Základy biostatistiky

statistika:

- **popisná** (data stručně popsat, něco z dat „vydolovat“)
- **induktivní** (tvrdit něco nového, zobecnit na větší soubor, záleží na interpretaci)

příklady dat:

- **výšky** (výška desetiletých chlapců/dívek)
- **děti** (pohlaví, porodní hmotnost a délka, hmotnost a délka v jednom roce, věk otce a matky, počet onemocnění otitidou v prvním roce věku)
- **kojení** (hmotnost a délka porodní a ve 24. týdnu, věk a výška obou rodičů, zda těhotenství plánováno, zda dudlík, porodnice)

znak - vlastnost měřená na objektu (statistické jednotce): délka, barva, . . .
možná **měřítka**:

- **nominální** (porodnice, pohlaví) seznam všech rozlišitelných hodnot, **faktor**
- **ordinální** (vzdělání matky,. . . , stupeň bolesti) hodnoty nominálního uspořádány, **uspořádaný faktor**
- **intervalové** (rok narození, teplota v $^{\mathrm{o}}$C) stejné vzdálenosti sousedních hodnot, o kolik se liší?
- **poměrové** (hmotnost, výška, věk) srovnání se zvolenou jednotkou, kolikrát je větší?

číselné **veličiny** (zápis hodnot znaků):

- **spojité**: intervalové, poměrové (ordinální) měřítko
- **diskrétní**: četnosti hodnot v nominálním nebo ordinálním měřítku

Popisné statistiky

statistika: též funkce pozorovaných hodnot

$x_1, x_2, \ldots, x_n$ zjištěné hodnoty

$x_1^*,\ x_2^*,\ \ldots,\ x_m^*$ možné hodnoty (různé)
$n_1,\ n_2,\ \ldots,\ n_m$ **četnosti** hodnot

$$n_1 + n_2 + \ldots + n_m = \sum_{j=1}^{m} n_j = n$$

$\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_m}{n}$ - relativní četnosti

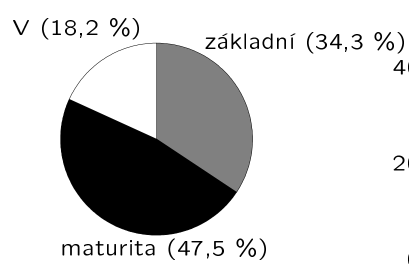
$N_j = \sum_{i=1}^{j} n_i$ **kumulativní** četnosti

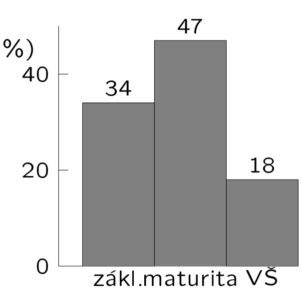
kumulativní četnosti - nutno aspoň ordinální měřítko

histogram: grafické znázornění četností
plocha (výška) obdélníku úměrná četnosti
(relativní četnosti – jiné měřítko)
podobně **výsečový diagram**

příklad **kojení** (vzdělání matky):

vzděl.	zákl.	maturita	VŠ	celkem
x_j^*	1	2	3	
n_j	34	47	18	99
n_j/n	0,343	0,475	0,182	1,000
n_j/n	34,3 %	47,5 %	18,2 %	100 %
N_j	34	81	99	



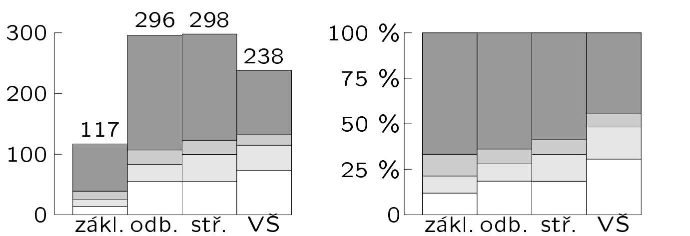


dvojice znaků –
možnost porovnání či zkoumání závislosti (kontingenční tabulka)
v procentech v dané skupině (pro danou hodnotu jednoho znaku)

příklad **kouření u mužů**

vzdělání	zákl.	odb.	mat.	VŠ	celkem
nekuřák	14	55	55	73	197
bývalý kuřák	11	28	44	42	125
kuřák	14	24	24	17	79
silný kuřák	78	189	175	106	548
celkem	117	296	298	238	949

vzdělání	zákl.	odb.	mat.	VŠ	celk.
nekuřák	12,0%	18,6%	18,5%	30,7%	20,6%
bývalý kuřák	9,4%	9,5%	14,8%	17,6%	13,2%
kuřák	12,0%	8,1%	8,1%	7,1%	8,3%
silný kuřák	66,7%	63,9%	58,7%	44,5%	57,8%
celkem	100%	100%	100%	100%	100%

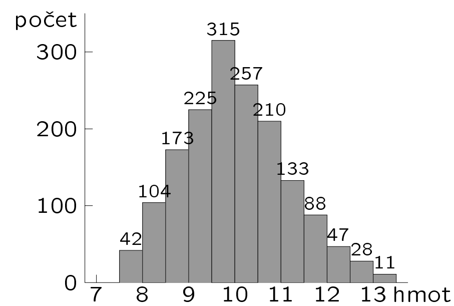


(zdola: nekuřák, bývalý kuřák, kuřák, silný kuřák)

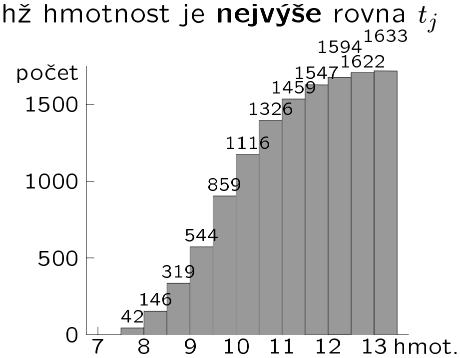
histogram u **spojité** veličiny – **třídění**: všechny hodnoty z daného intervalu $\langle t_{j-1}, t_j\rangle$ nahradíme prostřední hodnotou $x_j^* = (t_{j-1} + t_j)/2$

hmotnost dětí (příklad **děti**)

j	x_j^*	t_j	n_j	n_j/n	N_j	N_j/n
1	7750	8000	42	0,026	42	0,026
2	8250	8500	104	0,063	146	0,089
3	8750	9000	173	0,106	319	0,195
4	9250	9500	225	0,138	544	0,333
5	9750	10000	315	0,193	859	0,526
6	10250	10500	257	0,157	1116	0,683
7	10750	11000	210	0,129	1326	0,812
8	11250	11500	133	0,081	1459	0,893
9	11750	12000	88	0,054	1547	0,947
10	12250	12500	47	0,029	1594	0,976
11	12750	13000	28	0,017	1622	0,992
12	13250	∞	11	0,007	1633	1,000

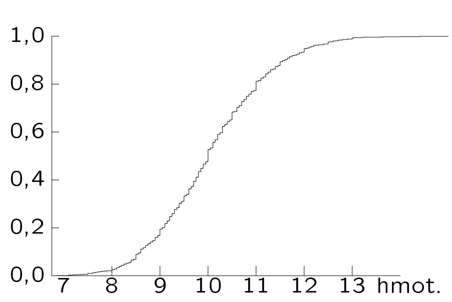


kumulativní četnosti ukazují vždy podíl dětí, jejichž hmotnost je **nejvýše** rovna t_j



empirická distribuční funkce: relativní četnost hodnot, které jsou nejvýše x

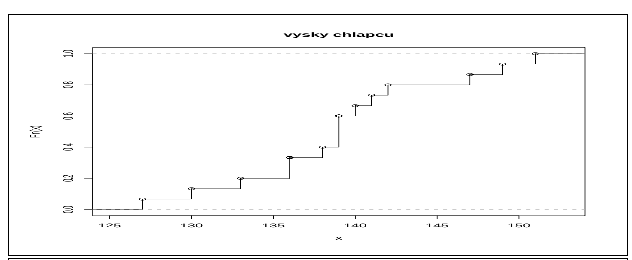
$$F_n(x) = \frac{\text{počet } (x_i \le x)}{n}$$

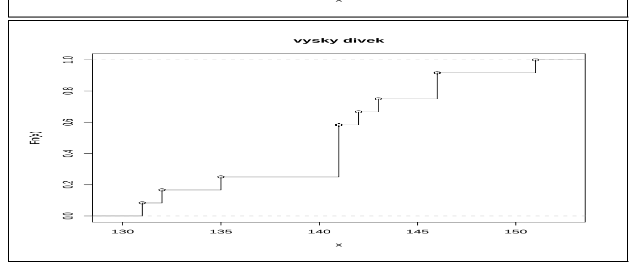


x – výšky desetiletých hochů
y – výšky desetiletých dívek

i	1	2	3	4	5	6	7	8
x_i	130	140	136	141	139	133	149	151
y_i	135	141	143	132	146	146	151	141

i	9	10	11	12	13	14	15	
x_i	139	136	138	142	127	139	147	
y_i	141	131	142	141				





uspořádaný seznam hodnot

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$$

pořadí na které místo se dané pozorování v uspořádaném seznamu dostane (při shodě průměrné pořadí)

míry polohy: $\mu(a + bX) = a + b\mu(X)$

- **průměr**

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **medián** (dolní a horní polovina hodnot)

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ liché} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & n \text{ sudé} \end{cases}$$

- **minimum, maximum**

$$x_{\text{min}} = x_{(1)}$$
$$x_{\text{max}} = x_{(n)}$$

- **variační průměr**

$$\frac{1}{2}\left(x_{(1)} + x_{(n)}\right) = \frac{1}{2}\left(x_{\text{min}} + x_{\text{max}}\right)$$

- p**-tý percentil** (dolních 100p % hodnot)

$r = [(n+1)p]$ celá část $(n+1)p$

$q = (n+1)p - r$ zlomková část $(n+1)p$

$$\boxed{x_p = (1-q)x_{(r)} + qx_{(r+1)}}$$

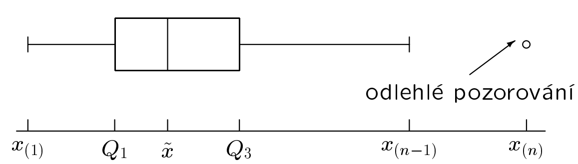
- **dolní kvartil** (oddělí dolní čtvrtinu)

$$Q_1 = x_{1/4}$$

- **horní kvartil** (oddělí dolní tři čtvrtiny)

$$Q_3 = x_{3/4}$$

krabicový diagram



výšky dívek

j	1	2	3	4	5	6	7	8
y_j^*	131	132	135	141	142	143	146	151
n_j	1	1	1	4	1	1	2	1
poř.	1	2	3	5,5	8	9	10,5	12

$$\bar{y} = \frac{1}{12}(131 + 132 + \ldots + 151) = 140{,}83$$
$$\tilde{y} = \frac{1}{2}\left(y_{(6)} + y_{(7)}\right) = \frac{1}{2}(141 + 141) = 141$$
$$r = [(12+1)/4] = 3 \qquad q = (12+1)/4 - 3 = 1/4$$
$$Q_1 = \frac{3}{4}y_{(3)} + \frac{1}{4}y_{(4)} = 0{,}75 \cdot 135 + 0{,}25 \cdot 141 = 136{,}5$$
$$Q_3 = 0{,}25 \cdot 143 + 0{,}75 \cdot 146 = 145{,}25$$
$$s_y^2 = \frac{1}{11}\left((131 - 140{,}3)^2 + \ldots + (151 - 140{,}83)^2\right) \doteq 33{,}788$$
$$s_y = \sqrt{33{,}788} \doteq 5{,}813$$
$$R = 151 - 131 = 20$$
$$R_Q = 145{,}25 - 136{,}5 = 8{,}75$$

vztah mužů ke kouření (základní vzdělání):

$$H = \frac{14}{117}\log\frac{14}{117} + \ldots + \;+\frac{78}{117}\log\frac{78}{117} = 1{,}000689$$

ostatní: $1{,}025939; 1{,}109783; 1{,}217334$

míry variability (měřítka)

$$\sigma(a + bX) = b\sigma(X) \quad (b > 0)$$

- **směrodatná odchylka**

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- **rozptyl** s_x^2 (nesplňuje vztah)
- **rozpětí** $R = x_{\text{max}} - x_{\text{min}}$
- **kvartilové rozpětí** $R_Q = Q_3 - Q_1$
- **variační koeficient**
porovnání variability při různých úrovních

$$V_x = \frac{s_x}{\bar{x}}$$

- **entropie** (nejistota nominální)

$$H = -\sum_{j=1}^{m}\frac{n_j}{n}\log\frac{n_j}{n}$$

(nezávisí na označení hodnot)

z-skór (normovaná veličina)

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

platí $\bar{z} = 0,\ s_z = 1 \Rightarrow$
vyšetřováním z hodnotíme jiné vlastnosti, na poloze a variabilitě nezávislé

- **šikmost**

$$g_1 = \frac{1}{n}\sum_{i=1}^{n} z_i^3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^3$$

- **špičatost**

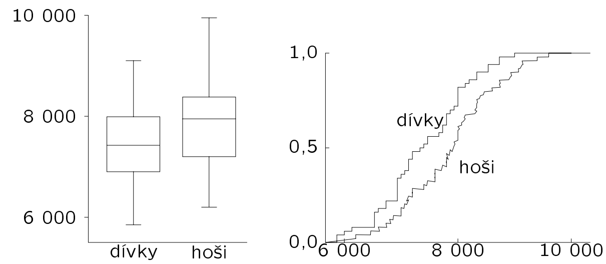
$$g_2 = \frac{1}{n}\sum_{i=1}^{n} z_i^4 - 3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^4 - 3$$

(někdy bez odečítání 3)

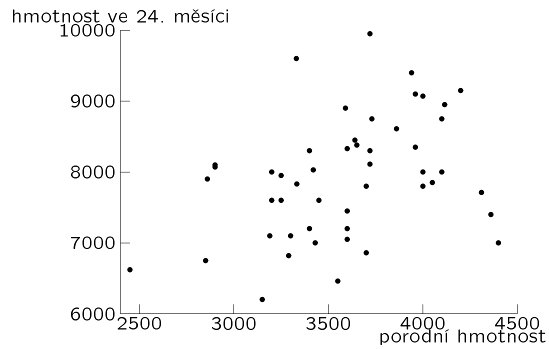
g_1, g_2 se používají k posouzení normality

další grafická znázornění

- srovnání souborů dat

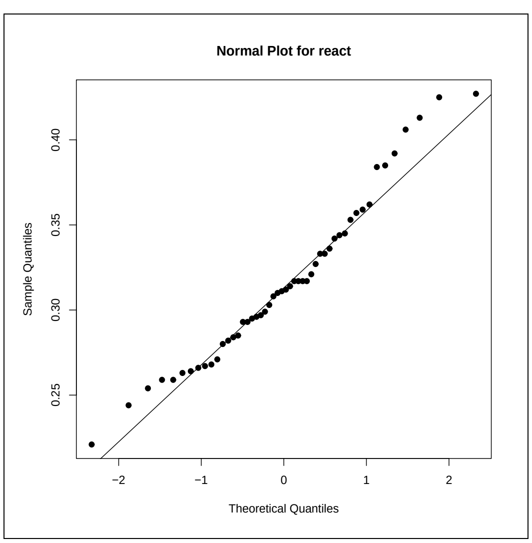


- závislost spojitých veličin (bodový diagram)

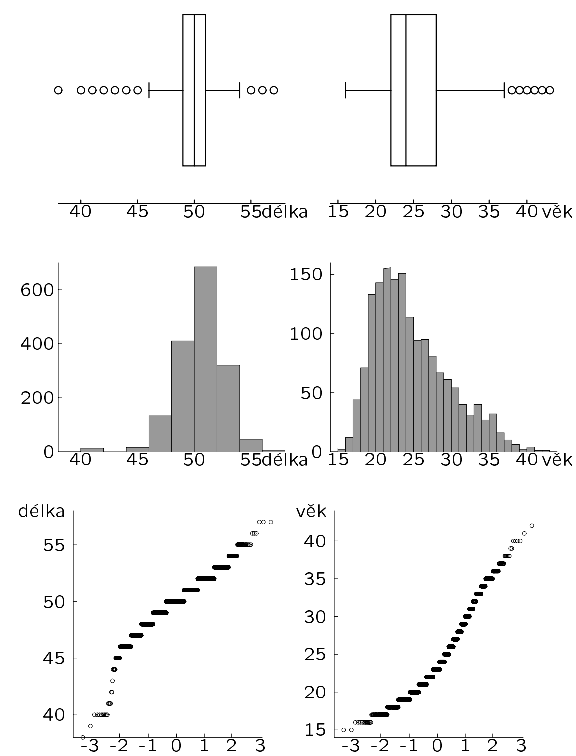


další grafická znázornění

- normální diagram
  - k ověřování předpokladu **normálního** rozdělení (častý předpoklad)
  - srovnání bodů s přímkou



$g_1 = 0{,}521, \qquad g_2 = -0{,}321$



$g_1 = -0{,}893, g_2 = 3{,}511 \qquad g_1 = 0{,}760, g_2 = 0{,}013$

Náhodné jevy

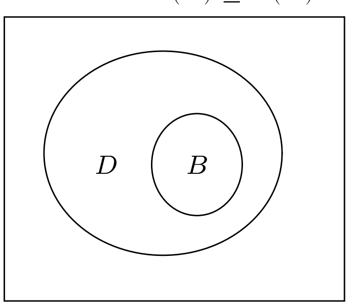
- **náhodný pokus** výsledek nejistý, při opakování stabilita frekvence možných výsledků
- **náhodný jev** tvrzení o výsledku náhodného pokusu, podmnožiny množiny Ω
- **jistý jev** Ω nastává vždy
- **nemožný jev** $\emptyset$ nenastává nikdy
- **podjev**: $B \subset D$ znamená $B \Rightarrow D$
- **jev opačný**: $\overline{D} \Leftrightarrow$ neplatí D
- **průnik jevů** $B \cap D$ nastaly oba jevy
- **sjednocení jevů** $D \cup B$ nastal aspoň jeden
- **neslučitelné jevy** $B \cap D = \emptyset$

Pravděpodobnost $\mathsf{P}(B)$

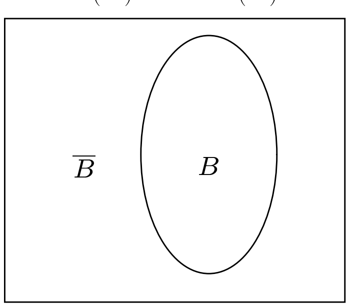
- objektivní číselné vyjádření „naděje“, že nastane B
- modelový protějšek relativní četnosti
- vlastnosti psti
  - $\boxed{0 \le \mathsf{P}(B) \le 1}$
  - $\boxed{\mathsf{P}(\Omega) = 1, \mathsf{P}(\emptyset) = 0}$
  - $\boxed{B \cap D = \emptyset \Rightarrow \mathsf{P}(B \cup D) = \mathsf{P}(B) + \mathsf{P}(D)}$
  - $\boxed{\mathsf{P}(B \cup D) = \mathsf{P}(B) + \mathsf{P}(D) - \mathsf{P}(B \cap D)}$
  - $\boxed{B \subset D \Rightarrow \mathsf{P}(B) \le \mathsf{P}(D)}$
  - $\boxed{\mathsf{P}(\overline{B}) = 1 - \mathsf{P}(B)}$
- **klasická definice psti**: m stejně pravděpodobných elementárních jevů, m_B příznivých B

$$\mathsf{P}(B) = \frac{m_B}{m}$$

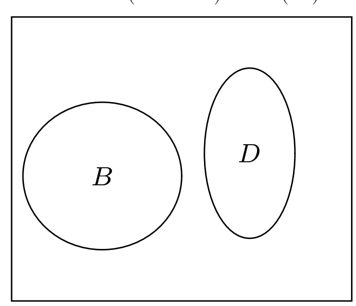
$B \subset D \Rightarrow \mathsf{P}(B) \le \mathsf{P}(D)$



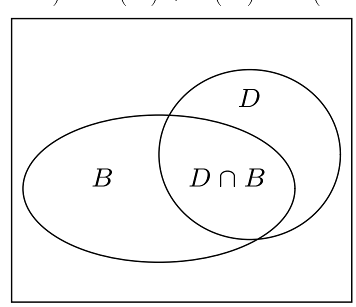
$\mathsf{P}(\overline{B}) = 1 - \mathsf{P}(B)$



$B \cap D = \emptyset \Rightarrow \mathsf{P}(B \cup D) = \mathsf{P}(B) + \mathsf{P}(D)$



$\mathsf{P}(B \cup D) = \mathsf{P}(B) + \mathsf{P}(D) - \mathsf{P}(B \cap D)$



příklad **rodina**: tři sourozenci, celkem 8 elementárních jevů $\omega_1, \ldots, \omega_8$

ω_i	D	B	$B \cap D$	$B \cup D$	C
(m,m,m)					$+$
(f,m,m)	$+$	$+$	$+$	$+$	$+$
(m,f,m)		$+$		$+$	$+$
(f,f,m)	$+$			$+$	$+$
(f,f,f)	$+$			$+$	
(m,f,f)					
(f,m,f)	$+$			$+$	
(m,m,f)		$+$		$+$	

D nejmladší je dívka, $\mathsf{P}(D) = 4/8 = 1/2$
B v rodině je jediná dívka, $\mathsf{P}(B) = 3/8$
$B \cap D$ jediná dívka je nejmladší, $\mathsf{P}(B \cap D) = 1/8$
$\mathsf{P}(B \cup D) = \mathsf{P}(B) + \mathsf{P}(D) - \mathsf{P}(B \cap D) = \frac{3}{8} + \frac{4}{8} - \frac{1}{8} = \frac{6}{8}$
C nejstarší je hoch, $\mathsf{P}(C) = 4/8 = 1/2$
Když víme, že nejstarší je hoch (C), jaká je pak pst, že nejmladší je dívka (D)?

$$\boxed{2/4 = 1/2}$$

stejně, jako když jsme nic nevěděli
pst jevu D **nezávisí** na tom, zda platí C

nezávislost: pst jevu D nezávisí na tom, zda B nastal či nenastal: D, B **nezávislé jevy**
podmíněná pst (pst D za podmínky B)

$$\boxed{\mathsf{P}(D|B) = \frac{\mathsf{P}(D \cap B)}{\mathsf{P}(B)} = \frac{m_{D \cap B}}{m_B} = \frac{m_{D \cap B}/m}{m_B/m}}$$

nezávislost D, B

$$\mathsf{P}(D \cap B) = \mathsf{P}(D)\mathsf{P}(B)$$

příklad **rodina**:

$$\mathsf{P}(B \cap D) = \frac{1}{8} \neq \frac{3}{8} \cdot \frac{4}{8} = \mathsf{P}(B)\mathsf{P}(D) \Rightarrow B, D \text{ závislé}$$

$$\mathsf{P}(B|D) = \frac{\mathsf{P}(B \cap D)}{\mathsf{P}(D)} = \frac{1/8}{4/8} = \frac{1}{4}$$

$$\mathsf{P}(B|\overline{D}) = \frac{\mathsf{P}(B \cap \overline{D})}{\mathsf{P}(\overline{D})} = \frac{2/8}{4/8} = \frac{1}{2}$$

$$\mathsf{P}(B) = \frac{3}{8}$$

HWE (zákon Hardyův-Weinbergův)

- diploidní populace
- na daném lokusu dvě alely: A, a
- pst alely A v populaci p
- pst alely a v populaci $q = 1 - p$
- nezávislé sdružování alel znamená

$$\mathsf{P}(AA) = \mathsf{P}(A)\mathsf{P}(A) = p^2$$
$$\mathsf{P}(aa) = \mathsf{P}(a)\mathsf{P}(a) = q^2$$
$$\mathsf{P}(Aa) = \mathsf{P}(A)\mathsf{P}(a) + \mathsf{P}(a)\mathsf{P}(A) = 2pq$$

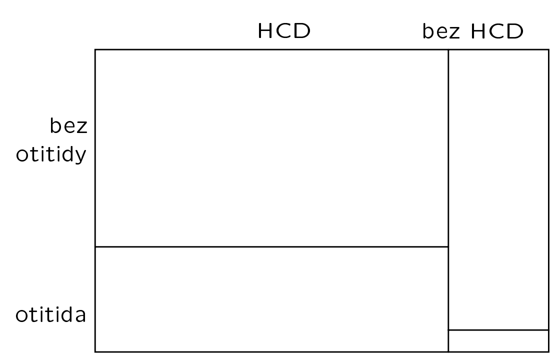
děti (otitidy a záněty HCD)

	HCD	bez HCD	celkem
bez otitidy	5168	2088	7256
otitida	2747	163	2910
celkem	7915	2251	10166

	HCD	bez HCD	celkem
bez otitidy	0,508	0,205	0,714
otitida	0,270	0,016	0,286
celkem	0,779	0,221	1,000

podmíněno HCD

	HCD	bez HCD	celkem
bez otitidy	0,653	0,928	0,714
otitida	0,347	0,072	0,286
celkem	1,000	1,000	1,000



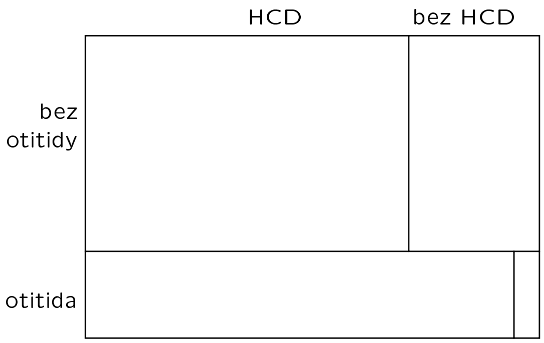
děti (otitidy a záněty HCD)

	HCD	bez HCD	celkem
bez otitidy	5168	2088	7256
otitida	2747	163	2910
celkem	7915	2251	10166

	HCD	bez HCD	celkem
bez otitidy	0,508	0,205	0,714
otitida	0,270	0,016	0,286
celkem	0,779	0,221	1,000

podmíněno otitidou

	HCD	bez HCD	celkem
bez otitidy	0,712	0,288	1,000
otitida	0,944	0,056	1,000
celkem	0,779	0,221	1,000



předpoklad:

– $H_1, \ldots, H_k$ neslučitelné

– sjednocení $H_1, \ldots, H_k$ – jev jistý

vzorec **pro úplnou pst**

$$\boxed{\mathsf{P}(C) = \sum_{j=1}^{k} \mathsf{P}(C|H_j)\mathsf{P}(H_j)}$$

Bayesův vzorec

$$\boxed{\mathsf{P}(H_i|C) = \frac{\mathsf{P}(C|H_i)\mathsf{P}(H_i)}{\mathsf{P}(C)}}$$

$$\boxed{\mathsf{P}(H_i|C) = \frac{\mathsf{P}(C|H_i)\mathsf{P}(H_i)}{\sum_{j=1}^{k} \mathsf{P}(C|H_j)\mathsf{P}(H_j)}}$$

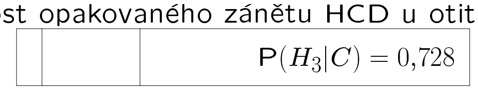
$H_1, \ldots, H_k$ – hypotézy
$\mathsf{P}(H_1), \ldots, \mathsf{P}(H_k)$ – apriorní psti
$\mathsf{P}(H_1|C), \ldots, \mathsf{P}(H_k|C)$ – aposteriorní psti

příklad **děti** C – otitida
H_j – výskyt zánětu HCD

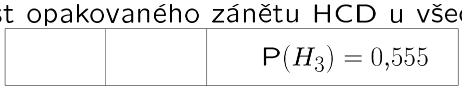
H_j	$\mathsf{P}(H_j)$	$\mathsf{P}(C\|H_j)$	součin
bez HCD	0,221	0,072	0,016
jednou HCD	0,223	0,276	0,061
opakovaně HCD	0,555	0,376	0,208
součet	1,000		0,286

$$\mathsf{P}(C) = 0{,}286$$
$$\mathsf{P}(H_3|C) = \frac{0{,}376 \cdot 0{,}555}{0{,}286} = 0{,}728$$

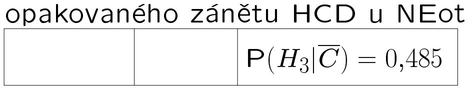
pst opakovaného zánětu HCD u otitid



pst opakovaného zánětu HCD u všech



pst opakovaného zánětu HCD u NEotitid



příklad: senzitivita, specificita testu

- $D, \overline{D}$ – nemocná/zdravá osoba
- $P, \overline{P}$ – pozitivní/negativní výsledek testu
- $\mathsf{P}(P|D)$ – **senzitivita** testu (0,98)
- $\mathsf{P}(\overline{P}|\overline{D})$ – **specificita** testu (0,99)
- $\mathsf{P}(D)$ – **incidence** nemoci (apriorní pst) (0,001)

$$\begin{aligned}
\mathsf{P}(D|P) &= \frac{\mathsf{P}(P|D)\mathsf{P}(D)}{\mathsf{P}(P|D)\mathsf{P}(D) + \mathsf{P}(P|\overline{D})\mathsf{P}(\overline{D})} \\
&= \frac{0{,}98 \cdot 0{,}001}{0{,}98 \cdot 0{,}001 + 0{,}01 \cdot 0{,}999} \\
&= \frac{0{,}00098}{0{,}01097} = 0{,}089 \\
\mathsf{P}(\overline{D}|\overline{P}) &= \frac{0{,}99 \cdot 0{,}999}{0{,}99 \cdot 0{,}999 + 0{,}02 \cdot 0{,}001} \\
&= 0{,}99998
\end{aligned}$$

náhodná veličina

- číselně vyjádřený výsledek náhodného pokusu
- každému elementárnímu jevu přiřadíme reálné číslo
- **diskrétní rozdělení**
  - možné hodnoty x^*
  - psti hodnot $\mathsf{P}(x_j^*)$ (pstní funkce)
- spojité rozdělení
  - interval možných hodnot
  - hustota $f(x)$

Příklad **rodina**

náhodná veličina – počet děvčat

ω_i	x_i	$x_i - \mu_X$	$(x_i - \mu_X)^2$	x_j^*
(m,m,m)	0	-1,5	2,25	0
(m,m,f)	1	-0,5	0,25	
(m,f,m)	1	-0,5	0,25	1
(f,m,m)	1	-0,5	0,25	
(f,f,m)	2	0,5	0,25	
(f,m,f)	2	0,5	0,25	2
(m,f,f)	2	0,5	0,25	
(f,f,f)	3	1,5	2,25	3
součet	12	0,0	6,00	

j	x_j^*	m_j	$\mathsf{P}(X = x_j^*)$
1	0	1	1/8
2	1	3	3/8
3	2	3	3/8
4	3	1	1/8
součet		8	8/8

distribuční funkce $\boxed{F_X(x) = \mathsf{P}(X \le x)}$

- diskrétní rozdělení $\boxed{F(x) = \sum_{t \le x} \mathsf{P}(X = t)}$
- spojité rozdělení $\boxed{F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t}$

  zřejmě pak:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

- vlastnosti distribuční funkce

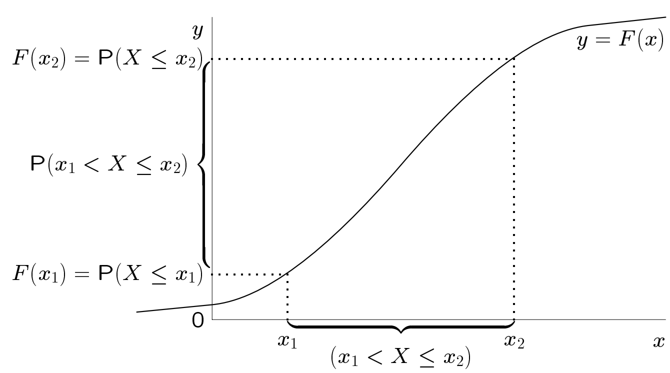
$$0 \le F(x) \le 1$$

neklesající: $\quad x_1 < x_2 \Rightarrow F(x_2) \ge F(x_1)$

$$\mathsf{P}(x_1 < X \le x_2) = F(x_2) - F(x_1)$$

$$\mathsf{P}(X \le x_2) = \mathsf{P}(X \le x_1) + \mathsf{P}(x_1 < X \le x_2)$$
$$F(x_2) = F(x_1) + \mathsf{P}(x_1 < X \le x_2)$$

geometrický význam **distribuční funkce**

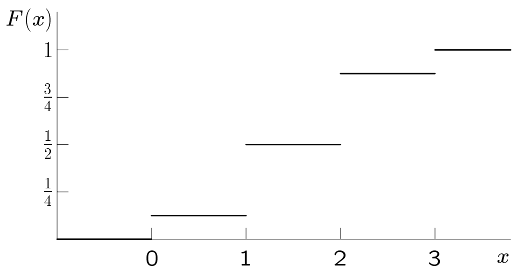


$$\begin{aligned} \mathsf{P}(X \le x_2) &= \mathsf{P}(X \le x_1) + \mathsf{P}(x_1 < X \le x_2) \\ F(x_2) &= F(x_1) + \mathsf{P}(x_1 < X \le x_2) \end{aligned}$$

příklad pro **diskrétní** rozdělení

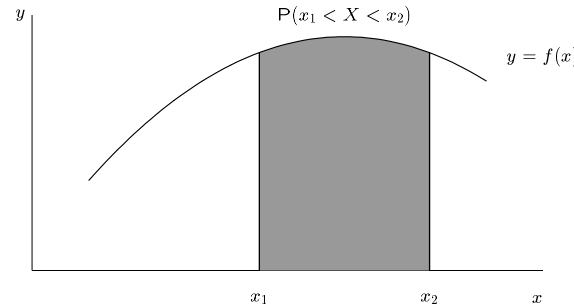
rozdělení počtu děvčat X

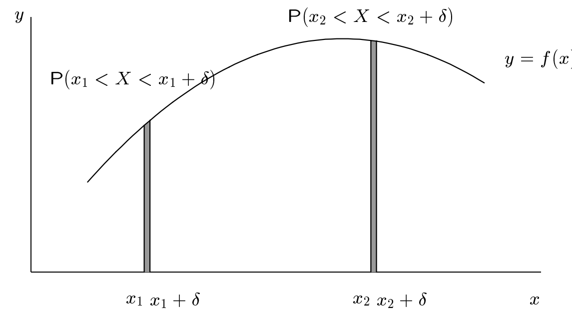
j	x_j^*	m_j	$\mathsf{P}(X = x_j^*)$	$F_X(x_j^*)$
1	0	1	1/8	1/8
2	1	3	3/8	4/8
3	2	3	3/8	7/8
4	3	1	1/8	8/8
součet		8	8/8	



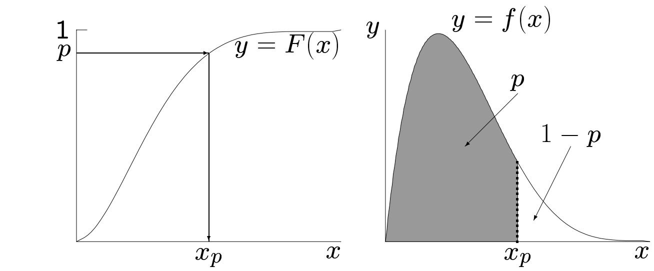
význam **hustoty spojitého** rozdělení:

$$f(x) \geq 0$$
$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

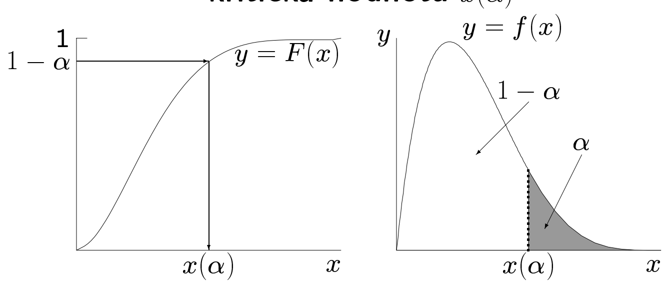




p-**kvantil** x_p



kritická hodnota $x(\alpha)$



$$\begin{aligned} x_{1-\alpha} + x(\alpha) &= 1 \\ x_p + x(1-p) &= 1 \end{aligned}$$

střední hodnota μ

- míra polohy, **populační průměr**
- vážený průměr možných hodnot
- diskrétní: $\mu_X = \sum_j x_j^* \mathsf{P}(X = x_j^*)$
- spojité $\mu_X = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$
- metoda výpočtu se značí $\mathsf{E}\,X$

příklad **rodina**

j	m_j	x_j^*	$\mathsf{P}(X = x_j^*)$	$x_j^* \cdot \mathsf{P}(X = x_j^*)$
1	1	0	0,125	0,000
2	3	1	0,375	0,375
3	3	2	0,375	0,750
4	1	3	0,125	0,375
součet			1,000	1,500

$$\begin{aligned} \mu_X &= 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} \\ &= 0 \cdot 0{,}125 + 1 \cdot 0{,}375 + 2 \cdot 0{,}375 + 3 \cdot 0{,}125 \\ &= 1{,}5 \end{aligned}$$

rozptyl σ^2 (σ **směr. odchylka**)

- míra variability, **populační rozptyl**
- velikost kolísání kolem střední hodnoty
- metoda výpočtu se značí var X
- pomocí střední hodnoty

$$\sigma^2 = \mathsf{E}\,(X-\mu_X)^2 = \mathsf{E}\,X^2 - \mu^2$$

- diskrétní $\quad \sigma^2 = \sum_j \left(x_j^* - \mu_X\right)^2 \mathsf{P}\left(X = x_j^*\right)$

j	x_j^*	p_j	$x_j^* - \mu_X$	$\left(x_j^* - \mu_X\right)^2$	$\left(x_j^* - \mu_X\right)^2 p_j$
1	0	0,125	-1,5	2,25	0,28150
2	1	0,375	-0,5	0,25	0,09375
3	2	0,375	0,5	0,25	0,09375
4	3	0,125	1,5	2,25	0,28150
$\sum$		1,000	0,0		0,75000

$$\begin{aligned}
\sigma_X^2 &= (0-1{,}5)^2 \cdot 0{,}125 + (1-1{,}5)^2 \cdot 0{,}375 \\
&\quad +(2-1{,}5)^2 \cdot 0{,}375 + (3-1{,}5)^2 \cdot 0{,}125 \\
&= 0{,}75 \\
\sigma_X &= \sqrt{0{,}75} \quad = \quad 0{,}866025
\end{aligned}$$

sdružené rozdělení:
zajímáme se o **společné** chování dvojice (trojice,. . .) náhodných veličin, tedy chování **náhodného vektoru**

Příklad **rodina**
X počet děvčat v rodině s třemi dětmi
Y počet děvčat mezi dvěma staršími dětmi
Z počet hochů v rodině s třemi dětmi

ω_i	x_i	y_i	z_i
(m,m,m)	0	0	3
(m,m,f)	1	1	2
(m,f,m)	1	1	2
(f,m,m)	1	0	2
(f,f,m)	2	1	1
(f,m,f)	2	1	1
(m,f,f)	2	2	1
(f,f,f)	3	2	0

rozdělení náhodného vektoru (X,Y)
proč nemá smysl uvažovat **vektor** (X,Z)?

sdružené rozdělení:
popisuje **společné chování** veličin pomocí jejich **sdruženého** rozdělení:

$$\boxed{\mathsf{P}(X = x_i^*, Y = y_j^*)} \text{ resp. } \boxed{f_{X,Y}(x,y)}$$

marginální rozdělení – chování jedné veličiny

$$\boxed{\mathsf{P}(X = x_i^*) = \sum_j \mathsf{P}(X = x_i^*, Y = y_j^*), \forall x_i^*}$$

kovariance vyjadřuje závislost náh. veličin:

$$\boxed{\sigma_{X,Y} = \mathsf{E}\,(X-\mu_X)(Y-\mu_Y)}$$

označení metody výpočtu: cov(X,Y)
zřejmě platí $\boxed{\text{cov}(X,X) = \text{var}\,X}$

nezávislost náhodných veličin:

$$\boxed{\mathsf{P}(X = x_i^*, Y = y_j^*) = \mathsf{P}(X = x_i^*)\mathsf{P}(Y = y_j^*), \forall (x_i^*, y_j^*)}$$

X,Y – nezávislé $\Rightarrow$ $\boxed{\sigma_{X,Y} = 0}$
(nikoliv obráceně)

Příklad **rodina**
X počet děvčat v rodině s třemi dětmi
Y počet děvčat mezi dvěma staršími dětmi

	y_j^*			
x_i^*	0	1	2	celkem
0	0,125	0	0	0,125
1	0,125	0,250	0	0,375
2	0	0,250	0,125	0,375
3	0	0	0,125	0,125
celkem	0,250	0,500	0,250	1,000

$$\begin{aligned}
\mu_X &= 0 \cdot 0{,}125 + 1 \cdot 0{,}375 + 2 \cdot 0{,}375 + 3 \cdot 0{,}125 = 1{,}5 \\
\mu_Y &= 0 \cdot 0{,}250 + 1 \cdot 0{,}500 + 2 \cdot 0{,}250 = 1
\end{aligned}$$

X,Y – závislé, např. $0{,}25 \cdot 0{,}125 \neq 0{,}125$

výpočet kovariance $\sigma_{XY} = \text{cov}(X,Y)$:

$$\begin{aligned}
\sigma_{XY} = &\quad (0-1{,}5)\cdot(0-1)\cdot 0{,}125 \\
&+(1-1{,}5)\cdot(0-1)\cdot 0{,}125 \\
&+(1-1{,}5)\cdot(1-1)\cdot 0{,}250 \\
&+(2-1{,}5)\cdot(1-1)\cdot 0{,}250 \\
&+(2-1{,}5)\cdot(2-1)\cdot 0{,}125 \\
&+(3-1{,}5)\cdot(2-1)\cdot 0{,}125 \\
= &\qquad 0{,}5
\end{aligned}$$

střední hodnota X (mean value)

$$\begin{aligned}\mu_X &= \mathsf{E}\,X \\ &= \sum_j x_j^* \mathsf{P}(X = x_j^*) \\ &= \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x\end{aligned}$$

střední hodnota $Y = g(X)$

$$\begin{aligned}\mu_Y &= \mathsf{E}\,g(X) \\ &= \sum_j g(x_j^*) \mathsf{P}(X = x_j^*) \\ &= \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x\end{aligned}$$

rozptyl X (variance, (standard deviation)^2)

$$\begin{aligned}\sigma_X^2 &= \mathrm{var}\,X = \mathsf{E}\,(X - \mu_X)^2 \\ &= \sum_j (x_j^* - \mu_X)^2 \mathsf{P}(X = x_j^*) \\ &= \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) \mathrm{d}x\end{aligned}$$

kovariance X a Y (covariance)

$$\begin{aligned}\sigma_{X,Y} &= \mathrm{cov}(X, Y) = \mathsf{E}\,(X - \mu_X)(Y - \mu_Y) \\ &= \sum_{i,j} (x_i^* - \mu_X)(y_j^* - \mu_Y) \mathsf{P}(X = x_i^*, Y = y_j^*) \\ &= \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y) \mathrm{d}x\mathrm{d}y\end{aligned}$$

vlastnosti populačního průměru a rozptylu

$$\begin{aligned}\mu_{\alpha+\beta X} &= \alpha + \beta \mu_X, \\ \sigma^2_{\alpha+\beta X} &= \beta^2 \sigma_X^2, \\ \sigma_{\alpha+\beta X} &= |\beta| \sigma_X, \\ \mu_{X+Y} &= \mu_X + \mu_Y, \\ \sigma^2_{X+Y} &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}.\end{aligned}$$

ukázka důkazu:

$$\begin{aligned}\mu_{\alpha+\beta X} &= \mathsf{E}\,(\alpha + \beta X) \\ &= \sum_i (\alpha + \beta x_i^*) \mathsf{P}(X = x_i^*) \\ &= \sum_i \alpha \mathsf{P}(X = x_i^*) + \sum_i \beta x_i^* \mathsf{P}(X = x_i^*) \\ &= \alpha \sum_i \mathsf{P}(X = x_i^*) + \beta \sum_i x_i^* \mathsf{P}(X = x_i^*) \\ &= \alpha + \beta\ \mathsf{E}\,X = \alpha + \beta\ \mu_X\end{aligned}$$

jsou-li X, Y **nezávislé**, pak

$$\begin{aligned}\sigma_{XY} &= 0 \\ \sigma^2_{X+Y} &= \sigma_X^2 + \sigma_Y^2\end{aligned}$$

normování náhodné veličiny X

$$Z = \frac{X - \mu_X}{\sigma_X} \qquad \text{bezrozměrné!}$$

$$\Rightarrow \qquad \mu_Z = 0 \quad \sigma_Z = 1$$

vlastnosti nezávislé na μ_X, σ_X^2:
(populační) **korelační koeficient**
(correlation coefficient)

$$\begin{aligned}\rho_{XY} &= \mathrm{cov}\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) \\ &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y}\end{aligned}$$

(populační) **šikmost** náhodné veličiny X
(skewness)

$$\begin{aligned}\gamma_1 &= \mathsf{E}\left(\frac{X - \mu_X}{\sigma_X}\right)^3 \\ &= \frac{\mathsf{E}\,(X - \mu_X)^3}{\sigma_X^3}\end{aligned}$$

(populační) **špičatost** náhodné veličiny X
(kurtosis, někdy se neodečítá 3)

$$\begin{aligned}\gamma_2 &= \mathsf{E}\left(\frac{X - \mu_X}{\sigma_X}\right)^4 - 3 \\ &= \frac{\mathsf{E}\,(X - \mu_X)^4}{\sigma_X^4} - 3\end{aligned}$$

Důležitá diskrétní rozdělení

alternativní (nula-jedničkové) rozdělení

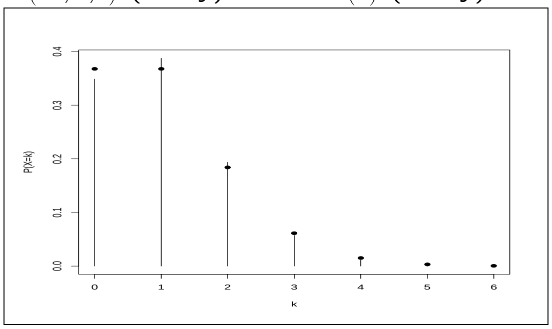
- *zdar* nebo *nezdar*
- $\mathsf{P}(X = 1) = \pi$, $\mathsf{P}(X = 0) = 1 - \pi$, $(0 < \pi < 1)$
- $\mathsf{E}\,X = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi$
- $\mathrm{var}\,X = (1 - \pi)^2 \cdot \pi + (0 - \pi)^2 \cdot (1 - \pi) = \pi(1 - \pi)$

binomické rozdělení $Y \sim \mathrm{bi}(n, \pi)$

- n **nezávislých** pokusů
- $\mathsf{P}(\text{zdar}) = \pi$, $\mathsf{P}(\text{nezdar}) = 1 - \pi$, $(0 < \pi < 1)$
- Y je počet zdarů v těchto pokusech
- $\mathsf{P}(Y = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}, \quad k = 0, 1, \ldots, n$
- $Y = \sum_{i=1}^n X_i$, X_i – zda zdar v i-tém pokusu
- $\mathsf{E}\,Y = \mathsf{E}\,(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathsf{E}\,X_i = n\pi$
- $\mathrm{var}\,Y = \mathrm{var}\,(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{var}\,X_i$
  $= n\pi(1 - \pi)$ (nezávislost X_i!)

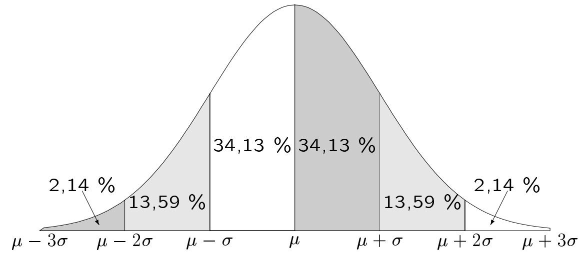
Poissonovo rozdělení $X \sim \mathsf{Po}(\lambda)$

- zákon vzácných (řídkých) jevů
- kolikrát nastal jev během jednotkového časového intervalu, na jednotkové ploše, v jednotkovém objemu ...
- $\mathsf{P}(X = k) = \frac{\lambda^k}{k!}\mathsf{e}^{-\lambda}, \qquad k = 0, 1, \ldots$
- $\mathsf{E}\, X = \lambda$, $\mathsf{var}\, X = \lambda$
- pro velké n a malé π lze rozdělení $\mathsf{bi}(n, \pi)$ aproximovat pomocí rozdělení $\mathsf{Po}(n\pi)$
- $\mathsf{bi}(10, 0{,}1)$ (hůlky) vers. $\mathsf{Po}(1)$ (tečky)



normální (Gaussovo) rozdělení $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$

- $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathsf{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$,
- $\mathsf{E}\, X = \mu$, $\mathsf{var}\, X = \sigma^2$



- $\mathsf{N}(0, 1)$: $\varphi(x) = \frac{1}{\sqrt{2\pi}}\mathsf{e}^{x^2/2}$, $\Phi(x) = \int_{-\infty}^{x} \varphi(t)\mathsf{d}t$
- $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$, pak

$$\mathsf{P}(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

- V má **logaritmicko-normální** rozdělení:

$$\ln V \sim \mathsf{N}\left(\mu, \sigma^2\right)$$

- aproximace binomického rozdělení $\mathsf{bi}(n, \pi)$ pomocí $\mathsf{N}(n\pi, n\pi(1-\pi)) \qquad (n\pi(1-\pi) > 9)$

- kritické hodnoty normálního rozdělení

$$Z \sim \mathsf{N}(0, 1): \qquad \mathsf{P}(Z > z(\alpha)) = \alpha$$

ze symetrie platí $\mathsf{P}(|Z| > z(\alpha/2)) = \alpha$

- kritické hodnoty Studentova t rozdělení

$$T \sim \mathsf{t}(k) : \mathsf{P}(|T| > t_k(\alpha)) = \alpha$$

α	0,10	0,05	0,01
$z(\alpha/2)$	1,645	1,960	2,576
$t_{100}(\alpha)$	1,660	1,984	2,626
$t_{20}(\alpha)$	1,725	2,086	2,845
$t_5(\alpha)$	2,015	2,571	4,032

- kritické hodnoty Fisherova F rozdělení

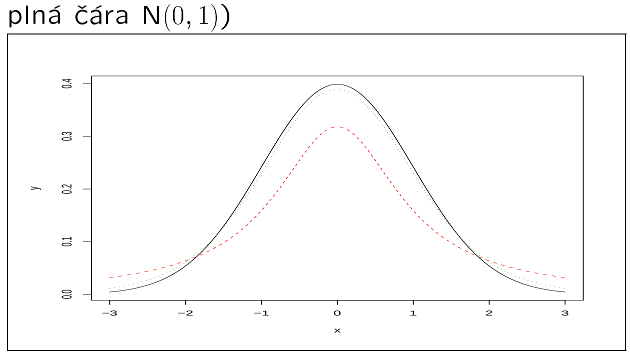
$$F \sim \mathsf{F}(k, m) : \mathsf{P}(F > F_{k,m}(\alpha)) = \alpha$$

- kritické hodnoty rozdělení chí-kvadrát

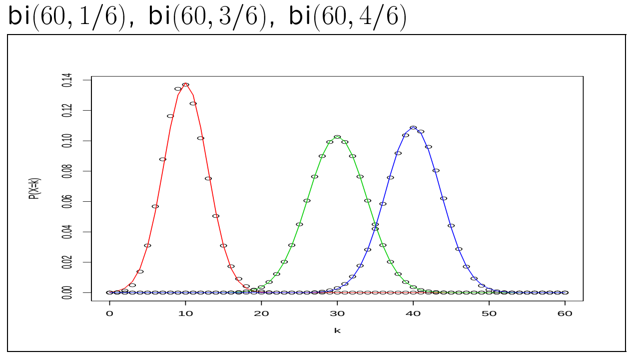
$$X^2 \sim \chi^2(k) : \mathsf{P}(X^2 > \chi^2_k(\alpha)) = \alpha$$

$\chi^2_1(0{,}05) = 1{,}960^2 = 3{,}841$

srovnání normálního a Studentova rozdělení
čárkovaně $\mathsf{t}(1)$, tečkovaně $\mathsf{t}(10)$,
plná čára $\mathsf{N}(0, 1)$)



srovnání binomického a normálního rozdělení
$\mathsf{bi}(60, 1/6)$, $\mathsf{bi}(60, 3/6)$, $\mathsf{bi}(60, 4/6)$



populace – výběr

- **populace (základní soubor)**
  soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku)
  $\Rightarrow$ rozdělení náhodné veličiny
- **výběr**
  náhodně vybraná část populace, kterou vyšetřujeme, vzorek populace
- **náhodný výběr**
  nezávislé náhodné veličiny se stejným rozdělením (neměřené na výběru)
- **parametr**
  neznámé číslo popisující nějaký rys populace, charakteristika rozdělení náh. vel.
- **statistika**
  funkce náhodného výběru
- **odhad**
  statistika použitá k odhadu parametru

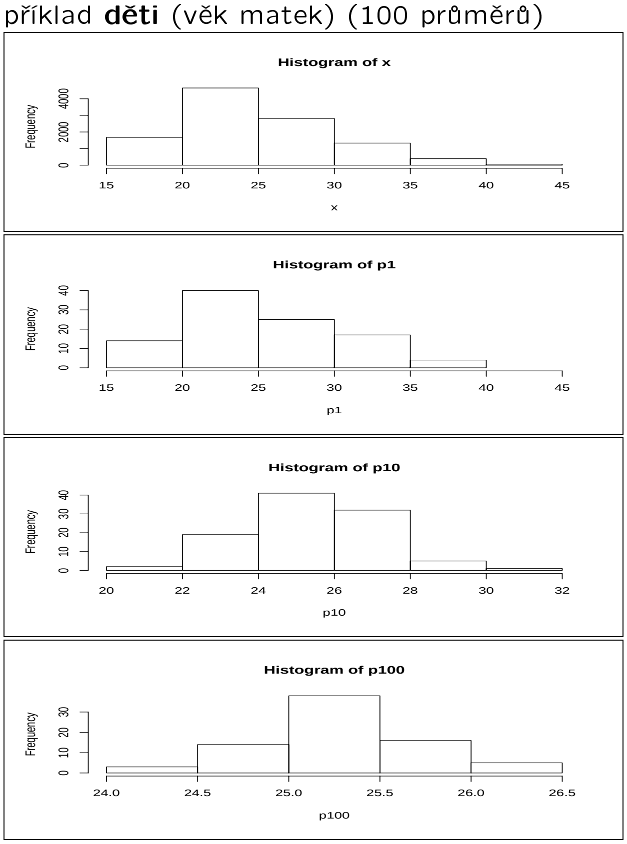
- $X_1, \dots, X_n$ nezávislé, stejné rozdělení
  $\mathsf{E}\, X_i = \mu$ populační průměr
  $\operatorname{var} X_i = \sigma^2$ populační rozptyl
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ výběrový průměr
- $\mathsf{E}\, \bar{X} = \mu$ výběrový průměr
  je **nestranným** odhadem populačního
- $\operatorname{var} \bar{X} = \frac{\sigma^2}{n}$ $= (\text{S.E.}(\bar{X}))^2$
  n-krát menší, než u jednoho pozorování!
- u **normálního** rozdělení: $X_i \sim \mathsf{N}\left(\mu, \sigma^2\right)$

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim \mathsf{N}(0,1)$$

- **interval spolehlivosti** pro μ:
  $(\bar{X} - \text{S.E.}(\bar{X}) z(\alpha/2),\ \bar{X} + \text{S.E.}(\bar{X}) z(\alpha/2))$
  $\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z(\alpha/2),\ \bar{X} + \frac{\sigma}{\sqrt{n}} z(\alpha/2)\right)$
- požadujeme int. spolehlivosti šířky $2c\sigma$:

$$n \geq \left(\frac{z(\alpha/2}{c}\right)^2$$

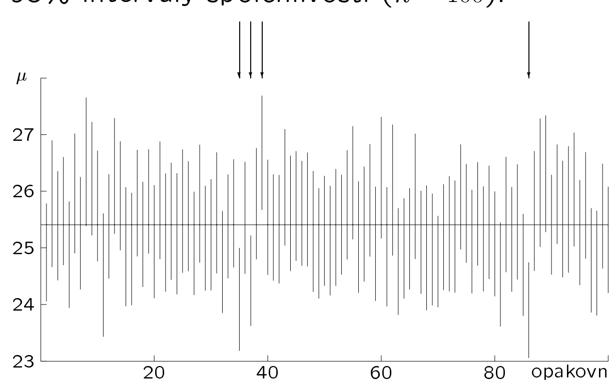
příklad **děti** (věk matek) (100 průměrů)



průměrný věk matek v opak. výběrech:

rozsah výběru n	průměr průměrů	směr. odch. průměrů	šikmost průměrů	špičatost průměrů
1	26,42	5,182	0,529	-0,679
10	25,56	1,475	0,140	-0,771
100	25,30	0,529	0,027	-0,303
1000	25,40	0,158	-0,040	-0,284
populace	25,40	4,943	0,773	0,192

95% intervaly spolehlivosti ($n = 100$):



statistické rozhodování

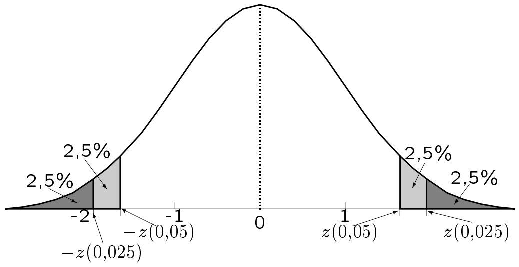
- **nulová hypotéza** H_0
  tvrzení o populaci (parametru), o jehož platnosti chceme rozhodnout, zpravidla zamítnout
- **alternativní hypotéza** H_1 (alternativa)
  zbývající možnost (k H_0)
- **kritický obor**
  možné výsledky pokusu, kdy H_0 zamítáme
- **obor přijetí**
  možné výsledky pokusu, kdy H_0 nezamítáme
- **chyba prvního druhu**
  rozhodnutí zamítnout H_0, když platí H_0
- **chyba druhého druhu**
  rozhodnutí nezamítnout H_0, když platí H_1
- **hladina testu** α (zpravidla 5 %, 1 %)
  maximální dovolená pst chyby prvního druhu

rozhodnutí	skutečnost: H_0 platí	skutečnost: H_0 neplatí
H_0 zamítnout (reject)	chyba 1. druhu ($\leq \alpha$)	správné rozhodnutí ($1-\beta$)
H_0 nezamítnout (accept)	správné rozhodnutí ($\geq 1-\alpha$)	chyba 2. druhu (β)

- hladina testu α se volí před pokusem (aby nezávisela na jeho výsledku)
- **síla testu** $1-\beta$
  pravděpodobnost zamítnutí neplatné H_0
- kritický obor zpravidla popsán pomocí statistiky (např. $|T| \geq t_{n-1}(\alpha)$)
- **dosažená hladina testu** p (p value)
  za platnosti H_0 určená pst, že dostaneme statistiku, která stejně nebo ještě méně podporuje H_0 (nejmenší hladina α, na které lze ještě H_0 zamítnout),
  např. $p = \mathsf{P}(|T| \geq t)$, kde t je skutečně realizovaná hodnota statistiky Z
- H_0 se **zamítá**, když $\boxed{p \leq \alpha}$

rozhodování o populačním průměru normálního rozdělení se známým rozptylem

- $X_1, \ldots, X_n \sim \mathsf{N}\left(\mu, \sigma^2\right)$ **nezávislé**
- $\sigma > 0$ známe
- $H_0 : \mu = \mu_0$ (dané číslo)
- platí-li H_0, pak
$$Z = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim \mathsf{N}(0,1)$$
- $H_1 : \mu \neq \mu_0 \Rightarrow$ kritický obor:
  $|Z|$ velké, tj. $|Z| \geq z(\alpha/2)$
- $H_1 : \mu > \mu_0$: zamítnout pro $Z \geq z(\alpha)$
- $H_1 : \mu < \mu_0$: zamítnout pro $Z \leq -z(\alpha)$



Hustota Z za platnosti H_0

příklad **výšky** desetiletých hochů ([cm])

130	140	136	141	139
133	149	151	139	136
138	142	127	139	147

$\sigma = 6{,}4$(známo z dřívějška), $\alpha = 0{,}05$
$H_0 : \mu = 136{,}1$ (před 10 lety), $H_1 : \mu \neq 136{,}1$

$$\bar{x} = \frac{1}{15}(130 + 140 + \ldots + 147) = 139{,}133$$
$$z = \frac{139{,}133 - 136{,}1}{6{,}4}\sqrt{15} = 1{,}835$$

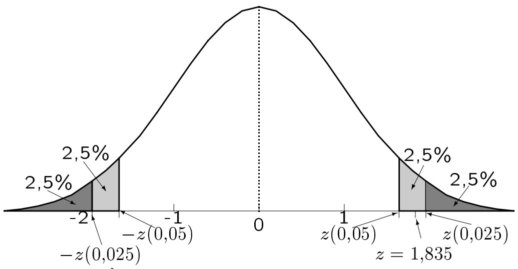
$|z| < z(0{,}05/2) = 1{,}960$
$\Rightarrow H_0$ nelze na 5% hladině zamítnout

ale $|z| \geq z(0{,}10/2) = 1{,}645$
$\Rightarrow H_0$ se na 10% hladině zamítá

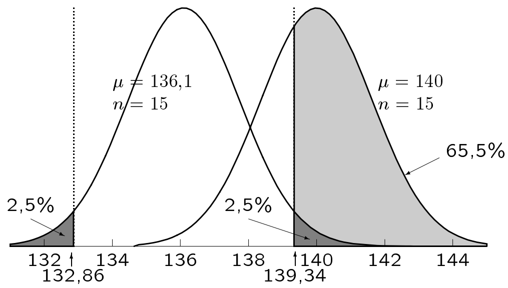
$p = \mathsf{P}(|Z| \geq 1{,}835) = 0{,}067$
dosažená hladina (p value) je 6,7 %

jednostranná alternativa (předem!):
$H_1 : \mu > 136{,}1$: $z \geq 1{,}645$ na 5 % zamítnout
$p = \mathsf{P}(Z \geq 1{,}835) = 0{,}033$ ($< 0{,}05$)

výšky desetiletých hochů



a) Hustota Z za platnosti H_0

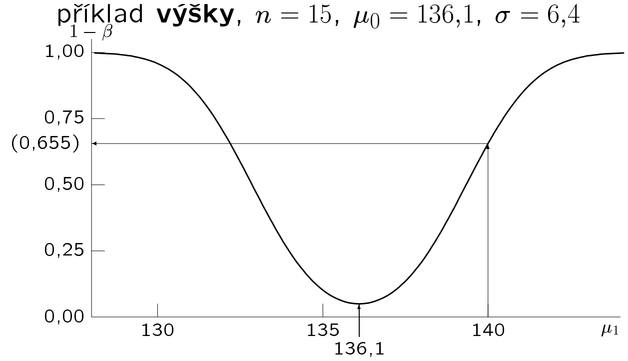


b) Hustota $\bar{X}$ při H_0 a při $\mathsf{H}_1 : \mu = 140,\ \sigma = 6{,}4$

$$\text{S.E.}(\bar{X}) = \sqrt{\frac{6{,}4^2}{15}} = 1{,}6525 \Rightarrow 136{,}1 - 1{,}6525 \cdot 1{,}96 = 132{,}86$$

$$\Rightarrow 136{,}1 + 1{,}6525 \cdot 1{,}96 = 139{,}34$$

síla testu $1 - \beta$
pravděpodobnost, že zamítneme nulovou hypotézu, když testovaný parametr je roven ...
(závisí na skutečné hodnotě parametru)
příklad **výšky**, $n = 15$, $\mu_0 = 136{,}1$, $\sigma = 6{,}4$



volba rozsahu výběru: pro μ_1 požadujeme sílu $1 - \beta$:

$$n \geq \left(\frac{z(\alpha/2) + z(\beta)}{\mu_1 - \mu_0}\right)^2 \sigma^2$$

aby pro $\mu_1 = 140$ byla síla 90 % ($z(0{,}1) = 1{,}282$), bude třeba aspoň

$$n \geq \left(\frac{1{,}96 + 1{,}282}{140 - 136{,}1}\right)^2 6{,}4^2 = 28{,}3$$

jednovýběrový t test

- n nezávislých pozorování $X_1, \ldots, X_n$
- stejné normální rozdělení $\mathsf{N}\big(\mu, \sigma^2\big)$
- $\mathsf{H}_0 : \mu = \mu_0$
  (populační průměr roven dané konstantě)
- nutno odhadnout neznámý rozptyl σ^2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- statistika

$$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n} \qquad = \frac{\bar{X} - \mu_0}{\text{S.E.}(\bar{X})}$$

- $\mathsf{H}_1 : \mu \neq \mu_0$ zamítat při $|T| \geq t_{n-1}(\alpha)$
- $\mathsf{H}_1 : \mu > \mu_0$ zamítat při $T \geq t_{n-1}(2\alpha)$
- $\mathsf{H}_1 : \mu < \mu_0$ zamítat při $T \leq -t_{n-1}(2\alpha)$
- interval spolehlivosti pro μ

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha), \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha)\right)$$

případ neznámého rozptylu ($\mathsf{H}_1 : \mu \neq 136{,}1$)

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x})^2 &= \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \\
&= 130^2 + \ldots + 147^2 - 15 \cdot 139{,}133^2 \\
&= 601{,}733 \\
s^2 &= \frac{601{,}733}{15 - 1} \\
&= 42{,}981 \qquad = 6{,}556^2 \\
t &= \frac{139{,}133 - 136{,}1}{6{,}556}\sqrt{15} \\
&= 1{,}792 \\
p &= \mathsf{P}(|T| \geq 1{,}792) = 0{,}0948 \qquad (9{,}48\ \%)
\end{aligned}$$

95% interval spolehlivosti ($t_{14}(0{,}05) = 2{,}145$):

$$\left(139{,}133 - \frac{6{,}556}{\sqrt{15}} \cdot 2{,}145 \quad , \quad 139{,}133 + \frac{6{,}556}{\sqrt{15}} \cdot 2{,}145\right)$$

$$(135{,}5 \quad , \quad 142{,}8)$$

jednostranná alternativa $\mathsf{H}_1 : \mu > 136{,}1$:
$t \geq t_{14}(2 \cdot 0{,}05) = 1{,}761$ zamítnout H_0 ($\alpha = 5\%$)
$t < t_{14}(2 \cdot 0{,}01) = 2{,}624$ nezamítnout H_0 ($\alpha = 1\%$)
$p = \mathsf{P}(T > t) = 0{,}0474$ $(4{,}74\ \%)$

pst výskytu jevu (binomické rozdělení)

- n **nezávislých** opakování dílčího pokusu
- v každém „zdar“ s pstí π
- počet zdarů $Y \sim \text{bi}(n, \pi)$
- odhad π: $\hat{\pi} = \dfrac{Y}{n}$ (relativní četnost)

  $\mathsf{E}\,\hat{\pi} = \pi, \qquad \text{var}\,\hat{\pi} = \frac{\pi(1-\pi)}{n} \qquad = (\text{S.E.}(\hat{\pi}))^2$
- intervalový odhad pro π (přibližný)

$$\left(\hat{\pi} - \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}z(\alpha/2),\ \hat{\pi} + \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}z(\alpha/2)\right)$$

- $\text{H}_0 : \pi = \pi_0$:

$$Z = \frac{Y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\text{S.E.}(\hat{\pi})}$$

- $\text{H}_1 : \pi \neq \pi_0$: zamítnout $\qquad |Z| \geq z(\alpha/2)$
- $\text{H}_1 : \pi > \pi_0$: zamítnout $\qquad Z \geq z(\alpha)$
- $\text{H}_1 : \pi < \pi_0$: zamítnout $\qquad Z \leq -z(\alpha)$

příklad **kalous**

z 50 případů dal kalous ve 33 případech přednost infikované myši před neinfikovanou

Y – počet „zdarů“, $n = 50$, π – pst, že zvolí infikovanou $\Rightarrow$ Y má **binomické rozdělení**

za $\text{H}_0 : \pi = 1/2$ (myši se neliší) $Y \sim \text{bi}(50, 1/2)$
alternativní hypotéza: $\text{H}_1 : \pi > 1/2$:

kritický obor: velká hodnota Y (velké $\hat{\pi}$)

$$z = \frac{33 - 50 \cdot 0{,}5}{\sqrt{50 \cdot 0{,}5 \cdot 0{,}5}} = 2{,}263 \qquad p = 0{,}0118$$

s opravou na spojitost (NCSS):

$$z = \frac{33 - 50 \cdot 0{,}5 - 0{,}5}{\sqrt{50 \cdot 0{,}5 \cdot 0{,}5}} = 2{,}121 \qquad p = 0{,}0169$$

dosažená hladina: za H_0 počítaná pst, že dostaneme výsledek aspoň tolik odporující nulové hypotéze, jako ve skutečném pokusu:

$$\begin{aligned} p &= \mathsf{P}(Y \geq 33) \\ &= \sum_{k=33}^{50} \binom{50}{k} 0{,}5^k (1-0{,}5)^{50-k} \\ &= 0{,}0164 \\ &= \mathsf{P}(Y > 32) \qquad \text{(NCSS, Prob. Calc.)} \end{aligned}$$

dvouvýběrový t test

- n_X nezávislých pozorování X
- n_Y nezávislých pozorování Y
- tyto výběry **nezávislé**
- rozptyly σ_X^2, σ_Y^2 shodné
  (odhady S_X^2, S_Y^2 podobné, lze ověřit)
- normální rozdělení v obou výběrech
  (lze ověřit pro velká n_X, n_Y,
  jinak podle zkušenosti)
- společný odhad rozptylu

$$S^2 = \frac{n_X - 1}{n_X + n_Y - 2}S_X^2 + \frac{n_Y - 1}{n_X + n_Y - 2}S_Y^2$$

- statistika

$$T = \frac{\bar{X} - \bar{Y}}{\text{S.E.}(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{S}\sqrt{\frac{n_X n_Y}{n_X + n_Y}}$$

- $\text{H}_0 : \mu_X = \mu_Y$ zamítnout ve prospěch alternativy $\text{H}_1 : \mu_X \neq \mu_Y$:

$$|T| \geq t_{n_X + n_Y - 2}(\alpha)$$

příklad **výšky** dětí (opět [cm])

hoši: $\quad n_x = 15,\ \bar{x} = 139{,}133,\ s_x^2 = 42{,}981$
dívky: $\quad n_y = 12,\ \bar{y} = 140{,}833,\ s_y^2 = 33{,}788$

H_0: shodné populační průměry, H_1: neshodné

$$\begin{aligned} s^2 &= \frac{14}{25}42{,}981 + \frac{11}{25}33{,}788 \\ &= 38{,}936 \end{aligned}$$

odhad $\text{S.E.}(\bar{X} - \bar{Y})$:

$$\sqrt{38{,}936\,\frac{15 + 12}{15 \cdot 12}} = \sqrt{5{,}8404} = 2{,}4167$$

$$\begin{aligned} t &= \frac{139{,}133 - 140{,}833}{\sqrt{38{,}936}}\sqrt{\frac{15 \cdot 12}{15 + 12}} \\ &= \frac{-1{,}7}{2{,}4167} \quad = \quad -0{,}703 \end{aligned}$$

$|t| < t_{25}(0{,}05) = 2{,}0595 \quad \Rightarrow$
na 5% hladině nezamítat
$p = 0{,}488$
95% int. spol. pro rozdíl popul. průměrů:

$$(-1{,}700 - 2{,}4167 \cdot 2{,}0595, \quad -1{,}700 + 2{,}4167 \cdot 2{,}0595)$$
$$(-6{,}7, \quad 3{,}3)$$

nula **je** intervalem pokryta

- test **Mannův-Whitneyův** (dvouvýběrový Wilcoxonův)
  - nahradí pozorování jejich pořadími
  - dva nezávislé výběry rozsahu n_X, n_Y
  - spojitá rozdělení
  - hypotéza: rozdělení jsou stejná, pak jsou výběry „dobře promíchané“
  - urči pořadí všech (promíchaných)
  - kritický obor: různá průměrná pořadí
  - W_X součet pořadí hodnot X

$$Z = \frac{W_X - n_X(n_X + n_Y + 1)/2}{\sqrt{n_X n_Y (n_X + n_Y + 1)/12}}$$

  - shodu zamítni pokud $|Z| \geq z(\alpha/2)$ (přibližný test)
  - citlivý vůči posunutí, nikoliv vůči nestejné variabilitě

dvouvýběrový Wilcoxonův (Mann-Whitney)

hoši	dívky	pořadí
127		1
130		2
	131	3
	132	4
133		5
	135	6
136 136		7,5
138		9
139 139 139		11
140		13
141	141 141 141 141	16
142	142	19,5
	143	21
	146 146	22,5
147		24
149		25
151	151	26,5

$$w_x = 1 + 2 + 5 + 2 \cdot 7{,}5 + 9 + 3 \cdot 11 + 13 + 16 + 19{,}5 + 24 + 25 + 26{,}5 = 189$$

$$w_y = 3 + 4 + 6 + 4 \cdot 16 + 19{,}5 + 21 + 2 \cdot 22{,}5 + 26{,}5 = 189$$

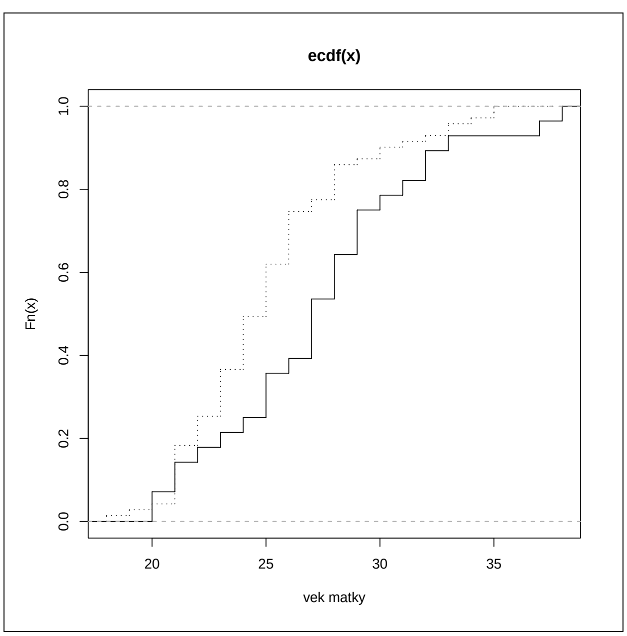
$$z = \frac{189 - 15 \cdot (15 + 12 + 1)/2}{\sqrt{15 \cdot 12(15 + 12 + 1)/12}} = -1{,}025$$

$$p = 0{,}3055$$

NCSS: $z = -1{,}029$ (korekce na shody)
$p = 0{,}3036$

přesně: $p = 0{,}3149$

- test **Kolmogorovův-Smirnovův**
  - porovná empirické distribuční funkce
  - citlivý vůči všem neshodám



párové testy

- $(U_1, V_1), \ldots, (U_n, V_n)$ **nezávislé** dvojice (možná závislých) náhodných veličin
- výhodná je těsná závislost uvnitř dvojic
- $X_i = U_i - V_i$ (označení rozdílů) $X_1, \ldots, X_n$ mají **stejné** rozdělení
- **párový t test**
  - **normální** rozdělení: $X_i \sim \mathsf{N}\left(\mu, \sigma^2\right)$
  - $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
  - $T = \frac{\bar{X}}{\mathsf{S.E.}(\bar{X})} = \frac{\bar{X}}{S}\sqrt{n} = \frac{\bar{U} - \bar{V}}{\mathsf{S.E.}(\bar{U} - \bar{V})}$
  - $\mathsf{H}_0 : \mu = 0$ (pak je $\mu_U = \mu_V$)
  - ve prospěch H_1:$\mu \neq 0$, když $|T| \geq t_{n-1}(\alpha)$
  - ve prospěch H_1:$\mu < 0$, když $T \leq -t_{n-1}(2\alpha)$
  - ve prospěch H_1:$\mu > 0$, když $T \geq t_{n-1}(2\alpha)$
  - jednovýběrový t test pro $X_i = U_i - V_i$

příklad: výšky rodičů (párová pozorování!)

- U – výška otce, V – výška matky
- $\alpha = 0{,}05, \mathsf{H}_0 : \mu_U - 10 = \mu_V$
- $n = 99, \bar{u} = 179{,}267, \bar{v} = 166{,}970$
- $\bar{x} = 2{,}293, s_X = s_{U-10-V} = s_{U-V} = 8{,}144$
- $t = \frac{2{,}293}{8{,}144}\sqrt{99} = 2{,}801$
- $t_{98}(0{,}05) = 1{,}9845$ $\Rightarrow$ zamítnout
- $p = \mathsf{P}(|T| \geq t) = 0{,}0061$ $(0{,}61\ \%)$
- 95% interval spolehlivosti pro $\mu_U - \mu_V$:

$$\left(12{,}293 - \frac{8{,}144}{\sqrt{99}}1{,}9845; 12{,}293\frac{8{,}144}{\sqrt{99}}1{,}9845\right)$$

$$(10{,}67; 13{,}92)$$

- 99% interval spolehlivosti: (10,14; 14,44)

- **znaménkový test**
  - stačí znát znaménka rozdílů $U_i - V_i$
  - pozorování s $U_i = V_i$ se zpravidla vynechají
  - Y – počet kladných znamének
  - H_0 : rozdělení U a V jsou stejná, pak je nutně $Y \sim \mathsf{bi}(n, 1/2)$
  - H_0 zamítáme pro velká nebo malá Y:

$$Z = \frac{|Y - n/2| - 0{,}5}{\sqrt{n/4}}, \qquad |Z| \geq z(\alpha/2)$$

- **jednovýběrový Wilcoxonův test**
  - nutné **symetrické** rozdělení $U_i - V_i$
  - vyloučíme případy $U_i = V_i$
  - určíme pořadí R_i^+ hodnot $|U_i - V_i|$
  - W součet pořadí, kde $U_i > V_i$

$$Z = \frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

příklad rozdíly dvou metod učení nazpaměť:

$$5, -1, 2, 3, -1, 4, 3, -3$$

- znaménkový test

$$y = 5 \qquad n = 8$$

$$z = \frac{|5 - 8/2| - 0{,}5}{\sqrt{8/4}} = 0{,}3536 \qquad p = 0{,}7237$$

- Wilcoxonův test
  (předpokládáme symetrii)

$u_i - v_i$	5	-1	2	3	-1	4	3	-3
r_i^+	8	1,5	3	5	1,5	7	5	5

$$w = 8 + 3 + 5 + 7 + 5 = 28$$

$$z = \frac{28 - 8 \cdot 9/4}{\sqrt{8 \cdot 9 \cdot 17/24}} = \frac{10}{\sqrt{51}} = 1{,}4$$

$$p = 0{,}1614$$

permutační testy - dva výběry

- H_0 : identické rozdělení v obou populacích
- příklad **hnojení**
  - x: 50,45,42,54 (klasicky)
  - y: 59,56,58,51,52 (nové)
  - H_0 stejné výnosy
  - H_1 výnosy jsou nestejné
- porovnejme průměry:

$$\bar{x} - \bar{y} = 47{,}75 - 55{,}2 = -7{,}45$$

- celkem $\binom{9}{4} = 126$ permutací – možností, kolikrát vybrat 4 hodnoty x z 9 hodnot
  - mezi nimi jsou 4 takové, že rozdíl průměrů nejvýše –7,45, což je 3,17 %
  - při oboustranné alternativě další dvě kombinace, kdy rozdíl aspoň 7,45, což je 1,59 % permutací, celkem

$$p = \frac{4+2}{126} = \frac{6}{126} = 0{,}0476 \qquad (4{,}76\ \%)$$

x				y					$\bar{x}-\bar{y}$	w_x
50	45	42	54	59	56	58	51	52	-7,45	
3	2	1	6	9	7	8	4	5		10
*	*	*					*		**−8,80**	**10**
*	*	*						*	**−8,35**	**11**
	*	*					*	*	**−7,90**	**12**
*	*	*	*						**−7,45**	**12**
	*	*	*				*		−7,00	13
	*	*	*					*	−6,55	14
*	*	*			*				−6,55	13
	*	*			*		*		−6,10	14
	...					...				
*	*			*				*	−0,70	19
		*			*	*	*		−0,25	20
		*	*	*				*	−0,25	21
	*			*			*	*	−0,25	20
	*		*		*			*	−0,25	20
*			*				*	*	−0,25	18
*		*		*	*				−0,25	20
*	*		*			*			−0,25	19
		*			*	*		*	0,20	21
		*		*	*		*		0,20	21
	...					...				
			*	*		*	*		6,50	27
			*	*		*		*	6,95	**28**
*				*	*	*			6,95	27
				*	*	*	*		7,40	**28**
				*	*	*		*	**7,85**	**29**
			*	*	*	*			**8,75**	**30**

$$p_{\text{perm}} = \frac{4+2}{126} = 0{,}0476 \qquad p_W = \frac{4+4}{126} = 0{,}0635$$

$$t = -2{,}5238 \qquad p = 0{,}0396$$

permutační testy - jeden výběr
příklad **učení nazpaměť**

- H_0 : rozdělení je symetrické kolem nuly
- rozdíly dvou metod učení nazpaměť:

  $5, -1, 2, 3, -1, 4, 3, -3$ průměr $=1{,}5$
- pokud jsou obě metody ekvivalentní, pak mají rozdíly náhodná znaménka
- případnou nulu lze předem vyloučit
- pro znaménka celkem $2^8 = 256$ možností
- ideál pro průměr 0
- v 27 případech průměr aspoň 1,5,
  v 27 případech průměr nejvýše -1,5
- dosažená hladina je rovna pravděpodobnosti, že aspoň stejně tak daleko od hypotézy, jako skutečná data

  $p_{\text{perm}} = \frac{27+27}{256} = \frac{54}{256} = 0{,}2109$ (21,09 %)

	data								$\bar{x}$	w
x	5	−1	2	3	−1	4	3	−3		
r	8	1,5	3	5	1,5	7	5	5		
1	-5	−1	−2	−3	−1	−4	−3	−3	**−2,75**	**0**
2	−5	−1	−2	−3	1	−4	−3	−3	**−2,50**	**1,5**
3	−5	1	−2	−3	−1	−4	−3	−3	**−2,50**	**1,5**
4	−5	−1	2	−3	−1	−4	−3	−3	**−2,25**	**3**
5	−5	1	−2	−3	1	−4	−3	−3	**−2,25**	**3**
					...					
18	−5	1	−2	−3	−1	−4	−3	3	**−1,75**	**6,5**
19	5	−1	−2	−3	−1	−4	−3	−3	**−1,50**	**8**
					...					
27	−5	1	−2	−3	1	-4	−3	3	**−1,50**	**8**
28	5	−1	−2	−3	1	−4	−3	−3	−1,25	9,5
					...					
229	−5	1	2	3	−1	4	3	3	1,25	26,5
230	5	−1	2	3	−1	4	3	−3	**1,50**	**28**
231	5	−1	2	3	−1	4	−3	3	**1,50**	**28**
232	5	−1	2	3	1	−4	3	3	**1,50**	27,5
233	5	−1	2	−3	−1	4	3	3	**1,50**	**28**
234	5	1	2	3	−1	−4	3	3	**1,50**	27,5
235	5	1	−2	3	1	4	3	−3	**1,50**	**28**
236	5	1	−2	3	1	4	−3	3	**1,50**	**28**
237	5	1	−2	−3	1	4	3	3	**1,50**	**28**
238	−5	1	2	3	1	4	3	3	**1,50**	**28**
239	5	−1	2	3	1	4	3	−3	**1,75**	**29,5**
					...					
255	5	1	2	3	−1	4	3	3	**2,50**	**34,5**
256	5	1	2	3	1	4	3	3	**2,75**	**36**

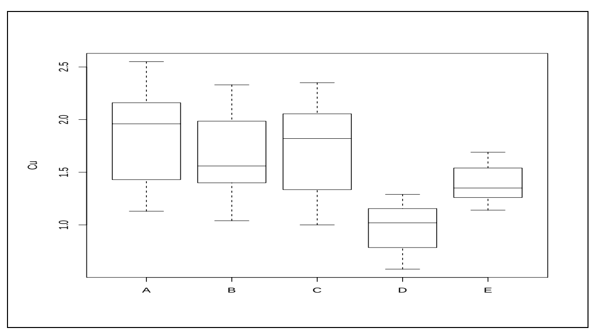
$$p_{\text{perm}} = \frac{27+27}{256} = 0{,}2109$$

$$p_W = \frac{25+25}{256} = 0{,}1953$$

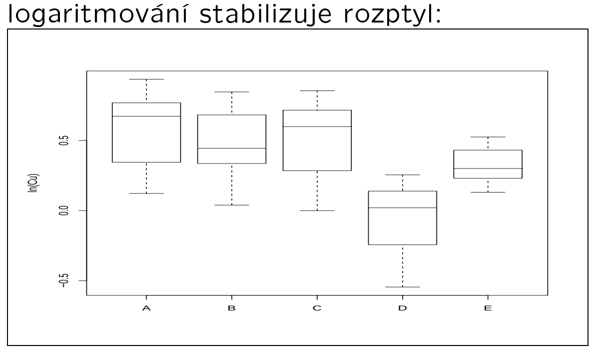
$$t = 1{,}5 \qquad p = 0{,}1773$$

příklad: játra

pět míst na řece, vždy vyloveno po 7 rybách,
zjišťována koncentrace mědi v játrech
liší se tato místa svým znečištěním?



logaritmování stabilizuje rozptyl:



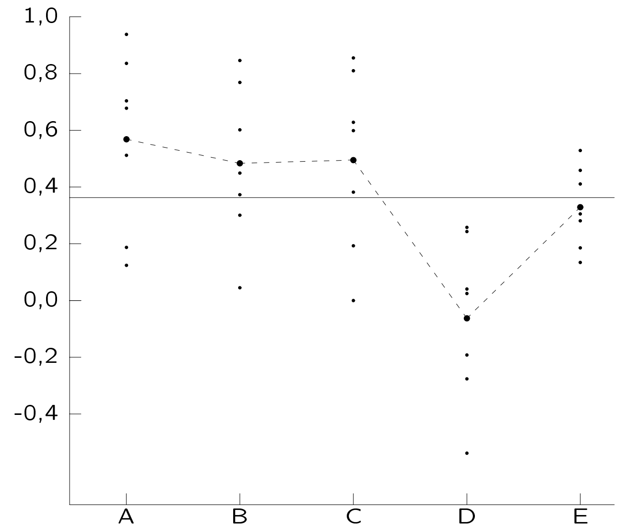
analýza rozptylu jednoduchého třídění

- $Y_{11}, \ldots, Y_{1n_1} \sim \mathsf{N}\left(\mu_1, \sigma^2\right)$
  $Y_{21}, \ldots, Y_{2n_2} \sim \mathsf{N}\left(\mu_2, \sigma^2\right)$
  $\ldots$
  $Y_{k1}, \ldots, Y_{kn_k} \sim \mathsf{N}\left(\mu_k, \sigma^2\right)$
- **nezávislé** výběry
  (shodné rozptyly, normální rozdělení)
- $\mathrm{H}_0 : \mu_1 = \ldots = \mu_k \quad (= \mu)$
  H_1 : neplatí H_0
- rozklad součtu čtverců
  $\sum\sum(Y_{it} - \bar{Y}_{\bullet\bullet})^2 =$
  $\sum n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum\sum(Y_{it} - \bar{Y}_{i\bullet})^2$
  (celková variabilita)=(mezi)+(uvnitř)

$$\begin{aligned} S_T &= S_A + S_e \\ f_T &= f_A + f_e \\ (n-1) &= (k-1) + (n-k) \end{aligned}$$

- H_0 zamítnout, je-li

$$F_A = \frac{S_A/f_A}{S_e/f_e} \geq F_{f_A,f_e}(\alpha)$$



- **model** (měření = úroveň + chyba)

$$\begin{aligned} Y_{it} &= \mu_i + E_{it} & 1 \leq t \leq n_i, \quad 1 \leq i \leq k \\ &= \mu + (\mu_i - \mu) + E_{it} & E_{it} \text{ nezávislé} \\ &= \mu + \alpha_i + E_{it} & E_{it} \sim \mathsf{N}\left(0, \sigma^2\right) \end{aligned}$$

- **reparametrizace** (α_i – efekty faktoru A):

$$\sum_{i=1}^{k} \alpha_i = 0$$

- **mnohonásobná srovnání**
  (které dvojice μ_i (resp. α_i) se liší?)

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| \geq q_{k,n-k}(\alpha)\sqrt{\frac{S^2}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$$S^2 = \frac{S_e}{f_e} = \frac{\sum\sum(Y_{it} - \bar{Y}_{i\bullet})^2}{n-k}$$

  (nutnost zachovat zvolenou hladinu testu)
- **ověření** shody rozptylů
  – Leveneův test
  – Bartlettův test (normalita!)

tabulka analýzy rozptylu

variabilita	S	f	S/f	F	p
výběry	S_A	$f_A = k-1$	S_A/f_A	F_A	p_A
reziduální	S_e	$f_e = n-k$	S_e/f_e		
celková	S_T	$f_T = n-1$			

příklad **játra**

variab.	S	f	S/f	F	p
místa	1,796	4	0,4490	5,862	0,0013
rezid.	2,285	30	0,0762		
celk.	4,081	34			

místo	počet	průměr	efekt	směr. odchylka
A	7	0,569	0,206	0,312
B	7	0,484	0,121	0,279
C	7	0,496	0,133	0,318
D	7	-0,063	-0,426	0,290
E	7	0,329	-0,034	0,144
celkem	35	0,363	0,000	0,104

$$q_{5,30}(0{,}05)\sqrt{\frac{0{,}0762}{2}\left(\frac{1}{7} + \frac{1}{7}\right)} = 4{,}10 \cdot 0{,}104 = 0{,}426$$

$-0{,}063 + 0{,}426 = 0{,}363 \Rightarrow$ na 5% hladině se liší místo D od každého z míst A, B, C

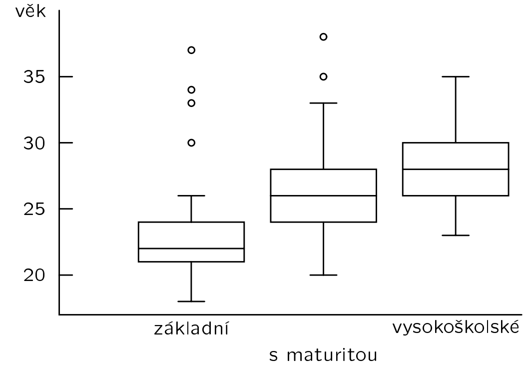
Kruskalův-Wallisův text

- zobecnění dvouvýběrového Wilcoxonova testu
  (pořadí místo původních hodnot)
- předpoklady:
  - k nezávislých výběrů
  - spojitá rozdělení
  - H_0: rozdělení jsou stejná
- T_i - součet pořadí v i-tém výběru

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$$

  H_0 se zamítá při $Q \geq \chi^2_{k-1}(\alpha)$
  (velká variabilita průměrných pořadí)

příklad **kojení** (věk matek podle vzdělání)



vzděl.	n_i	prům. věk	stř. chyba	souč. poř.	prům. poř.
zákl.	34	23,412	0,638	1025	30,15
mat.	47	26,278	0,543	2618	55,70
VŠ	18	28,50	0,877	1307	72,61
celk.	99	25,697		4 950	50

$$Q = \frac{12}{99 \cdot 100}\left(\frac{1025^2}{34} + \frac{2618^2}{47} + \frac{1307^2}{18}\right) - 3 \cdot 100 = 29{,}25$$

$\chi^2_2(0{,}05) = 5{,}99$ $\qquad p < 0{,}0001$

porovnání populačních měr polohy

rozdělení	normální	spojité
populační parametr (o čem je hypotéza)	populační průměr	populační medián (distribuční funkce)
jeden výběr	jednovýběrový t test	znaménkový Wilcoxon
výběr dvojic	párový t test	znaménkový Wilcoxon
dva nezávislé výběry	dvouvýběrový t test	Mann-Whitney (Kolmogorov-Smirnov)
k nezávislých výběrů	analýza rozptylu jednoduchého třídění	Kruskal-Wallis

vyšetřování závislosti

nezávisle proměnná(é)	závisle proměnná	
	spojitá	**nominální**
spojitá	regrese korelace	(*logistická regrese*)
nominální	analýza rozptylu	kontingenční tabulky

příklady:

- hmotnost na výšce
- rakovina plic na počtu vykouřených cigaret
- hmotnost obilky na živném roztoku
- barva očí a barva vlasů

Korelace a regrese

- **korelace**
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost v X, Y
- **regrese**
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** hodnotu Y pro zvolenou hodnoty x

korelační koeficient

- (populační) korelační koeficient ρ_{XY}
  - $|\rho_{XY}| \le 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy}
  - pro test nutno **normální** rozdělení

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

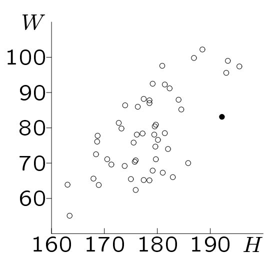
  - $\mathsf{H}_0 : \rho_{XY} = 0$ se na hladině α zamítá:

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}, \qquad |T| \ge t_{n-2}(\alpha)$$

- **Spearmanův** korelační koeficient
  - měří sílu **monotonní** závislosti
  - založen na **pořadích** R_i, Q_i hodnot X_i, Y_i

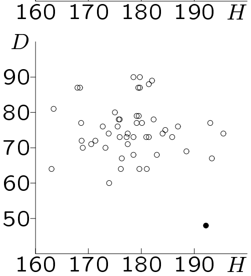
$$r_{XY}^{(S)} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$

příklad **tuk**



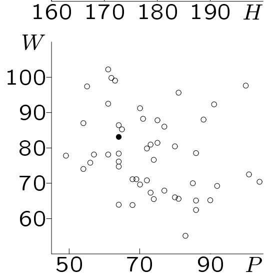
výška vers. hmotnost

$$\begin{aligned} r &= 0{,}643 && (0{,}654) \\ t &= 5{,}814 && (5{,}921) \\ p &< 0{,}0001 && (p < 0{,}0001) \end{aligned}$$



výška vers. diast. tlak

$$\begin{aligned} r &= -0{,}145 && (-0{,}018) \\ t &= -1{,}019 && (-0{,}124) \\ p &= 0{,}3135 && (0{,}9017) \end{aligned}$$



puls vers. hmotnost

$$\begin{aligned} r &= -0{,}245 && (-0{,}241) \\ t &= -1{,}752 && (-1{,}701) \\ p &= 0{,}0862 && (0{,}0955) \end{aligned}$$

Fisherova Z transformace

$$Z = \frac{1}{2}\log\frac{1+r}{1-r} \sim \mathsf{N}\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$$

- příklad **děti**: porodní délka, hmotnost
  - dívky: $r_1 = 0{,}5687$, $n_1 = 51$

$$z_1 = \frac{1}{2}\log\frac{1+0{,}5687}{1-0{,}5687} = 0{,}6456$$

  - hoši: $r_2 = 0{,}5967$, $n_2 = 49$, $z_2 = 0{,}6880$
  - test shody

$$z^* = \frac{0{,}6456 - 0{,}6880}{\sqrt{\frac{1}{51-3} + \frac{1}{49-3}}} = -0{,}2055.$$

  srovnej se $z(0{,}05/2) = 1{,}960$, $p = 0{,}8376$

  - 95% interval spolehlivosti pro ρ_1

$$\left(0{,}6456 - \frac{1{,}960}{\sqrt{51-3}}, \quad 0{,}6456 + \frac{1{,}960}{\sqrt{51-3}}\right)$$

$$(0{,}363, \quad 0{,}929)$$

$$\left(\frac{\mathsf{e}^{2\cdot 0{,}363} - 1}{\mathsf{e}^{2\cdot 0{,}363} + 1}, \quad \frac{\mathsf{e}^{2\cdot 0{,}929} - 1}{\mathsf{e}^{2\cdot 0{,}929} + 1}\right)$$

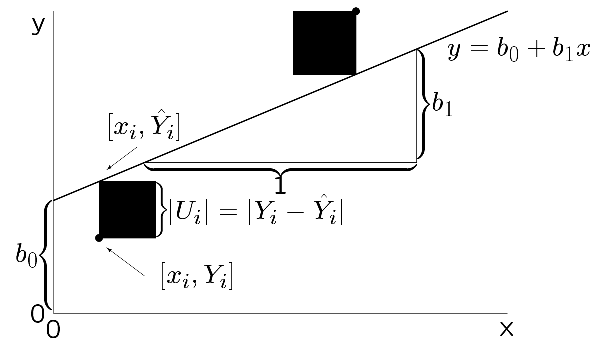
$$(0{,}348, \quad 0{,}730)$$

regrese (původ pojmu)

- tendence (návrat) k průměrnosti
  - F. Galton (1886): Family likeness in stature. Proc. Roy. Soc. XL, 42
  - F. Galton (1886): Regression towards mediocrity in hereditary stature. Journ. Anthrop. Inst. XV, 246
- uvažujme otce, jejichž výška je rovna průměrné výšce generace **všech** otců; průměrná výška synů těchto otců bude rovna průměrné výšce **všech** synů
- uvažujme otce o 10 cm **vyšší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen asi o 5 cm **vyšší**, než průměrná výška generace synů
- uvažujme otce o 10 cm **nižší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen o asi 5 cm **nižší**, než průměrná výška generace synů

regresní přímka

- odhadovaná závislost: $\mathsf{E}\,Y = \beta_0 + \beta_1 x$
- k daným $x_1, \ldots, x_n$ zjistíme $Y_1, \ldots, Y_n$
  - **nezávislá** pozorování
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (pro testy)
- b_0, b_1 – odhady metodou **nejmenších čtverců**: minimalizovat $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$



- odhad závislosti $\mathsf{E}\,Y = \beta_0 + \beta_1 x$

$$\hat{Y} = b_0 + b_1 x$$

- b_1 – odhad směrnice β_1, odhad změny střední hodnoty závisle proměnné Y při **jednotkové změně** nezávisle proměnné x
- reziduum $U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i)$
- reziduální součet čtverců:

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} U_i^2$$

- reziduální rozptyl

$$S^2 = \frac{S_e}{n-2}$$

- **koeficient determinace** (podíl variability Y vysvětlené uvažovanou závislostí)

$$R^2 = 1 - \frac{S_e}{\sum (Y_i - \bar{Y})^2}$$

- nezávislost $\mathsf{E}\,Y$ na x znamená $\mathrm{H}_0 : \beta_1 = 0$

$$T = \frac{b_1}{\mathrm{S.E.}(b_1)} \qquad |T| \geq t_{n-2}(\alpha)$$

příklad závislost procenta tuku FAT na výšce HEIGHT u mladých mužů

regresor	b_j	S.E.(b_j)	t	p
abs. člen	–53,870	24,657	–2,185	0,0338
HEIGHT	0,379	0,138	2,742	0,0086

předpověď: $\hat{Y}_i = -53{,}870 + 0{,}379 x_i$,
tedy $\widehat{\text{FAT}}$= –53,870 + 0,379·HEIGHT
(na každý centimetr výšky *v průměru* 0,379 procentního hodu)

variabilita	součet čtverců	st. vol.	prům. čtverec	F	p
regrese	362,54	1	362,54	7,519	0,0086
rezid.	2314,41	48	48,22		
celk.	2676,95	49	(54,63)		

$$R^2 = \frac{362{,}54}{2676{,}95} = 1 - \frac{2314{,}41}{2676{,}95} = 0{,}135$$

mnohonásobná lineární regrese

- závislost na dvou nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \dots, (x_n, v_n, Y_n)$
- $Y_1, \dots, Y_n$ jsou **nezávislé** náhodné veličiny
- stejný rozptyl σ^2
- normální rozdělení Y_i pro dané x_i, v_i
- střední hodnoty Y_i vysvětleny pomocí x_i, v_i

$$\mathsf{E}\, Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i$$

- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$
- b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- U_i – **reziduum**

$$\begin{aligned} U_i &= Y_i - \hat{Y}_i \\ &= Y_i - (b_0 + b_1 x_i + b_2 v_i) \end{aligned}$$

- **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = S_R + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- **koeficient determinace** R^2
(podíl celkové variability, který se podařilo vysvětlit závislostí Y na x, v)

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$$

uvažujeme závislost $\boxed{\mathsf{E}\, Y = \beta_0 + \beta_1 x + \beta_2 v}$

- $\mathsf{H}_0 : \beta_2 = 0$ (k vysvětlení Y stačí x)

$T_2 = \dfrac{b_2}{\mathsf{S.E.}(b_2)}$, zamítat pro $|T_2| \geq t_{n-3}(\alpha)$

- $\mathsf{H}_0 : \beta_1 = 0$ (k vysvětlení Y stačí v)

$T_1 = \dfrac{b_1}{\mathsf{S.E.}(b_1)}$, zamítat pro $|T_1| \geq t_{n-3}(\alpha)$

- $\mathsf{H}_0 : \beta_1 = \beta_2 = 0$ (nezáv. ani na x ani na v)

$$F = \frac{S_R/2}{S_e/(n-3)} \geq F_{2,n-3}(\alpha)$$

příklad závislost FAT na HEIGHT a WEIGHT

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
HEIGHT	–0,262	0,110	–2,376	0,0216
WEIGHT	0,624	0,0690	9,050	<0,0001

– při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku

– u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku

variabilita	součet čtverců	st. vol.	prům. čtverec	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

$$R^2 = \frac{1833{,}11}{2676{,}95} = 1 - \frac{843{,}85}{2676{,}95} = 0{,}685$$

χ^2 testy

- pro znaky v **nominálním** měřítku
- **příklady**
  - krevní skupiny A, B, AB, 0 u n osob
  - počty dětí narozených v jednotlivých měsících
  - počty matek se základním, středním, vysokoškolským vzděláním
- **multinomické** rozdělení
  - v dílčím pokusu k možných výsledků $A_1, \dots, A_k$ (neslučitelné)
  - π_j je pst, že vyjde A_j $(\sum \pi_j = 1)$
  - n **nezávislých** dílčích pokusů
  - N_j – počet dílčích pokusů, kdy A_j
  - $(N_1, \dots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \dots, \pi_k$
  - samotné N_j má binomické rozdělení
- **pravděpodobnost** $N_1 = n_1, \dots, N_k = n_k$

$$\frac{n!}{n_1! \dots n_k!} \pi_1^{n_1} \dots \pi_k^{n_k}$$

- hlavní vlastnost (pokud $n\pi_j \geq 5$ pro $\forall\ j$)

$$\boxed{X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j}}$$

má přibližně rozdělení $\chi^2(k-1)$

- **test shody** $\mathsf{H}_0 : \pi_1 = \pi_1^0, \dots, \pi_k = \pi_k^0$ (pravděpodobnosti dány **jednoznačně**)
  - platí-li H_0, očekáváme četnosti blízké hodnotám $n\pi_j^0$:

$$\boxed{X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}}$$

  - H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(\alpha)$
  - N_j experimentální četnost
  - $n\pi_j^0$ **teoretická** četnost
  - statistika X^2 porovnává experimentální a teoretické četnosti

příklad **měsíce**
počty studentů biologie narozených v jednotlivých měsících
hypotéza:
dětí se rodí během roku **rovnoměrně**

měsíc	n_j	$n\pi_j^0$	přínos
1	11	9,43	0,2623
2	9	8,52	0,0276
3	13	9,43	1,3539
4	11	9,12	0,3861
5	8	9,43	0,2161
6	5	9,12	1,8635
7	10	9,43	0,0348
8	6	9,43	1,2461
9	13	9,12	1,6473
10	8	9,43	0,2161
11	8	9,12	0,1383
12	9	9,43	0,0194
celkem	111	111,00	7,4115

$X^2 = 7{,}4115 < \chi^2_{12-1}(0{,}05) = 19{,}675 \qquad p = 0{,}765$

kontingenční tabulka

- nominální znak s hodnotami $A_1, \dots, A_r$
- nominální znak s hodnotami $B_1, \dots, B_c$
- N_{ij} kolikrát současně A_i a B_j
- **marginální** četnosti

$$N_{i\bullet} = \sum_{j=1}^{c} N_{ij} \qquad N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$$

- **nezávislost** znaků: pro všechna i, j

$$\mathsf{P}(A_i \cap B_j) = \mathsf{P}(A_i)\mathsf{P}(B_j)$$

- teoretické četnosti (protějšek N_{ij})

$$o_{ij} = n \cdot \widehat{\mathsf{P}(A_i)} \cdot \widehat{\mathsf{P}(B_j)} = n \cdot \frac{N_{i\bullet}}{n} \cdot \frac{N_{\bullet j}}{n} = \frac{N_{i\bullet} N_{\bullet j}}{n}$$

- H_0 : znaky jsou **nezávislé**

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - o_{ij})^2}{o_{ij}}$$

- nezávislost se zamítá pro $X^2 \geq \chi^2_{(r-1)(c-1)}(\alpha)$
- musí být $o_{ij} \geq 5\ \forall\ (i, j)$

příklad **Baden**

barva očí	barva vlasů				celkem
	světlá	hnědá	černá	ryšavá	
modrá	1 768	807	189	47	2 811
šedá/zelená	946	1 387	746	53	3 132
hnědá	115	438	288	16	857
celkem	2 829	2 632	1 223	116	6 800

- barva očí $r = 3$
- barva vlasů $c = 4$
- $n = 6800$
- $o_{11} = 2811 \cdot 2829/6800 = 1169$
- $o_{12} = 2811 \cdot 2632/6800 = 1088$
- $o_{13} = \dots$

$$\begin{aligned} \chi^2 &= \frac{(1768-1169)^2}{1169} + \frac{(807-1088)^2}{1088} + \dots \\ &= 1073{,}5 \\ &> \chi^2_6(0{,}05) = 12{,}5916 \\ p &< 0{,}0001 \end{aligned}$$

závislost je na každé rozumné hladině **prokázána**

- test **homogenity**
  - hodnoty znaku $B_1, \ldots, B_c$
  - r **nezávislých** výběrů z různých populací
  - H_0 : populace se **neliší**
  - dál stejně jako pro nezávislost
- příklad **krevní skupiny**

populace	skupina				celkem
	0	A	B	AB	
C	121	120	79	33	353
D	118	95	121	30	364
celkem	239	215	200	63	717

$$\chi^2 = \frac{(121 - 353 \cdot 239/717)^2}{353 \cdot 239/717} + \ldots = 11{,}742$$

- $\chi^2_3(0{,}05) = 7{,}815$ $\qquad p = 0{,}008$
- nejmenší teoretická četnost: $353 \cdot 63/717 = 31{,}02 > 5$

McNemarův test (test symetrie)

- **párový** test pro nominální veličinu s hodnotami $B_1, \ldots, B_k$
- zjišťujeme hodnoty nominálního znaku na **stejných** objektech za **dvojích** okolností (před ošetřením, po ošetření)
- N_{ij} počet objektů, u nichž před ošetřením B_i a po ošetření B_j
- **hypotéza**: pravděpodobnosti možných hodnot znaku jsou **stejné** za obojích okolností (před ošetřením i po něm)

$$X^2 = \sum \sum_{i<j} \frac{(N_{ij} - N_{ji})^2}{N_{ij} + N_{ji}}$$

- hypotézu zamítneme při $X^2 \geq \chi^2_{k(k-1)/2}(\alpha)$
- výrazy ve jmenovateli kladné!
- nezávisí na počtu objektů, kdy vyšly oba výsledky stejně

příklad **stromy**

1994	1995			celkem
	1	2	3	
1	4	3	3	10
2	7	21	11	39
3	1	15	35	51
celkem	12	39	49	100

- stav týchž stromů ve dvou sezónách
- celkem 100 stromů

$$\chi^2 = \frac{(3-7)^2}{3+7} + \frac{(3-1)^2}{3+1} + \frac{(11-15)^2}{11+15} = 3{,}215$$

- $\chi^2_3(0{,}05) = 7{,}8147$ $\qquad p = 0{,}3597$
- rozdíl mezi sezónami jsme neprokázali

čtyřpolní tabulka

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

- speciální případ kontingenční tabulky pro $r = c = 2$
- test nezávislosti/homogenity

$$X^2 = \frac{n(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

zamítá se pro $X^2 \geq \chi^2_1(\alpha) = z(\alpha/2)^2$

- **Yatesova korekce**

$$X^2_Y = \frac{n(|ad - bc| - n/2)^2}{(a+c)(b+d)(a+b)(c+d)}$$

- **Fisherův faktoriálový (exaktní) test**
  - počítá přímo dosaženou hladinu p
  - malé četnosti nevadí

příklad **hraboš**

Frenkelia spp.	*Sarcocystis spp.* +	*Sarcocystis spp.* −	celkem
+	4	27	31
-	11	473	484
celkem	15	500	515

- souvisí spolu nákazy dvěma cizopasníky?
- nulová hypotéza: **nezávislost**

$$\chi^2 = \frac{515(4 \cdot 473 - 11 \cdot 27)^2}{15 \cdot 500 \cdot 31 \cdot 484} = 11{,}643$$

$$p = 0{,}0006$$

- **ale**: $15 \cdot 31/515 = 0{,}9 < 5$
- **Yates**: $\chi^2 = 8{,}187$ $p = 0{,}0042$
- **Fisherův test**: $p = 0{,}0092$
- na 5% hladině závislost **prokázána**

jak použijeme statistiku

- co o problému zjistili jiní? (přečti, sepiš)
- co chceš zjistit?
  - zformuluj otázku (to určí možné statistické metody)
  - zformuluj nulovou a alternativní hypotézu
- zvol hladinu testu α
- pořiď data
  - proveď měření (podrobné záznamy!)
  - převeď do elektronické formy (kódování)
  - vyčisti data (grafy, popisné statistiky,. . .)
- proveď výpočty, kresli grafy
- použij výsledky a grafy, interpretuj

dvojí původ dat

- **plánovaný** (organizovaný) **pokus**
  - aktivně zasahujeme
  - fixujeme okolnosti (stálá teplota, světelný režim)
  - nastavujeme úrovně zvoleného faktoru (dva živné roztoky)
  - jedincům náhodně přiřazujeme ošetření
  - zjistíme-li rozdíl, známe jeho příčinu
- **šetření** (sledování dění)
  - pouze sledujeme, nezasahujeme
  - rozdělení do skupin nemůžeme ovlivnit
  - rozdíl mezi skupinami může být způsoben matoucí (confounding) veličinou, která souvisí s rozdělením do skupin i s měřeným znakem
  - příklad: plánované těhotenství na vzdělání matky, matoucí je věk matky

jaké úlohy řešíme

- **popsat stav**
  - poloha (průměr, medián, kvartily,. . .)
  - variabilita (směr. odchylka, rozptyl, kvartilové rozpětí)
  - závislost (korelační koeficient, Spearmanův korelační koeficient)
  - tvar rozdělení (šikmost, špičatost)
- **prokázat vliv ošetření**
  - změna polohy (t testy, ANOVA)
  - změna variability (Levene, F test, Bartlettův test)
  - jiná změna (Kolmogorov-Smirnov)
- **prokázat závislost**
  - obě spojité (korelační koeficient)
  - spojitá na kvalitativními (ANOVA)
  - obě kvalitativní (kontingenční tabulka)
- **popsat závislost** spojitých – regrese

výběr metody

- jakou úlohu řešíme?
- jsou výběry nezávislé?
  - z organizace pokusu
- lze předpokládat normální rozdělení?
  - ze zkušenosti
  - lze ověřovat (ve skupinách pozorování, z reziduí)
  - lze soudit z grafu (normální diagram)
- je rozptyl stálý?
  - lze ověřovat (ve skupinách pozorování, z reziduí)
  - lze soudit z grafu (normální diagram)

volba nulové a alternativní hypotézy

- H_0 zjednodušuje model
  - populace se neliší (výběry se liší jen náhodně)
  - veličiny jsou nezávislé
  - H_0 zpravidla chceme vyvrátit
- H_1 je opak nulové hypotézy
  - zpravidla obsahuje tvrzení, které chceme dokázat
  - pokud existuje jednostranná alternativní hypotéza, musíme ji zvolit **před pokusem** na základě úvah, které **nejsou** založeny na použitých datech
- pouze zamítnutím H_0 něco dokazujeme