Počet hodnôt $n = 8$, zoradené dáta: $2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9$
- Aritmetický priemer: $\bar{x} = \dfrac{40}{8} = 5$
- Smerodajná odchýlka

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} = \sqrt{\frac{32}{7}} \approx 2{,}138$$

- Každú hodnotu štandardizujeme

$$z_i = \frac{x_i - \bar{x}}{s} = \frac{x_i - 5}{2{,}138}$$
$$T_i = 50 + 10 z_i$$

- z-skóre a T-skóre hodnôt

x_i	z_i	T_i
2	$-1{,}40$	$36{,}0$
4	$-0{,}47$	$45{,}3$
4	$-0{,}47$	$45{,}3$
4	$-0{,}47$	$45{,}3$
5	$0{,}00$	$50{,}0$
5	$0{,}00$	$50{,}0$
7	$0{,}94$	$59{,}4$
9	$1{,}87$	$68{,}7$

- z-skóre sú vypočítané z priemeru a smerodajnej odchýlky týchto dát. Na porovnanie s populáciou sa použijú priemer a SD normy (model „Prevody štandardných skóre“).

Korelácia a regresná priamka – počet hodín učenia a body v teste.

Zadanie: `data(x y; 1 52; 2 58; 3 61; 4 70; 5 74; 6 79)`

Výsledok

$$r \approx 0{,}9937$$

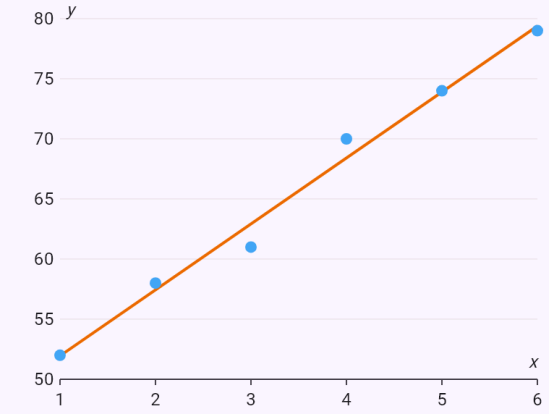
$$\hat{y} = 46{,}47 + 5{,}486x$$

Postup

- Počet hodnôt $n = 6$, zoradené dáta
  $[1;\,52],\ [2;\,58],\ [3;\,61],\ [4;\,70],\ [5;\,74],\ [6;\,79]$
- Súčty a priemery
  $\sum x_i = 21$
  $\sum y_i = 394$
  $\sum x_i^2 = 91$
  $\sum y_i^2 = 26\,410$
  $\sum x_i y_i = 1475$
  $\bar{x} = 3{,}5$
  $\bar{y} = 65{,}67$
- Smerodajná odchýlka
  $$s_x = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} \approx 1{,}871$$
  $$s_y = \sqrt{\frac{\sum y_i^2 - n\bar{y}^2}{n-1}} \approx 10{,}33$$
- Kovariancia
  $$s_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{1475 - 6 \cdot 3{,}5 \cdot 65{,}67}{5} = 19{,}2$$
- Pearsonov korelačný koeficient:
  $r = \frac{s_{xy}}{s_x s_y} = \frac{19{,}2}{1{,}871 \cdot 10{,}33} \approx 0{,}9937$ – závislosť je veľmi silná, kladná (s rastúcim x rastie y).
- Koeficienty regresnej priamky metódou najmenších štvorcov
  $b = \frac{s_{xy}}{s_x^2} = \frac{19{,}2}{3{,}5} \approx 5{,}486$
  $a = \bar{y} - b\bar{x} = 65{,}67 - 5{,}486 \cdot 3{,}5 \approx 46{,}47$
- Regresná priamka: $\hat{y} = 46{,}47 + 5{,}486x$
- Koeficient determinácie $r^2 = 0{,}9937^2 \approx 0{,}9874$ – priamka vysvetľuje $98{,}7\,\%$ rozptylu y.
- Test nulovej hypotézy ρ = 0 (t-rozdelenie s 4 stupňami voľnosti) – korelácia je štatisticky významná.
  $$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}9937\sqrt{4}}{\sqrt{1-0{,}9937^2}} \approx 17{,}73$$
  $p < 0{,}0001 \le \alpha = 0{,}05$
- Korelácia opisuje len lineárnu súvislosť: nehovorí, čo je príčinou, a nemusí zachytiť súvislosť nelineárnu. Regresnú priamku je rozumné používať len v rozsahu nameraných x.

Graf

Bodový graf s regresnou priamkou



Tri premenné vedľa seba.

Zadanie: `data(A B C; 12 15 9; 14 18 11; 11 16 10; 15 19 12; 13 17 10)`

Výsledok

$$\bar{x}_{\mathrm{A}} = 13$$
$$\bar{x}_{\mathrm{B}} = 17$$
$$\bar{x}_{\mathrm{C}} = 10{,}4$$

Postup

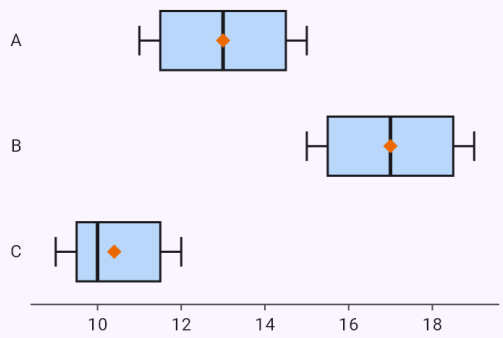
- Popisná štatistika jednotlivých premenných

	A	B	C
n	5	5	5
$\bar{x}$	13	17	10,4
$\tilde{x}$	13	17	10
Q_1	11,5	15,5	9,5
Q_3	14,5	18,5	11,5
s	1,581	1,581	1,14
σ	1,414	1,414	1,02
x_{min}	11	15	9
x_{max}	15	19	12

- Výberový rozptyl s² delí n − 1 a odhaduje rozptyl celej populácie z výberu – obvyklá voľba. Populačný σ² delí n a použijeme ho, len keď dáta sú celá populácia.

Graf

Krabicový graf (medián, kvartily, odľahlé hodnoty; ◆ priemer)



95% interval spoľahlivosti priemeru desiatich meraní.

Zadanie: `data(102 98 110 105 99 107 101 104 96 108)` Interval spoľahlivosti priemeru

Výsledok

$$95\,\%: \ \langle 99{,}71;\ 106{,}3 \rangle$$

Postup

- Rozsah, priemer a výberová smerodajná odchýlka:
  $n = 10, \quad \bar{x} = 103, \quad s \approx 4{,}595$
- Smerodajná chyba: $SE = \frac{s}{\sqrt{n}} = \frac{4{,}595}{\sqrt{10}} \approx 1{,}453$
- Kvantil pre zvolenú spoľahlivosť: $t_{0{,}975}(9) \approx 2{,}262$
- Interval spoľahlivosti
  $\bar{x} \pm t \cdot SE = 103 \pm 2{,}262 \cdot 1{,}453 = 103 \pm 3{,}287 \Rightarrow \langle 99$
- Spoľahlivosť 95 % znamená: keby sme výber mnohokrát opakovali, asi 95 % takto vypočítaných intervalov by skutočný priemer μ obsahovalo. Neznamená to, že μ leží v tomto konkrétnom intervale s pravdepodobnosťou 95 %.
- Pre malý výber (n < 30) t-rozdelenie predpokladá, že dáta pochádzajú z približne normálneho rozdelenia. Pri výrazne zošikmených dátach alebo odľahlých hodnotách je vhodnejší neparametrický test.

Líši sa priemer od 100? t-test jedného výberu.

Zadanie: `data(102 98 110 105 99 107 101 104 96 108)` t-test

Výsledok

$$t \approx 2{,}065$$

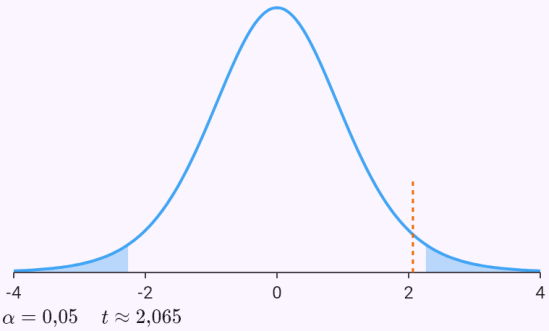
$$p \approx 0{,}06895 > \alpha = 0{,}05$$

Postup

- Rozsah, priemer a výberová smerodajná odchýlka: $n = 10, \quad \bar{x} = 103, \quad s \approx 4{,}595$
- Nulová a alternatívna hypotéza: $H_0: \mu = 100, \quad H_1: \mu \neq 100$
- Smerodajná chyba: $SE = \frac{s}{\sqrt{n}} = \frac{4{,}595}{\sqrt{10}} \approx 1{,}453$
- Testové kritérium – pri platnosti H_0 má t-rozdelenie s 9 stupňami voľnosti: $t = \frac{\bar{x} - \mu_0}{SE} = \frac{103 - 100}{1{,}453} \approx 2{,}065$
- Kritická hodnota na hladine α: $t_{0{,}975}(9) \approx 2{,}262$, kritický obor $|t| \geq 2{,}262$
- p-hodnota: $p = 2\,P(T \geq 2{,}065) \approx 0{,}06895$
- Keďže $p \approx 0{,}06895 > \alpha = 0{,}05$, nulovú hypotézu nezamietame – dáta neposkytujú dosť dôkazov proti nej.
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Pre malý výber (n < 30) t-rozdelenie predpokladá, že dáta pochádzajú z približne normálneho rozdelenia. Pri výrazne zošikmených dátach alebo odľahlých hodnotách je vhodnejší neparametrický test.

Graf

t-rozdelenie – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



$\alpha = 0{,}05 \quad t \approx 2{,}065$

Dva nezávislé výbery rôznej veľkosti – Welchov t-test.

Zadanie: `data(A = 23 25 28 30 26 27; B = 20 22 21 25 19 24 23)` t-test

Výsledok

$$t \approx 3{,}503$$

$$p \approx 0{,}005546 \leq \alpha = 0{,}05$$

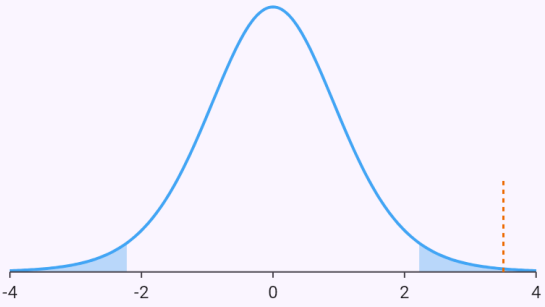
Postup

- Rozsah, priemer a výberová smerodajná odchýlka
  $n_1 = 6,\ \bar{x}_1 = 26{,}5,\ s_1 \approx 2{,}429$
  $n_2 = 7,\ \bar{x}_2 = 22,\ s_2 \approx 2{,}16$
- Nulová a alternatívna hypotéza:
  $H_0: \mu_1 = \mu_2, \quad H_1: \mu_1 \neq \mu_2$
- Smerodajná chyba
  $$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{2{,}429^2}{6} + \frac{2{,}16^2}{7}} \approx 1{,}285$$
- Stupne voľnosti podľa Welcha a Satterthwaitea
  $$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \approx 10{,}18$$
- Testové kritérium – pri platnosti H_0 má t-rozdelenie s $10{,}18$ stupňami voľnosti:
  $$t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = \frac{26{,}5 - 22}{1{,}285} \approx 3{,}503$$
- Kritická hodnota na hladine α: $t_{0{,}975}(10{,}18) \approx 2{,}223$, kritický obor $|t| \geq 2{,}223$
- p-hodnota: $p = 2\,P(T \geq 3{,}503) \approx 0{,}005546$
- Keďže $p \approx 0{,}005546 \leq \alpha = 0{,}05$, nulovú hypotézu zamietame – výsledok je štatisticky významný.

- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Welchov test nepredpokladá zhodné rozptyly oboch skupín; jeho stupne voľnosti v nemusia byť celé číslo.

Graf

t-rozdelenie – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



$\alpha = 0{,}05 \quad t \approx 3{,}503$

Párový t-test – meranie pred a po.

Zadanie: `data(x1 x2; 72 70; 75 73; 68 69; 80 76; 77 74; 74 71)` t-test · Párové

Výsledok

$$t \approx 3{,}081$$

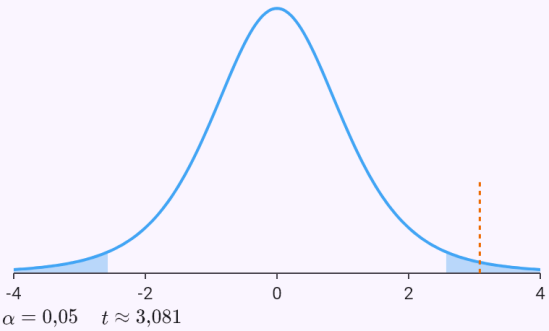
$$p \approx 0{,}02743 \leq \alpha = 0{,}05$$

Postup

- Rozdiely dvojíc d_i (1. stĺpec − 2. stĺpec): $2,\ 2,\ -1,\ 4,\ 3,\ 3;\ n = 6, \quad \bar{d} \approx 2{,}167, \quad s_d \approx 1{,}722$
- Nulová a alternatívna hypotéza: $H_0: \mu_d = 0, \quad H_1: \mu_d \neq 0$
- Smerodajná chyba: $SE = \frac{s}{\sqrt{n}} = \frac{1{,}722}{\sqrt{6}} \approx 0{,}7032$
- Testové kritérium – pri platnosti H_0 má t-rozdelenie s 5 stupňami voľnosti: $t = \frac{\bar{d} - \mu_0}{SE} = \frac{2{,}167 - 0}{0{,}7032} \approx 3{,}081$
- Kritická hodnota na hladine α: $t_{0{,}975}(5) \approx 2{,}571$, kritický obor $|t| \geq 2{,}571$
- p-hodnota: $p = 2\,P(T \geq 3{,}081) \approx 0{,}02743$
- Keďže $p \approx 0{,}02743 \leq \alpha = 0{,}05$, nulovú hypotézu zamietame – výsledok je štatisticky významný.
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Pre malý výber (n < 30) t-rozdelenie predpokladá, že dáta pochádzajú z približne normálneho rozdelenia. Pri výrazne zošikmených dátach alebo odľahlých hodnotách je vhodnejší neparametrický test.
- Párový t-test je t-test jedného výberu pre rozdiely (1. stĺpec − 2. stĺpec) – každá dvojica patrí k rovnakému subjektu (napr. pred a po).

Graf

t-rozdelenie – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



Je kocka spravodlivá? Test dobrej zhody χ².

Zadanie: `data(8 12 9 11 6 14)` Test χ²

Výsledok

$$\chi^2 = 4{,}2$$
$$p \approx 0{,}521 > \alpha = 0{,}05$$

Postup

- Nulová hypotéza: všetkých 6 kategórií je rovnako pravdepodobných ($H_0: p_1 = \cdots = p_6 = \frac{1}{6}$), alternatívna: aspoň jedna sa líši.
- Očakávané početnosti pri platnosti H_0 $E_i = \frac{N}{k} = \frac{60}{6} = 10$ a príspevky jednotlivých kategórií

i	O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
1	8	10	0,4
2	12	10	0,4
3	9	10	0,1
4	11	10	0,1
5	6	10	1,6
6	14	10	1,6
$\sum$	60	60	4,2

- Testové kritérium – pri platnosti H_0 má rozdelenie χ² s 5 stupňami voľnosti

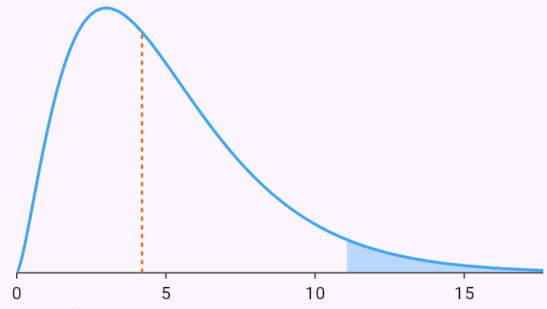
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 0{,}4 + 0{,}4 + 0{,}1 + 0{,}1 + 1{,}6 + 1$$

- Kritická hodnota na hladine α: $\chi^2_{0{,}95}(5) \approx 11{,}07$, kritický obor $\chi^2 \geq 11{,}07$
- p-hodnota: $p = P(\chi^2 \geq 4{,}2) \approx 0{,}521$
- Keďže $p \approx 0{,}521 > \alpha = 0{,}05$, nulovú hypotézu nezamietame – dáta neposkytujú dosť dôkazov proti nej.

- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Test χ² pracuje s početnosťami (počtami prípadov), nie s percentami alebo priemermi.

Graf

Rozdelenie χ² – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



Test nezávislosti v tabuľke 2×2.

Zadanie: `data(20 30; 25 15)` Test χ²

Výsledok

$$\chi^2 = 4{,}5$$
$$p \approx 0{,}03389 \leq \alpha = 0{,}05$$
$$V \approx 0{,}2236$$

Postup

- Nulová hypotéza: riadková a stĺpcová premenná sú nezávislé, alternatívna: sú závislé.
- Očakávané početnosti pri nezávislosti $E_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{N}$, v zátvorke pozorované

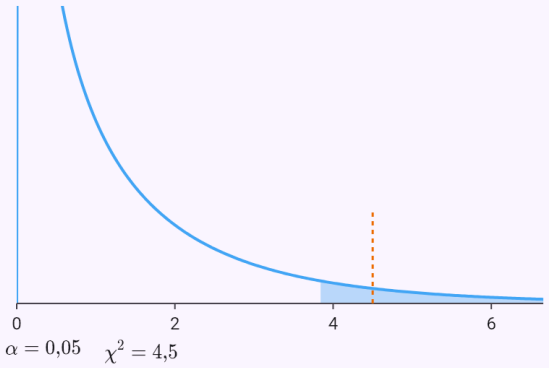
	1	2	∑
1	25 (20)	25 (30)	50
2	20 (25)	20 (15)	40
∑	45	45	90

- Testové kritérium – pri platnosti H_0 má rozdelenie χ^2 s 1 stupňami voľnosti:
  $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 1 + 1 + 1{,}25 + 1{,}25 = 4{,}5$
- Kritická hodnota na hladine α: $\chi^2_{0{,}95}(1) \approx 3{,}841$, kritický obor $\chi^2 \geq 3{,}841$
- p-hodnota: $p = P(\chi^2 \geq 4{,}5) \approx 0{,}03389$
- Keďže $p \approx 0{,}03389 \leq \alpha = 0{,}05$, nulovú hypotézu zamietame – výsledok je štatisticky významný.
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.

- Sila závislosti – Cramérovo V (0 žiadna, 1 úplná)
  $V = \sqrt{\frac{\chi^2}{N(\min(r,c)-1)}} = \sqrt{\frac{4{,}5}{90 \cdot 1}} \approx 0{,}2236$
- Test χ² pracuje s početnosťami (počtami prípadov), nie s percentami alebo priemermi.

Graf

Rozdelenie χ² – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



Dve skupiny Mannovým–Whitneyho testom – malé vzorky, presná p-hodnota.

Zadanie: `data(A = 23 25 28 30 26 27; B = 20 22 21 25 19 24 23)`

Poradový test (Mann–Whitney, Wilcoxon, Kruskal–Wallis)

Výsledok

$$U = 3$$
$$z \approx -2{,}579$$
$$p \approx 0{,}009922 \leq \alpha = 0{,}05$$

Postup

- Nulová hypotéza: obe skupiny pochádzajú z rovnakého rozdelenia, alternatívna: hodnoty jednej bývajú systematicky vyššie

  $H_0: F_A = F_B$
  $H_1: F_A \neq F_B$

- Hodnoty oboch skupín spolu zoradíme a očíslujeme poradím R (rovnakým hodnotám priemer ich poradí)

x	R	
19	1	B
20	2	B
21	3	B
22	4	B
23	5,5	A
23	5,5	B
24	7	B
25	8,5	A
25	8,5	B
26	10	A
27	11	A
28	12	A
30	13	A

- Rozsahy skupín a súčty ich poradí:
  $n_1 = 6, \quad n_2 = 7, \quad R_1 = 60, \quad R_2 = 31$

- Mannova–Whitneyho štatistika U – počet dvojíc, v ktorých má hodnota jednej skupiny vyššie poradie ako druhej; testujeme menšiu z oboch

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 60 - \frac{6 \cdot 7}{2} = 39$$
$$U_2 = n_1 n_2 - U_1 = 6 \cdot 7 - 39 = 3$$
$$U = \min(U_1, U_2) = 3$$

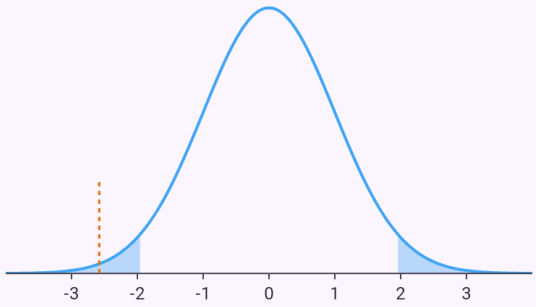
- Testové kritérium – pri platnosti H_0 má normované normálne rozdelenie:

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sigma_U} = \frac{3 - 21}{6{,}981} \approx -2{,}579$$

- Kritická hodnota na hladine α: $z_{0{,}975} \approx 1{,}96$, kritický obor $|z| \geq 1{,}96$
- p-hodnota: $p = 2\,P(Z \geq 2{,}579) \approx 0{,}009922$
- Keďže $p \approx 0{,}009922 \leq \alpha = 0{,}05$, nulovú hypotézu zamietame – výsledok je štatisticky významný.
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Veľkosť účinku r (orientačne 0,1 malá, 0,3 stredná, 0,5 veľká): $r = \frac{|z|}{\sqrt{N}} = \frac{2{,}579}{\sqrt{13}} \approx 0{,}7152$
- Mannov–Whitneyho test je poradový test: nepredpokladá normálne rozdelenie, hodí sa aj pre poradové dáta a odľahlé hodnoty. Porovnáva celé rozdelenie skupín – keď majú podobný tvar, je to test rozdielu mediánov.

Graf

Normované normálne rozdelenie – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



$\alpha = 0{,}05 \quad z \approx -2{,}579$

Párové hodnoty pred a po Wilcoxonovým testom.

Zadanie: `data(pred po; 72 70; 75 73; 68 69; 80 76; 77 74; 74 71; 79 72)`

Poradový test (Mann–Whitney, Wilcoxon, Kruskal–Wallis) · Párové

Výsledok

$$T = 1$$
$$z \approx -2{,}205$$
$$p \approx 0{,}02743 \leq \alpha = 0{,}05$$

Postup

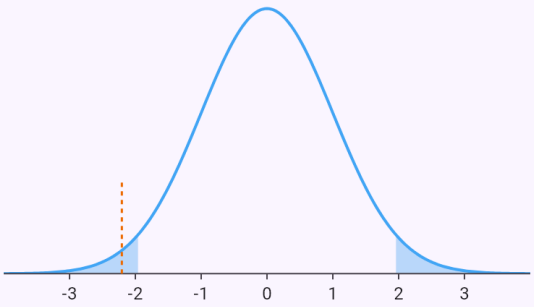
- Nulová hypotéza: medián rozdielov je nulový (žiadna zmena), alternatívna: nie je: $H_0: \tilde{d} = 0, \quad H_1: \tilde{d} \neq 0$
- Rozdiely dvojíc d_i (1. stĺpec − 2. stĺpec): $2,\ 2,\ -1,\ 4,\ 3,\ 3,\ 7$; nulové rozdiely vynecháme: $n = 7$
- Nenulové rozdiely zoradíme podľa veľkosti (bez znamienka) a očíslujeme poradím R; znamienko zostáva pri poradí

d_i	R	
-1	1	$-$
2	$2{,}5$	$+$
2	$2{,}5$	$+$
3	$4{,}5$	$+$
3	$4{,}5$	$+$
4	6	$+$
7	7	$+$

- Súčet poradí kladných a záporných rozdielov
  $T^+ = 27$
  $T^- = 1$
  $T = \min\left(T^+, T^-\right) = 1$
- Testové kritérium − pri platnosti H_0 má normované normálne rozdelenie:
  $$z = \frac{T - \frac{n(n+1)}{4}}{\sigma_T} = \frac{1 - 14}{5{,}895} \approx -2{,}205$$
- Kritická hodnota na hladine α: $z_{0{,}975} \approx 1{,}96$, kritický obor $|z| \geq 1{,}96$
- p-hodnota: $p = 2\,P(Z \geq 2{,}205) \approx 0{,}02743$
- Keďže $p \approx 0{,}02743 \leq \alpha = 0{,}05$, nulovú hypotézu zamietame − výsledok je štatisticky významný.
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí − nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Veľkosť účinku r (orientačne 0,1 malá, 0,3 stredná, 0,5 veľká): $r = \frac{|z|}{\sqrt{n}} = \frac{2{,}205}{\sqrt{7}} \approx 0{,}8335$
- Wilcoxonov test je poradová obdoba párového t-testu: nepredpokladá normálne rozdelenie rozdielov, berie do úvahy ich znamienka a veľkosť poradím.

Graf

Normované normálne rozdelenie − vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



$\alpha = 0{,}05 \quad z \approx -2{,}205$

Tri skupiny Kruskalovým–Wallisovým testom.

Zadanie: `data(A = 12 15 14 11 13; B = 16 18 17 15 19; C = 13 14 12 16 15)`

Poradový test (Mann–Whitney, Wilcoxon, Kruskal–Wallis)

Výsledok

$$H \approx 8{,}263$$
$$p \approx 0{,}01606 \leq \alpha = 0{,}05$$

Postup

- Nulová hypotéza: všetky skupiny pochádzajú z rovnakého rozdelenia, alternatívna: aspoň jedna sa líši
  $H_0: F_1 = \cdots = F_3$
  $H_1: \exists\, i, j: F_i \neq F_j$

- Hodnoty všetkých skupín spolu zoradíme a očíslujeme poradím R (rovnakým hodnotám priemer ich poradí)

x	R	
11	1	A
12	2,5	A
12	2,5	C
13	4,5	A
13	4,5	C
14	6,5	A
14	6,5	C
15	9	A
15	9	B
15	9	C
16	11,5	B
16	11,5	C
17	13	B
18	14	B
19	15	B

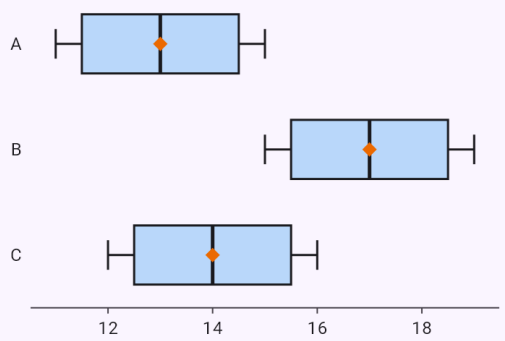
- Súčty poradí jednotlivých skupín
  $R_{\mathrm{A}} = 23{,}5\ (n = 5)$
  $R_{\mathrm{B}} = 62{,}5\ (n = 5)$
  $R_{\mathrm{C}} = 34\ (n = 5)$

- Testové kritérium – pri platnosti H_0 má rozdelenie χ^2 s 2 stupňami voľnosti
  $$H = \frac{12}{N(N+1)} \sum \frac{R_j^2}{n_j} - 3(N+1) = \frac{12}{15 \cdot 16}\left(\frac{23{,}5^2}{5}\right.$$
  $$C = 1 - \frac{\sum (t^3 - t)}{N^3 - N} = 1 - \frac{48}{3360} \approx 0{,}9857$$
  $$H_c = \frac{H}{C} = \frac{8{,}145}{0{,}9857} \approx 8{,}263$$

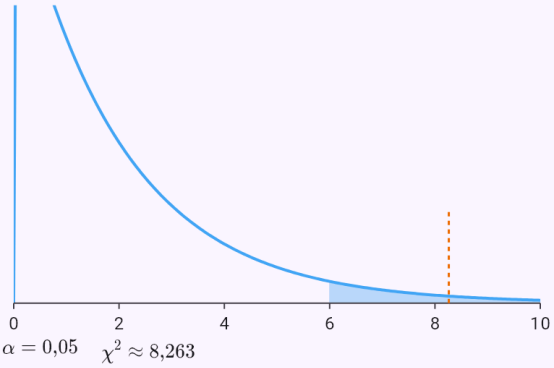
- Kritická hodnota na hladine α: $\chi^2_{0{,}95}(2) \approx 5{,}991$, kritický obor $\chi^2 \geq 5{,}991$
- p-hodnota: $p = P(\chi^2 \geq 8{,}263) \approx 0{,}01606$
- Keďže $p \approx 0{,}01606 \leq \alpha = 0{,}05$, nulovú hypotézu zamietame – výsledok je štatisticky významný.
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Veľkosť účinku ε² – podiel rozptylu poradí vysvetlený skupinami: $\varepsilon^2 = \frac{H}{N-1} = \frac{8{,}263}{14} \approx 0{,}5902$
- Kruskalov–Wallisov test je poradová obdoba jednofaktorovej ANOVY. Hovorí len, že sa niektoré skupiny líšia – ktoré, ukáže až porovnanie dvojíc (Mannovým–Whitneyho testom s Bonferroniho korekciou).

Graf

Krabicový graf (medián, kvartily, odľahlé hodnoty; ◆ priemer)



Rozdelenie χ² – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



Priemery troch skupín jednofaktorovou ANOVOU.

Zadanie: `data(A = 12 15 14 11 13; B = 16 18 17 15 19; C = 13 14 12 16 15)`

ANOVA (analýza rozptylu)

Výsledok

$$F \approx 8{,}667$$
$$p \approx 0{,}004687 \leq \alpha = 0{,}05$$
$$\eta^2 \approx 0{,}5909$$

Postup

- Nulová hypotéza: priemery všetkých skupín sú rovnaké, alternatívna: aspoň dva sa líšia

  $H_0: \mu_1 = \cdots = \mu_3$

  $H_1: \exists\, i, j: \mu_i \neq \mu_j$

- Rozsahy, priemery a smerodajné odchýlky skupín a celkový priemer $\bar{x} \approx 14{,}67$

	n_j	$\bar{x}_j$	s_j
A	5	13	1,581
B	5	17	1,581
C	5	14	1,581
$\sum$	15	14,67	

- Analýza rozptylu – súčet štvorcov medzi skupinami (B) a vnútri skupín (W), stupne voľnosti a priemerné štvorce

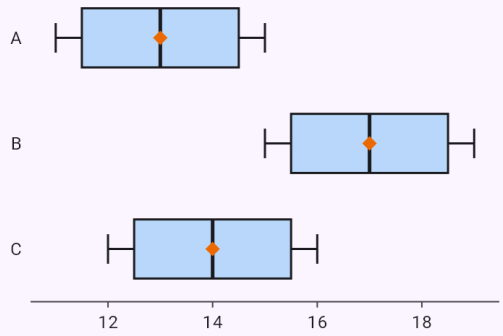
  $$SS_B = \sum n_j (\bar{x}_j - \bar{x})^2$$
  $$SS_W = \sum \sum (x_{ij} - \bar{x}_j)^2$$

	SS	df	MS	F
B	43,33	2	21,67	8,667
W	30	12	2,5	
$\sum$	73,33	14		

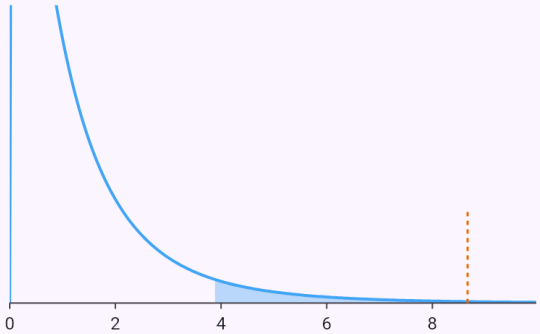
- Testové kritérium – pri platnosti H_0 má F-rozdelenie s 2 a 12 stupňami voľnosti: $F = \frac{MS_B}{MS_W} = \frac{21{,}67}{2{,}5} \approx 8{,}667$
- Kritická hodnota na hladine α: $F_{0{,}95}(2;\, 12) \approx 3{,}885$, kritický obor $F \geq 3{,}885$
- p-hodnota: $p = P(F \geq 8{,}667) \approx 0{,}004687$
- Keďže $p \approx 0{,}004687 \leq \alpha = 0{,}05$, nulovú hypotézu zamietame – výsledok je štatisticky významný.
- Veľkosť účinku η² – podiel rozptylu vysvetlený rozdielmi medzi skupinami (orientačne 0,01 malá, 0,06 stredná, 0,14 veľká): $\eta^2 = \frac{SS_B}{SS_T} = \frac{43{,}33}{73{,}33} \approx 0{,}5909$
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- ANOVA predpokladá nezávislé skupiny, približne normálne rozdelenie v každej z nich a podobné rozptyly. Keď to neplatí (malé skupiny, odľahlé hodnoty, poradové dáta), použite Kruskalov–Wallisov test.
- Zamietnutie H_0 hovorí len, že sa aspoň dva priemery líšia. Ktoré, ukáže až porovnanie dvojíc (post hoc testy, napr. t-testy s Bonferroniho korekciou).

Graf

Krabicový graf (medián, kvartily, odľahlé hodnoty; ◆ priemer)



F-rozdelenie – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



$\alpha = 0{,}05$ $F \approx 8{,}667$

Test dobrej zhody s očakávanými podielmi 1 : 2 : 1.

Zadanie: `data(30 50 20)` Test χ² · Očakávané podiely 1:2:1

Výsledok

$$\chi^2 = 2$$
$$p \approx 0{,}3679 > \alpha = 0{,}05$$

Postup

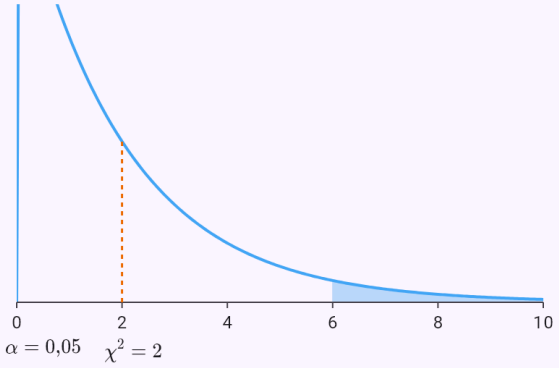
- Nulová hypotéza: kategórie majú zadané podiely, alternatívna: aspoň jeden podiel sa líši:
  $H_0: p_1 = 0{,}25,\ p_2 = 0{,}5,\ p_3 = 0{,}25$
- Očakávané početnosti podľa zadaných podielov $E_i = N \cdot p_i, \quad N = 100$ a príspevky kategórií

i	O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
1	30	25	1
2	50	50	0
3	20	25	1
$\sum$	100	100	2

- Testové kritérium – pri platnosti H_0 má rozdelenie χ^2 s 2 stupňami voľnosti:
  $$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 1 + 0 + 1 = 2$$
- Kritická hodnota na hladine α: $\chi^2_{0{,}95}(2) \approx 5{,}991$, kritický obor $\chi^2 \geq 5{,}991$
- p-hodnota: $p = P(\chi^2 \geq 2) \approx 0{,}3679$
- Keďže $p \approx 0{,}3679 > \alpha = 0{,}05$, nulovú hypotézu nezamietame – dáta neposkytujú dosť dôkazov proti nej.
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Test χ² pracuje s početnosťami (počtami prípadov), nie s percentami alebo priemermi.

Graf

Rozdelenie χ² – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



Kvadratická regresia.

Zadanie: `data(x y; 1 2.1; 2 3.9; 3 9.2; 4 15.8; 5 25.1; 6 36.2)`

Korelácia a regresia · Kvadratická

Výsledok

$$\hat{y} = 1{,}73 - 0{,}9354x + 1{,}116x^2$$

$$R^2 \approx 0{,}9997$$

Postup

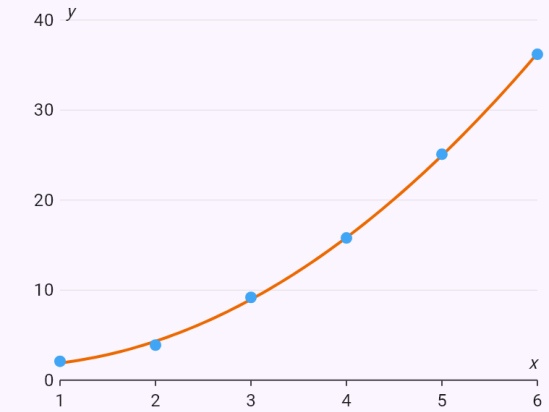
- Počet hodnôt $n = 6$, zoradené dáta
  $[1;\,2{,}1],\ [2;\,3{,}9],\ [3;\,9{,}2],\ [4;\,15{,}8],\ [5;\,25{,}1],\ [6;\,36{,}2]$
- Súčty a priemery
  $$\sum x_i = 21$$
  $$\sum x_i^2 = 91$$
  $$\sum x_i^3 = 441$$
  $$\sum x_i^4 = 2275$$
  $$\sum y_i = 92{,}3$$
  $$\sum x_i y_i = 443{,}4$$
  $$\sum x_i^2 y_i = 2284$$
- Normálne rovnice metódy najmenších štvorcov pre ŷ = a + bx + cx²
  $$6a + 21b + 91c = 92{,}3$$
  $$21a + 91b + 441c = 443{,}4$$
  $$91a + 441b + 2275c = 2284$$
- Regresná krivka: $\hat{y} = 1{,}73 - 0{,}9354x + 1{,}116x^2$
- Koeficient determinácie – krivka vysvetľuje $99{,}97\,\%$ rozptylu y.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{0{,}3012}{874{,}5} \approx 0{,}9997$$

- Regresnú krivku je rozumné používať len v rozsahu nameraných x – mimo neho (extrapolácia) môže byť odhad veľmi nepresný.

Graf

Bodový graf s regresnou krivkou



Exponenciálna regresia linearizáciou.

Zadanie: `data(x y; 0 2.0; 1 2.7; 2 3.7; 3 5.0; 4 6.8; 5 9.1)` Korelácia a regresia · Exponenciálna

Výsledok

$$\hat{y} = 2{,}002\, e^{0{,}3042x}$$

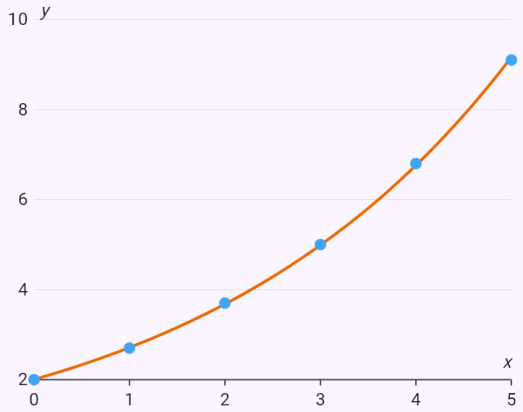
$$R^2 \approx 0{,}9998$$

Postup

- Počet hodnôt $n = 6$, zoradené dáta
  $[0;\ 2],\ [1;\ 2{,}7],\ [2;\ 3{,}7],\ [3;\ 5],\ [4;\ 6{,}8],\ [5;\ 9{,}1]$
- Model prevedieme logaritmovaním na priamku v premenných (u, v): $y = a\, e^{bx} \Rightarrow \ln y = \ln a + bx$; transformované dvojice
  $[0;\ 0{,}6931],\ [1;\ 0{,}9933],\ [2;\ 1{,}308],\ [3;\ 1{,}609],\ [4;\ 1{,}91$
- Koeficienty regresnej priamky metódou najmenších štvorcov
  $$b = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{\sum (u_i - \bar{u})^2} = \frac{5{,}324}{17{,}5} \approx 0{,}3042$$
  $$\ln a = \overline{\ln y} - b\bar{x} = 1{,}455 - 0{,}3042 \cdot 2{,}5 \approx 0{,}6943$$
- Koeficient a dostaneme spätnou transformáciou: $a = e^{0{,}6943} \approx 2{,}002$
- Regresná krivka: $\hat{y} = 2{,}002\, e^{0{,}3042x}$
- Koeficient determinácie – krivka vysvetľuje $99{,}98\,\%$ rozptylu y.
  $$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{0{,}006563}{35{,}95} \approx 0{,}9998$$
- Koeficienty sú vypočítané metódou najmenších štvorcov pre transformované dáta (logaritmy), nie pre pôvodné hodnoty – výsledok sa preto môže trochu líšiť od nelineárnej regresie v štatistických programoch.
- Regresnú krivku je rozumné používať len v rozsahu nameraných x – mimo neho (extrapolácia) môže byť odhad veľmi nepresný.

Graf

Bodový graf s regresnou krivkou



Pravostranný t-test zo súhrnných hodnôt: n = 25, x̄ = 104, s = 15, μ₀ = 100.

Zadanie: `model(ttest; find=test[one]; n=25[one]; mean=104[one]; sd=15[one]; mu0=100[one];`

`alpha=0.05[one]; alternative=greater; sigma=sample)`

Výsledok

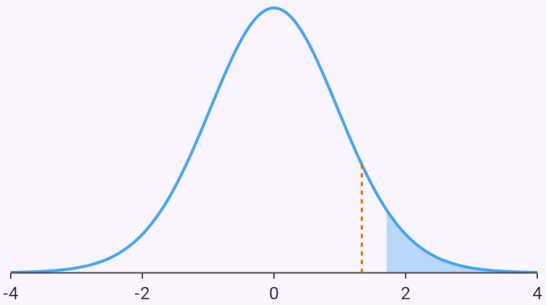
$$t \approx 1{,}333$$
$$p \approx 0{,}09747 > \alpha = 0{,}05$$

Postup

- Hľadáme t; zadané:
  $n = 25, \quad \bar{x} = 104, \quad s = 15, \quad \mu_0 = 100, \quad \alpha = 0{,}05$
- t-test overuje, či sa priemer populácie líši od zadanej hodnoty μ₀. Testové kritérium t = (x̄ − μ₀)/(s/√n) má pri platnosti nulovej hypotézy t-rozdelenie s n − 1 stupňami voľnosti; p-hodnota hovorí, aký pravdepodobný je taký alebo extrémnejší výsledok, keby H_0 platila. Pri známom σ ide o z-test. Rovnica modelu: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- Nulová a alternatívna hypotéza:
  $H_0: \mu = 100, \quad H_1: \mu > 100$
- Smerodajná chyba: $SE = \frac{s}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$
- Testové kritérium – pri platnosti H_0 má t-rozdelenie s 24 stupňami voľnosti: $t = \frac{\bar{x} - \mu_0}{SE} = \frac{104 - 100}{3} \approx 1{,}333$
- Kritická hodnota na hladine α: $t_{0{,}95}(24) \approx 1{,}711$, kritický obor $t \geq 1{,}711$
- p-hodnota: $p = P(T \geq 1{,}333) \approx 0{,}09747$
- Keďže $p \approx 0{,}09747 > \alpha = 0{,}05$, nulovú hypotézu nezamietame – dáta neposkytujú dosť dôkazov proti nej.
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Pre malý výber (n < 30) t-rozdelenie predpokladá, že dáta pochádzajú z približne normálneho rozdelenia. Pri výrazne zošikmených dátach alebo odľahlých hodnotách je vhodnejší neparametrický test.

Graf

t-rozdelenie – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



$\alpha = 0{,}05 \quad t \approx 1{,}333$

Welchov t-test dvoch skupín zo súhrnných hodnôt.

Zadanie: `model(twosample; find=test[one]; n1=30[one]; mean1=52[one]; sd1=8[one];`

```
n2=28[one]; mean2=47[one]; sd2=10[one]; alpha=0.05[one]; variances=welch; alternative=two)
```

Výsledok

$$t \approx 2{,}093$$

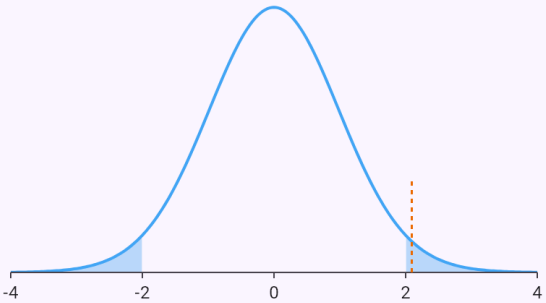
$$p \approx 0{,}04124 \leq \alpha = 0{,}05$$

Postup

- Hľadáme t; zadané
  $n_1 = 30, \quad \bar{x}_1 = 52, \quad s_1 = 8, \quad n_2 = 28, \quad \bar{x}_2 = 47,$
- Porovnanie priemerov dvoch nezávislých skupín. Welchov test nepredpokladá zhodné rozptyly a je dnes odporúčanou voľbou; klasický t-test so spoločným (váženým) rozptylom predpokladá, že obe skupiny majú rozptyl rovnaký. Stačia počty, priemery a smerodajné odchýlky oboch skupín. Rovnica modelu:
  $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
- Nulová a alternatívna hypotéza:
  $H_0 \colon \mu_1 = \mu_2, \quad H_1 \colon \mu_1 \neq \mu_2$
- Smerodajná chyba
  $$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{8^2}{30} + \frac{10^2}{28}} \approx 2{,}388$$
- Stupne voľnosti podľa Welcha a Satterthwaitea
  $$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \approx 51{,}71$$
- Testové kritérium – pri platnosti H_0 má t-rozdelenie s $51{,}71$ stupňami voľnosti:
  $$t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = \frac{52 - 47}{2{,}388} \approx 2{,}093$$
- Kritická hodnota na hladine α: $t_{0{,}975}(51{,}71) \approx 2{,}007$, kritický obor $|t| \geq 2{,}007$
- p-hodnota: $p = 2\,P(T \geq 2{,}093) \approx 0{,}04124$
- Keďže $p \approx 0{,}04124 \leq \alpha = 0{,}05$, nulovú hypotézu zamietame – výsledok je štatisticky významný.
- p-hodnota je pravdepodobnosť výsledku aspoň takého extrémneho ako pozorovaný, ak H_0 platí – nie je to pravdepodobnosť, že H_0 platí. Nezamietnutie H_0 neznamená, že H_0 je pravdivá; štatistická významnosť nehovorí nič o veľkosti ani praktickom význame rozdielu.
- Welchov test nepredpokladá zhodné rozptyly oboch skupín; jeho stupne voľnosti v nemusia byť celé číslo.

Graf

t-rozdelenie – vyfarbený je kritický obor, čiarkovane hodnota testového kritéria



$\alpha = 0{,}05 \quad t \approx 2{,}093$

Normálne rozdelenie: P(X < 85) pre N(100; 15²).

Zadanie: `model(normal; find=below[one]; mu=100[one]; sigma=15[one]; x=85[one])`

Výsledok

$$P(X < 85) \approx 0{,}1587 = 15{,}87\,\%$$

Postup

- Hľadáme P; zadané: $\mu = 100, \quad \sigma = 15, \quad x = 85$
- Mnoho veličín (výška, výkon v teste, chyba merania) kolíše okolo priemeru μ súmerne a čím ďalej od neho, tým zriedkavejšie – tak to opisuje normálne (Gaussovo) rozdelenie N(μ; σ²) so smerodajnou odchýlkou σ. Pravdepodobnosti sa počítajú cez štandardizáciu z = (x − μ)/σ a distribučnú funkciu Φ normovaného rozdelenia N(0; 1): P(X < x) = Φ(z). Kvantil je hodnota, pod ktorou leží zadaný podiel p. Rovnica modelu

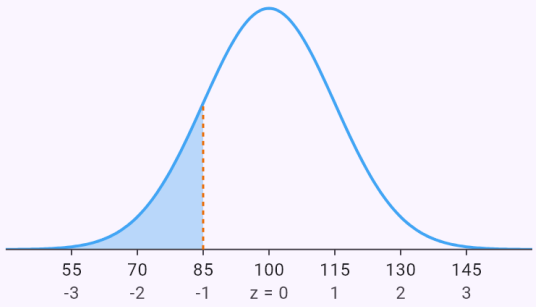
  $$X \sim N\left(\mu;\, \sigma^2\right)$$
  $$Z = \frac{X - \mu}{\sigma} \sim N(0;\, 1)$$

- Vyjadríme hľadanú veličinu: $z = \dfrac{x - \mu}{\sigma}$
- Dosadíme: $z = \dfrac{85 - 100}{15} = -1$
- Dosadíme

  $$P(X < 85) = \Phi(-1) = (1 - \Phi(1)) \approx (1 - 0{,}8413) \approx 0$$

Graf

Hustota normálneho rozdelenia, vyfarbená plocha je hľadaná pravdepodobnosť



$P \approx 0{,}1587 \quad x = 85$

Podiel ľudí s výškou medzi 160 a 185 cm pri μ = 170, σ = 8.

Zadanie: `model(normal; find=between[one]; mu=170[one]; sigma=8[one]; a=160[one];`

```
b=185[one])
```

Výsledok

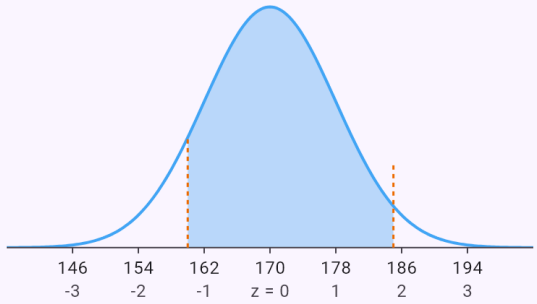
$$P(160 < X < 185) \approx 0{,}864 = 86{,}4\,\%$$

Postup

- Hľadáme P; zadané:
  $\mu = 170, \quad \sigma = 8, \quad a = 160, \quad b = 185$
- Mnoho veličín (výška, výkon v teste, chyba merania) kolíše okolo priemeru μ súmerne a čím ďalej od neho, tým zriedkavejšie – tak to opisuje normálne (Gaussovo) rozdelenie N(μ; σ²) so smerodajnou odchýlkou σ. Pravdepodobnosti sa počítajú cez štandardizáciu z = (x − μ)/σ a distribučnú funkciu Φ normovaného rozdelenia N(0; 1): P(X < x) = Φ(z). Kvantil je hodnota, pod ktorou leží zadaný podiel p. Rovnica modelu
  $X \sim N(\mu;\, \sigma^2)$
  $Z = \dfrac{X - \mu}{\sigma} \sim N(0;\, 1)$
- Vyjadríme hľadanú veličinu: $z_a = \dfrac{a - \mu}{\sigma}, \quad z_b = \dfrac{b - \mu}{\sigma}$
- Dosadíme:
  $z_a = \dfrac{160 - 170}{8} = -1{,}25, \quad z_b = \dfrac{185 - 170}{8} = 1{,}875$
- Dosadíme
  $P(160 < X < 185) = \Phi(1{,}875) - \Phi(-1{,}25) \approx 0{,}9696$

Graf

Hustota normálneho rozdelenia, vyfarbená plocha je hľadaná pravdepodobnosť



$P \approx 0{,}864 \quad a \quad b$

Kvantil: skóre, pod ktorým leží 95 % populácie.

Zadanie: `model(normal; find=quantile[one]; mu=100[one]; sigma=15[one]; p=95[pct])`

Výsledok

$$x \approx 124{,}7$$

Postup

- Hľadáme x; zadané: $\mu = 100, \quad \sigma = 15, \quad p = 95\,\%$
- Prevedieme jednotky: $p = 95\,\% = 0{,}95$
- Mnoho veličín (výška, výkon v teste, chyba merania) kolíše okolo priemeru μ súmerne a čím ďalej od neho, tým zriedkavejšie – tak to opisuje normálne (Gaussovo) rozdelenie N(μ; σ²) so smerodajnou odchýlkou σ. Pravdepodobnosti sa počítajú cez štandardizáciu z = (x − μ)/σ a distribučnú funkciu Φ normovaného rozdelenia N(0; 1): P(X < x) = Φ(z). Kvantil je hodnota, pod ktorou leží zadaný podiel p. Rovnica modelu

$$X \sim N\left(\mu;\, \sigma^2\right)$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0;\, 1)$$

- Vyjadríme hľadanú veličinu

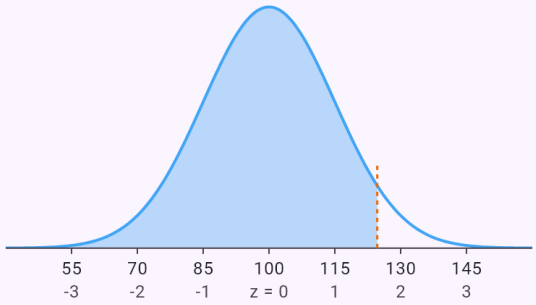
$$P(X < x) = \Phi\left(\frac{x - \mu}{\sigma}\right) = p \Rightarrow x = \mu + z_p\sigma, \quad z_p = \text{`}$$

- Dosadíme

$$z_{0{,}95} = \Phi^{-1}(0{,}95) \approx 1{,}645, \quad x = 100 + 1{,}645 \cdot 15 \approx 1\text{`}$$

Graf

Hustota normálneho rozdelenia, vyfarbená plocha je hľadaná pravdepodobnosť



$p = 0{,}95 \quad x \approx 124{,}7$

Známe obmedzenia

- Kvartily sa počítajú ako mediány polovíc (školský postup); štatistické programy (Excel, SPSS) používajú iné vzorce a pri malých súboroch im môžu vyjsť trochu inak.
- Porovnanie dvojíc skupín po ANOVE alebo Kruskalovom–Wallisovom teste (post hoc) a viacfaktorová ANOVA zatiaľ nie sú – porovnajte dvojice t-testom alebo Mannovým–Whitneyho testom s Bonferroniho korekciou; nelineárna regresia sa počíta linearizáciou.
- Dáta sa zadávajú len ručne, mriežkou alebo súborom CSV – z fotky tabuľky ich aplikácia zatiaľ nerozpozná.

Vygenerované z aplikácie Mathtrace