Základy biostatistiky

(S710P09)

akademický rok 2004/2005

Karel Zvára

Karel.Zvara@mff.cuni.cz

http://www.karlin.mff.cuni.cz/~zvara

poslední úprava 11. května 2005

- **cvičení na počítačích**
  - od čtvrtka 24. února ve Viničné 7, 1. patro B5
  - v paralelce 16–18 studentů
  - nutno se přihlásit, přednost mají studenti bez zápočtu
  - zápočet za aktivní účast (včetně odevzdání souborů podle požadavků cvičícího)
  - nutno mít aktivní účet na síti, znát svoje heslo
  - NCSS (Number Cruncher Statistical System)
- **zkouška v B5**
  - jen se zápočtem, přihlašování přes SIS
  - kombinace písemné a ústní
- **literatura**
  - K. Zvára: Biostatistika. Karolinum
- **konzultace** pátek 10:00–11:00, 2. patro, Sokolovská 83, Karlín

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)

co měříme (zjišťujeme) a kde

- měříme na mnoha statistických **jednotkách** (osoba, obec, stát, pokusné pole . . .)
- měříme (zjišťujeme) hodnoty **znaků**
- **znak** - vlastnost měřená na objektu (statistické jednotce)
- zjištěnou hodnotu vyjadřujeme ve zvoleném **měřítku** (stupnici)
- na jedné jednotce můžeme měřit několik znaků (možná závislost)
- měříme na skupinách jednotek – **souborech**
- zajímají nás **hromadné** vlastnosti (les, ne jednotlivé stromy)
- můžeme porovnávat vlastnosti znaku mezi soubory

měřítka

- **nula-jedničkové** (muž/žena) pouze dvě možné hodnoty
- **nominální** (porodnice, pohlaví, odrůda) seznam všech rozlišitelných hodnot, **faktor**
- **ordinální** (vzdělání matky,. . . , stupeň bolesti) hodnoty nominálního měřítka uspořádány, **uspořádaný faktor**
- **intervalové** (rok narození, teplota ve $^{\circ}$C) stejné vzdálenosti sousedních hodnot, „o kolik je menší?“
- **poměrové** (hmotnost, výška, věk) srovnání se zvolenou jednotkou, „kolikrát je větší?“

měřítka (2)

- **kvalitativní**: nula-jedničkové, nominální, často i ordinální
- u kvalitativních se zpravidla udávají **četnosti** jednotlivých hodnot (kolikrát která nastala)
- **kvantitativní** (spojité): intervalové, poměrové, někdy ordinální (není spojité)
- hodnoty kvantitativních – čísla

veličina

- číselně vyjádřený výsledek měření, pokusu
- hodnoty znaků v intervalovém, poměrovém měřítku jsou husté – **spojitá veličina**
- četnosti hodnot znaků v nula-jedničkovém, nominálním (či ordinálním) měřítku – **diskrétní veličina**
- u veličin používáme charakteristiky některých hromadných vlastností (**charakteristiky polohy, variability**)
- **statistika** (též) funkce pozorovaných hodnot

popisné statistiky

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

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

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

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

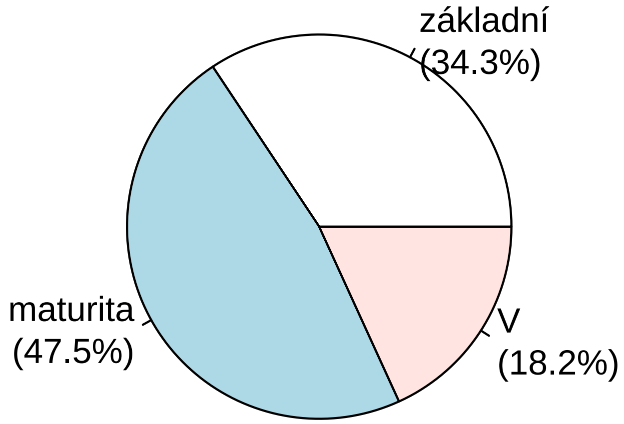
pro kumulativní četnosti nutno aspoň ordinální měřítko

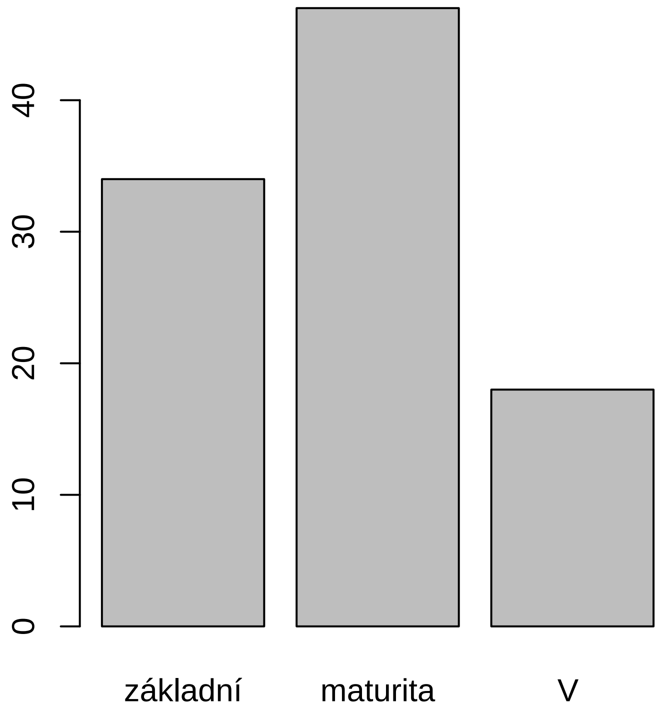
histogram (histogram, barplot)

- grafické znázornění četností (počtů hodnot kvalitativní veličiny)
- plocha (výška) obdélníku úměrná četnosti
- relativní četnosti dají jen jiné měřítko
- podobně **výsečový diagram** pro relativní četnosti (podíly nějakého celku)

příklad **kojení** (vzdělání 99 matek):

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


základní
(34.3%)
maturita
(47.5%)
V
(18.2%)


0
10
20
30
40
základní
maturita
V

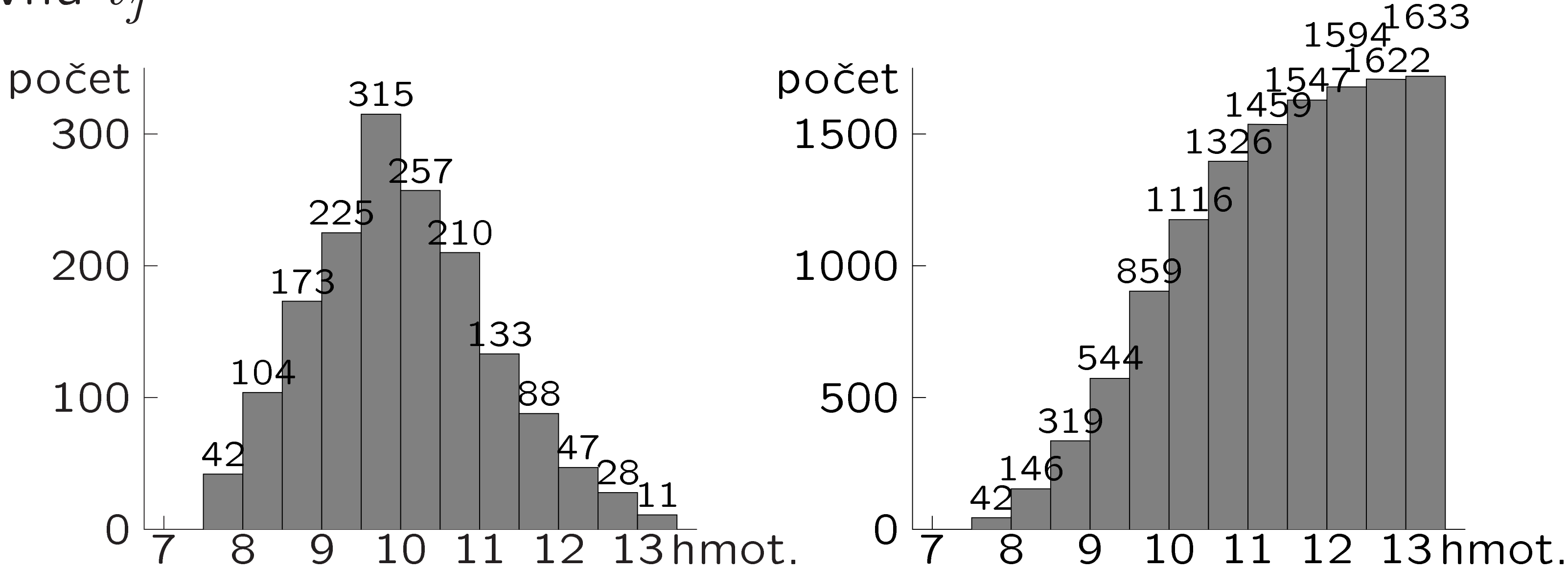
histogram u **spojité** veličiny – **třídění**: všechny hodnoty z daného intervalu $(t_{j-1}, t_j\rangle$ nahradíme prostřední hodnotou $x_j^* = (t_{j-1} + t_j)/2$

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

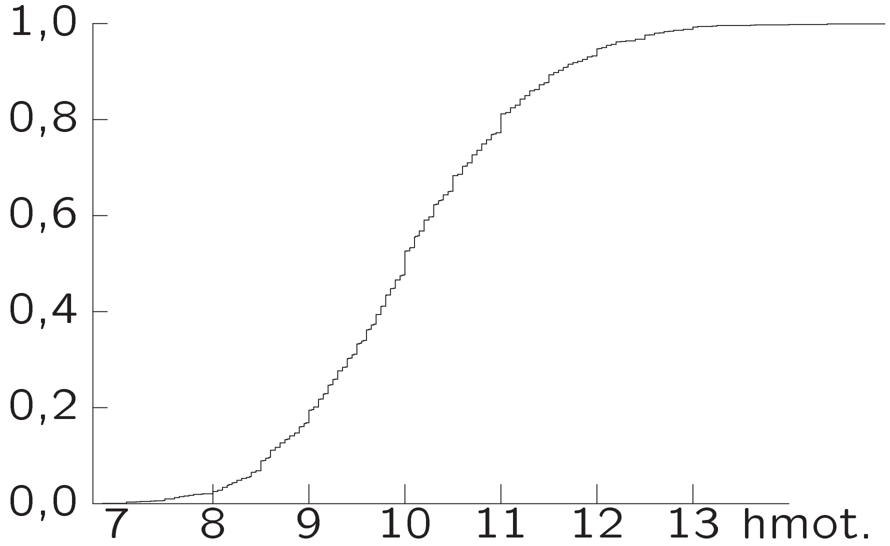
absolutní a kumulativní četnosti

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



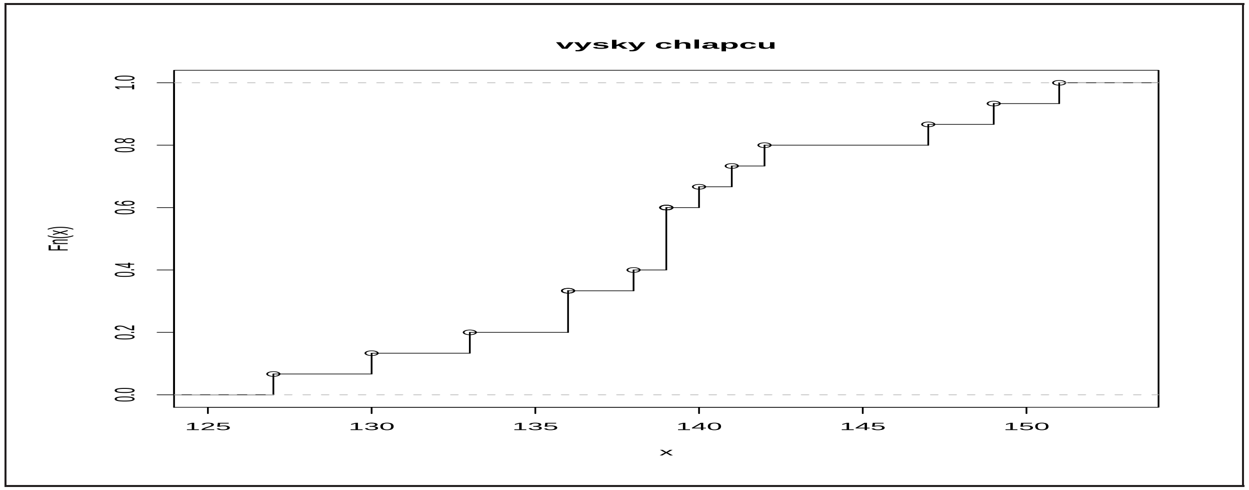
empirická distribuční funkce: relativní četnost hodnot, které jsou nejvýše x

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



x − výšky desetiletých hochů

i	1	2	3	4	5	6	7	8
x_i	130	140	136	141	139	133	149	151
i	9	10	11	12	13	14	15	
x_i	139	136	138	142	127	139	147	



uspořádaný seznam hodnot, variační řada

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

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

míry polohy **požadavky na míry polohy**:

$$\mu(a + X) = a + \mu(X)$$
$$\mu(b \cdot X) = b \cdot \mu(X) \qquad b > 0$$

– přičteme-li ke každé hodnotě konstantu a, přičte se stejná konstanta k míře polohy
– vynásobíme-li každou hodnotu kladnou konstantou b, vynásobí se míra polohy stejnou konstantou

- **průměr**

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

- **vážený průměr** s využitím četností ($n = \sum_j n_j$)

$$\boxed{\bar{x} = \frac{1}{n}\left(n_1 x_1^* + n_2 x_2^* + \ldots + n_m x_m^*\right) = \frac{1}{n}\sum_{j=1}^{m} n_j x_j^*}$$

- obecněji s nezápornými vahami w_j hodnot x_j^*

$$\boxed{\bar{x} = \frac{\sum_j w_j x_j^*}{\sum_j w_j}}$$

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

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

- **minimum, maximum**

$$\begin{aligned} x_{\text{min}} &= x_{(1)} \\ x_{\text{max}} &= x_{(n)} \end{aligned}$$

- **variační průměr**

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

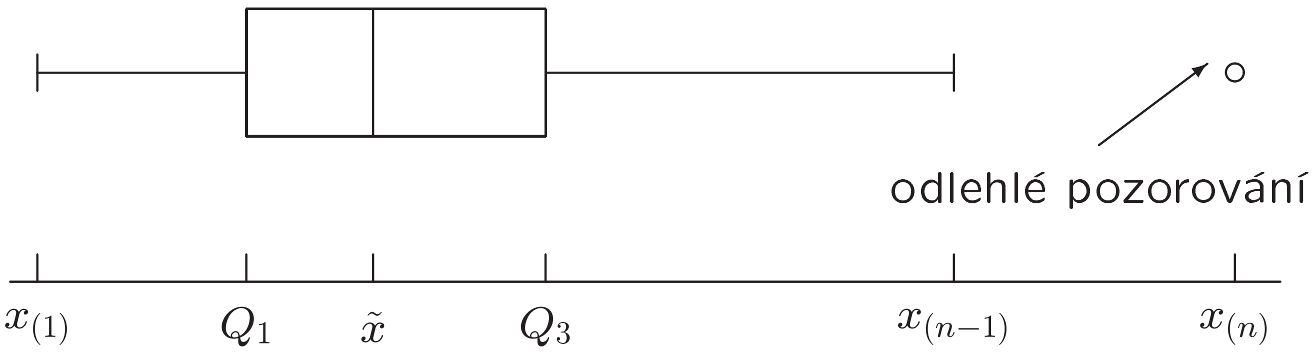
- p**-tý percentil** (dolních $100p$ % hodnot)
  (řada postupů, zde interpolace)

$$r = [(n+1)p] \qquad \text{celá část } (n+1)p$$
$$q = (n+1)p - r \qquad \text{zlomková část } (n+1)p$$

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

- **dolní kvartil** (hodnota oddělující dolní čtvrtinu): $\boxed{Q_1 = x_{1/4}}$
- **horní kvartil** (hodnota oddělující dolní tři čtvrtiny): $\boxed{Q_3 = x_{3/4}}$
- **modus** $\hat{x}$ mejčastější hodnota (s největší četností)

krabicový diagram (box-plot, box and whisker plot)



pozorování je **odlehlé**, je-li vzdáleno od bližšího kvartilu o více než

$$\boxed{\frac{3}{2}(Q_3 - Q_1)}$$

příklad **výšky dívek** (n=12)

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

$$\bar{y} = \frac{1}{12}(131 + 132 + \ldots + 151) = 140{,}83$$

$$\tilde{y} = \frac{1}{2}\Big(y_{(6)} + y_{(7)}\Big) = \frac{1}{2}(141 + 141) = 141$$

$$r = [(12+1)/4] = 3 \qquad q = (12+1)/4 - 3 = 1/4$$

$$Q_1 = \frac{3}{4}y_{(3)} + \frac{1}{4}y_{(4)} = 0{,}75 \cdot 135 + 0{,}25 \cdot 141 = 136{,}5$$

$$Q_3 = 0{,}25 \cdot 143 + 0{,}75 \cdot 146 = 145{,}25$$

míry variability

požadavky na míry variability (měřítka):

$$\sigma(a + X) = \sigma(X)$$
$$\sigma(b \cdot X) = b \cdot \sigma(X) \qquad\qquad b > 0$$

– přičtení konstanty míru variability neovlivní
– vynásobení všech hodnot kladnou konstantou b znamená vynásobení míry stejnou konstantou

- **směrodatná odchylka**

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

- **rozptyl** s_x^2 (nesplňuje druhý požadavek, $s_{b\cdot x}^2 = b^2 s_x^2$)
- **rozpětí** $R = x_{\mathsf{max}} - x_{\mathsf{min}}$
- **kvartilové rozpětí** $\boxed{R_Q = Q_3 - Q_1}$
- **variační koeficient** (nesplňuje ani jeden požadavek)
  porovnání variability při různých úrovních
$$V_x = \frac{s_x}{\bar{x}}$$
- **entropie** (pro nominální, požadavky nemají smysl)
$$H = -\sum_{j=1}^{m} \frac{n_j}{n} \ln \frac{n_j}{n}$$
  (nezávisí na označení hodnot)

příklad **výšky dívek**

$$
\begin{aligned}
s_y^2 &= \frac{1}{11}\Big((131-140{,}83)^2+\ldots+(151-140{,}83)^2\Big)\\
&\doteq 33{,}788\\
s_y &= \sqrt{33{,}788}\doteq 5{,}813\\
R &= 151-131=20\\
R_Q &= 145{,}25-136{,}5=8{,}75
\end{aligned}
$$

příklad **vztah mužů ke kouření** (základní vzdělání):

$$H=-\left(\frac{14}{117}\ln\frac{14}{117}+\ldots+\frac{78}{117}\ln\frac{78}{117}\right)=1{,}000689$$

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

z-skór (normovaná veličina)

$$\boxed{z_i = \frac{x_i - \bar{x}}{s_x}} \Rightarrow \bar{z} = 0, \qquad s_z = 1$$

hodnocení vlastností na poloze a variabilitě nezávislých

- **šikmost** (průměr 3. mocnin z-skórů)

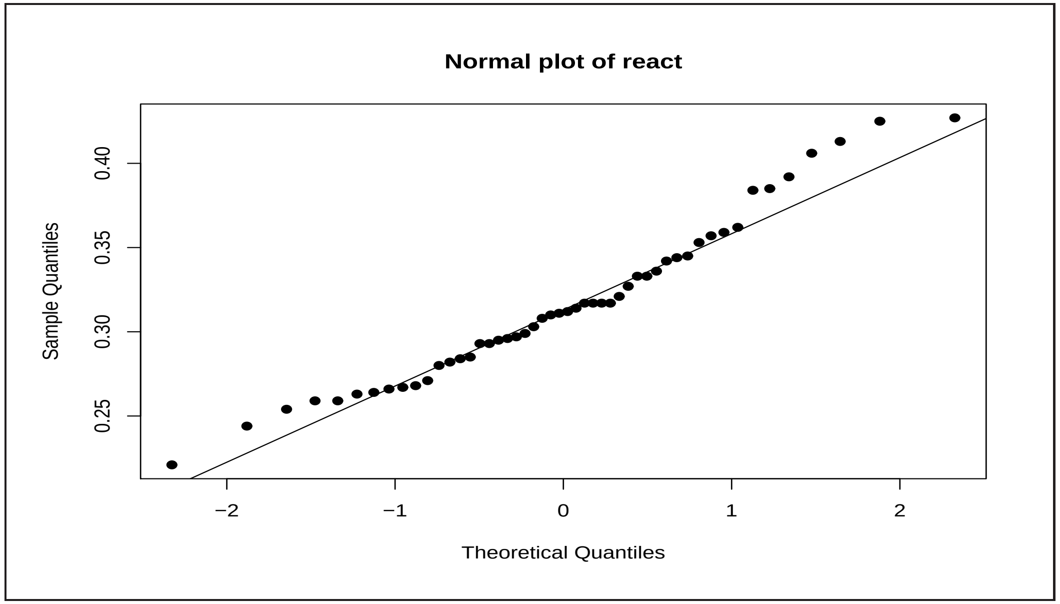
$$\boxed{g_1 = \frac{1}{n}\sum_{i=1}^{n} z_i^3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^3}$$

- **špičatost** (průměr 4. mocnin z-skórů, někdy bez -3)

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

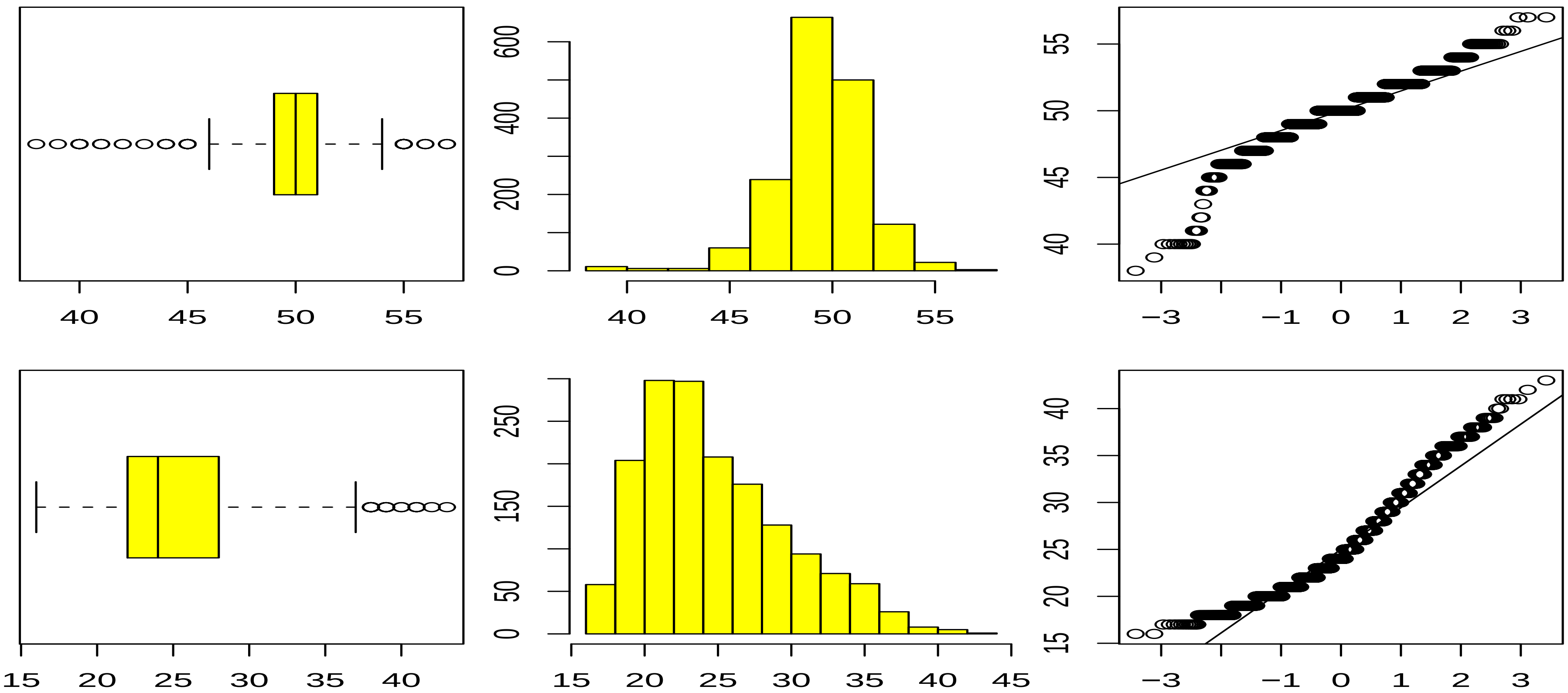
g_1, g_2 se používají k posouzení normality

- **normální diagram**
  - k ověřování předpokladu **normálního** rozdělení
  - pomocí srovnání bodů s přímkou
  - reakční doba: $g_1 = 0{,}521, \qquad g_2 = -0{,}321$



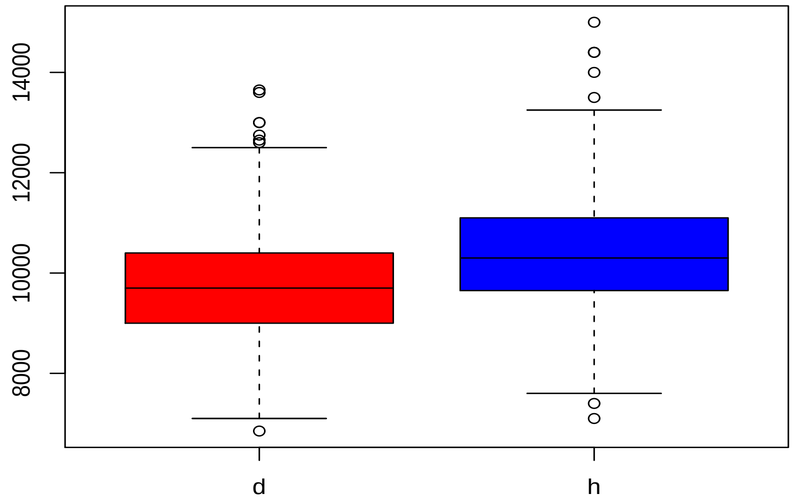
nahoře **porodní délka**: $g_1 = -0{,}893, g_2 = 3,511$

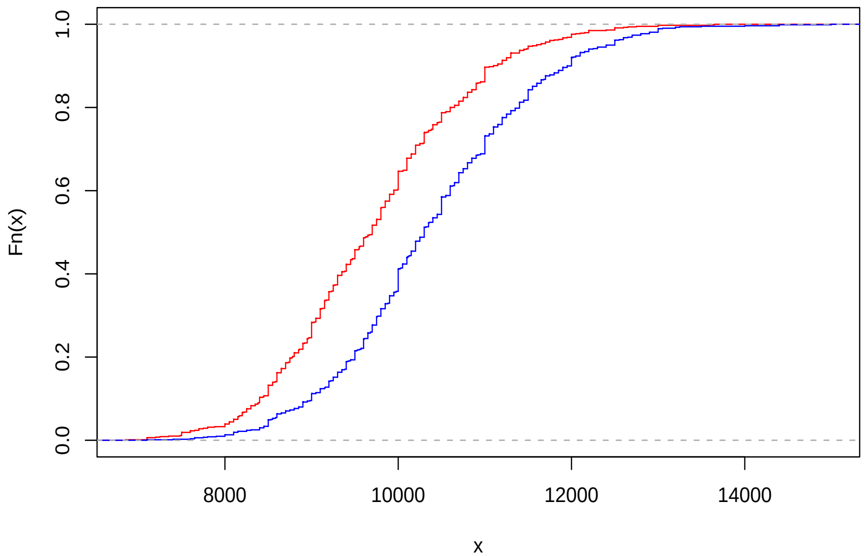
dole **věk matek**: $g_1 = 0{,}760, g_2 = 0{,}013$



srovnání souborů dat (spojitá veličina)

- lze chápat jako závislost spojité veličiny na nula-jedničkové
- krabicové diagramy resp. empirické distribuční funkce
- příklad: hmotnost chlapců a dívek (vlevo) v jednom roce)





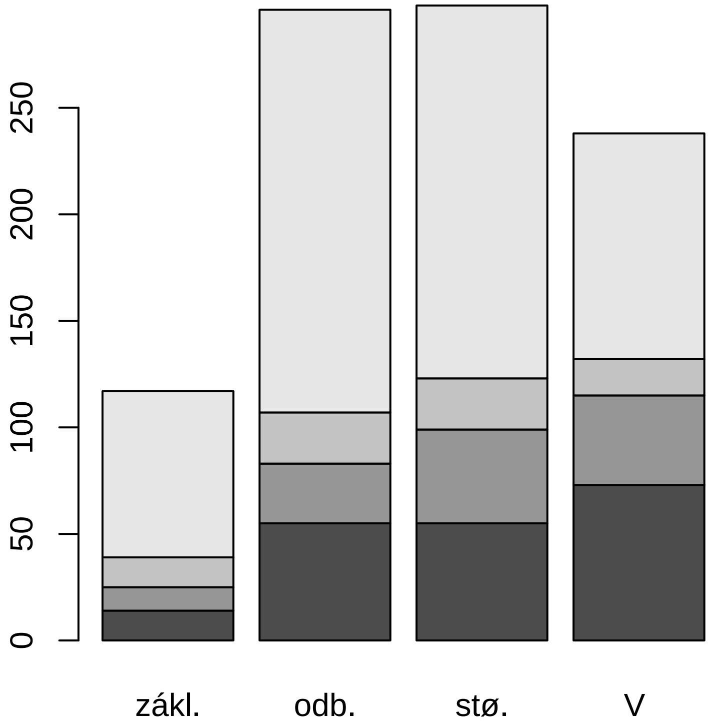
závislost dvojice znaků

- **dvojice znaků** – (není totéž co dva znaky!)
  - možnost porovnání hodnot jednoho pří různých hodnotách druhého
  - zkoumání závislosti
- **kontingenční tabulka**
  - četnosti všech kombinací obou kvalitativních znaků
  - v procentech v dané skupině (pro danou hodnotu jednoho znaku)

příklad **kouření u mužů**

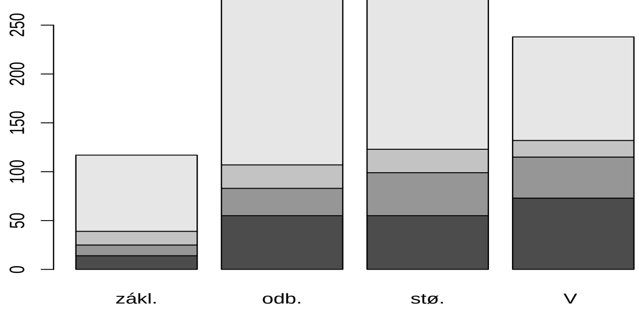
vzdělání	zákl.	odb.	mat.	VŠ	celk.
nekuřák	14	55	55	73	197
bývalý kuřák	11	28	44	42	125
kuřák	14	24	24	17	79
silný kuřák	78	189	175	106	548
celkem	117	296	298	238	949

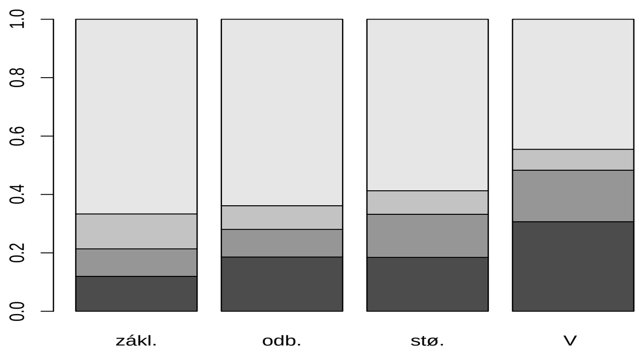
(zdola: nekuřák, bývalý kuřák, kuřák, silný kuřák)



procenta vztahu ke kouření podle vzdělání

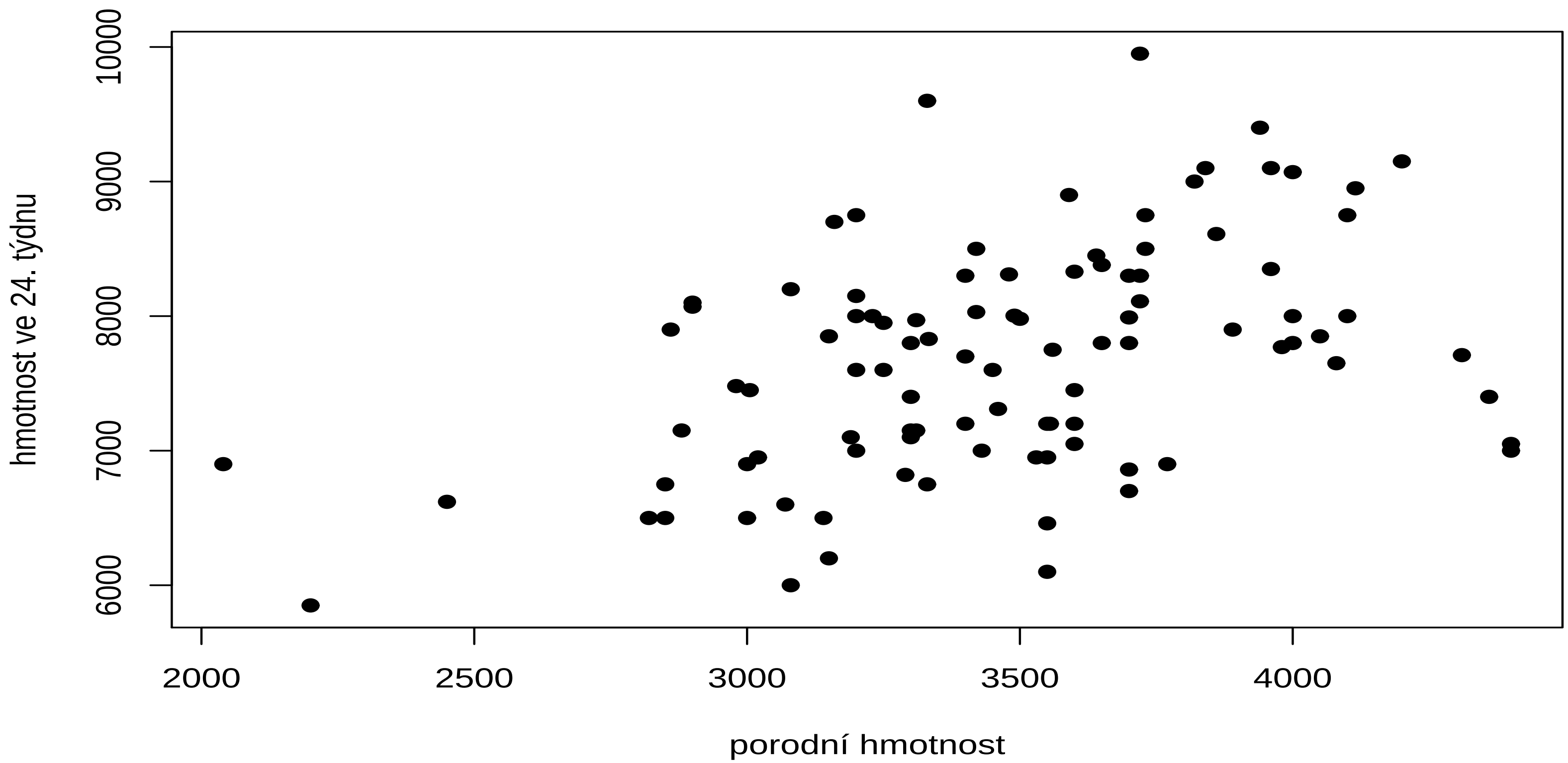
vzdělání	zákl.	odb.	mat.	VŠ	celk.
nekuřák	12,0%	18,6%	18,5%	30,7%	20,6%
bývalý kuřák	9,4%	9,5%	14,8%	17,6%	13,2%
kuřák	12,0%	8,1%	8,1%	7,1%	8,3%
silný kuřák	66,7%	63,9%	58,7%	44,5%	57,8%
celkem	100%	100%	100%	100%	100%





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

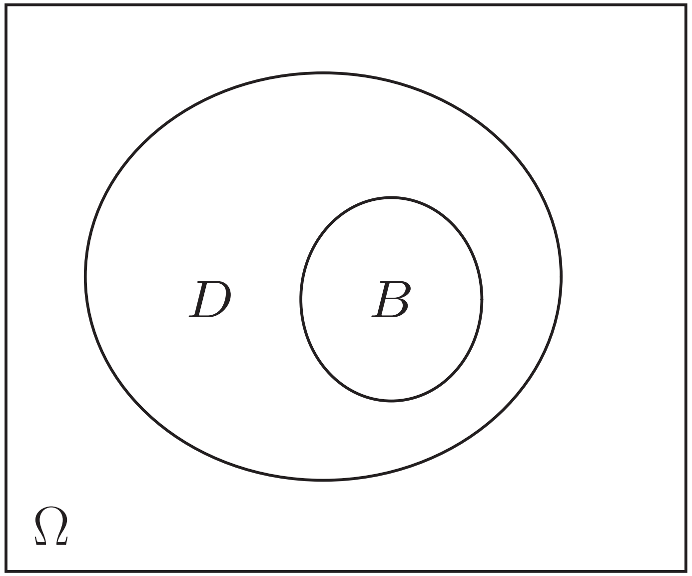
- **závislost spojitých veličin** (bodový diagram)



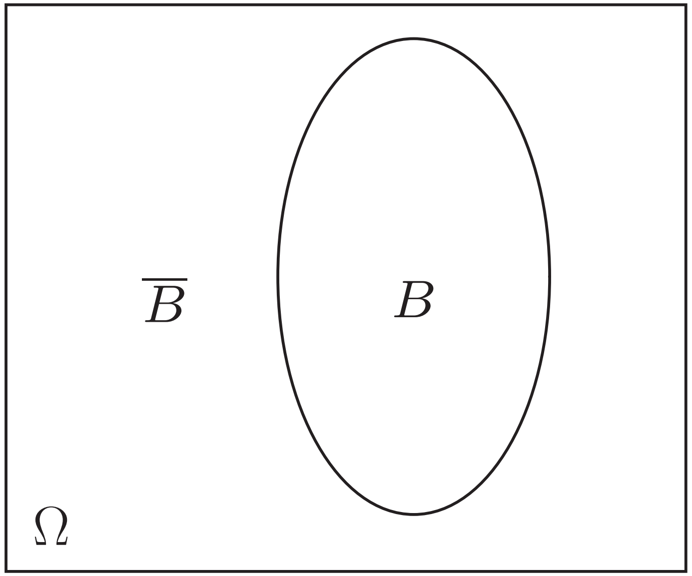
Náhodné jevy

- **náhodný pokus** výsledek nejistý, s opakováním roste 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** $B \cap D = \emptyset$

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

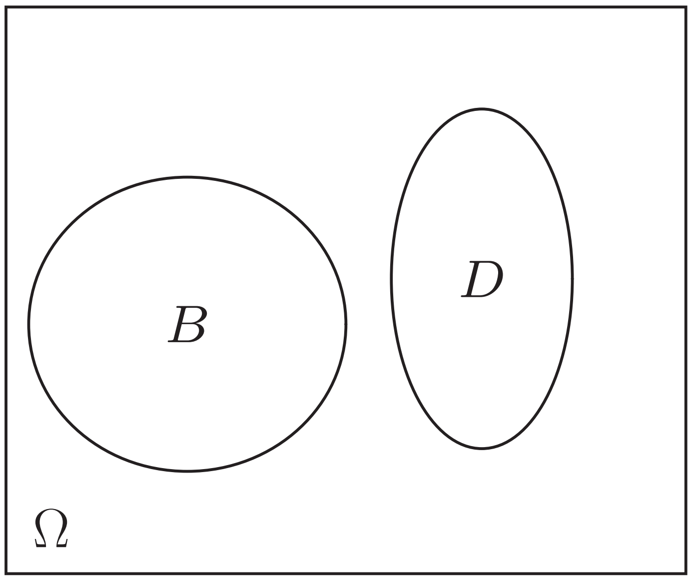


$\mathsf{P}(\overline{B}) = 1 - \mathsf{P}(B)$

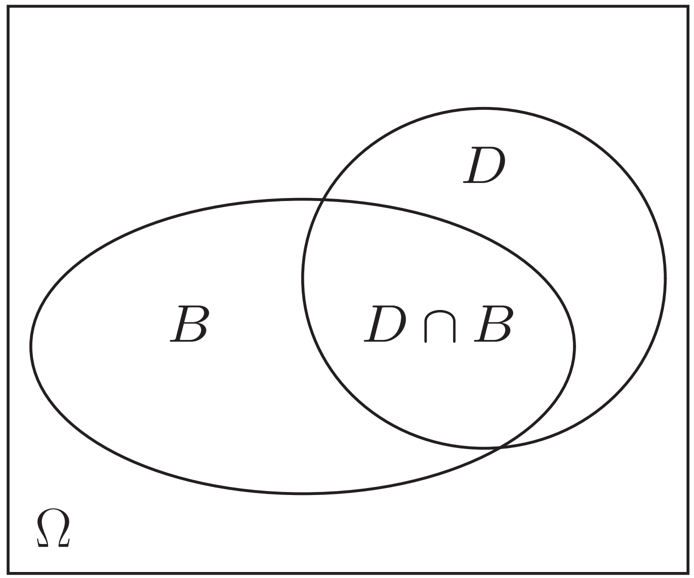


velikost plochy odpovídá pravděpodbnosti

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



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



velikost plochy odpovídá pravděpodbnosti

Pravděpodobnost $\mathsf{P}(B)$

- objektivní číselné vyjádření „naděje“, že nastane B
- modelový protějšek relativní četnosti
- vlastnosti psti
  - $\boxed{0 \leq \mathsf{P}(B) \leq 1}$
  - $\boxed{\mathsf{P}(\Omega) = 1, \mathsf{P}(\emptyset) = 0}$
  - $\boxed{B \cap D = \emptyset \Rightarrow \mathsf{P}(B \cup D) = \mathsf{P}(B) + \mathsf{P}(D)}$
  - $\boxed{\mathsf{P}(B \cup D) = \mathsf{P}(B) + \mathsf{P}(D) - \mathsf{P}(B \cap D)}$
  - $\boxed{B \subset D \Rightarrow \mathsf{P}(B) \leq \mathsf{P}(D)}$
  - $\boxed{\mathsf{P}(\overline{B}) = 1 - \mathsf{P}(B)}$

- **klasická definice psti**
  - m stejně pravděpodobných elementárních jevů ω_i
  - m_B elementárních jevů příznivých B

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

- **příklad**
  - hází se dvěma kostkami (modrá, zelená)
  - B – součet aspoň 10

$$m = 6 \cdot 6 = 36; \quad m_B = 6 \quad \Rightarrow \quad \mathsf{P}(B) = \frac{6}{36}$$

(příznivé možnosti: 6+4,6+5,6+6,5+5,5+6,4+6) 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)?

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

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

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

nezávislost D, B

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

D, B **nezávislé jevy**

příklad **rodina**: (B – v rodině je jediná dívka, D – nejmladší je dívka)

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

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

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

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

$$\mathsf{P}(B|D) < \mathsf{P}(B) < \mathsf{P}(B|\overline{D})$$

příklad **HWE** (zákon Hardyův-Weinbergův)

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

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

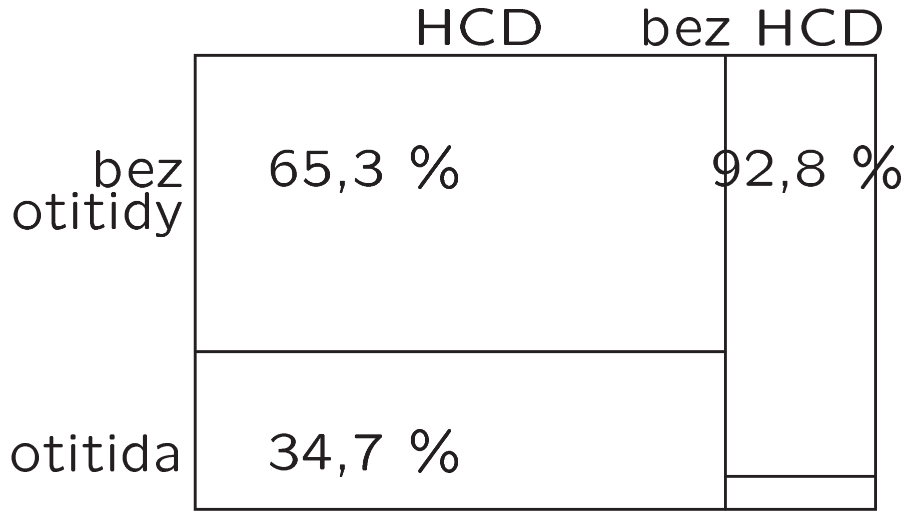
děti (otitidy a záněty HCD)

podmíněno 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

	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



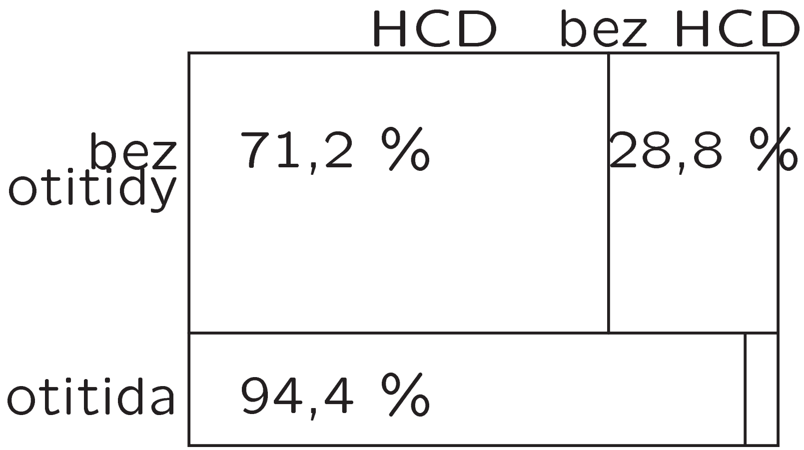
děti (otitidy a záněty HCD)

podmíněno otitídou

	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

	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



předpoklady:

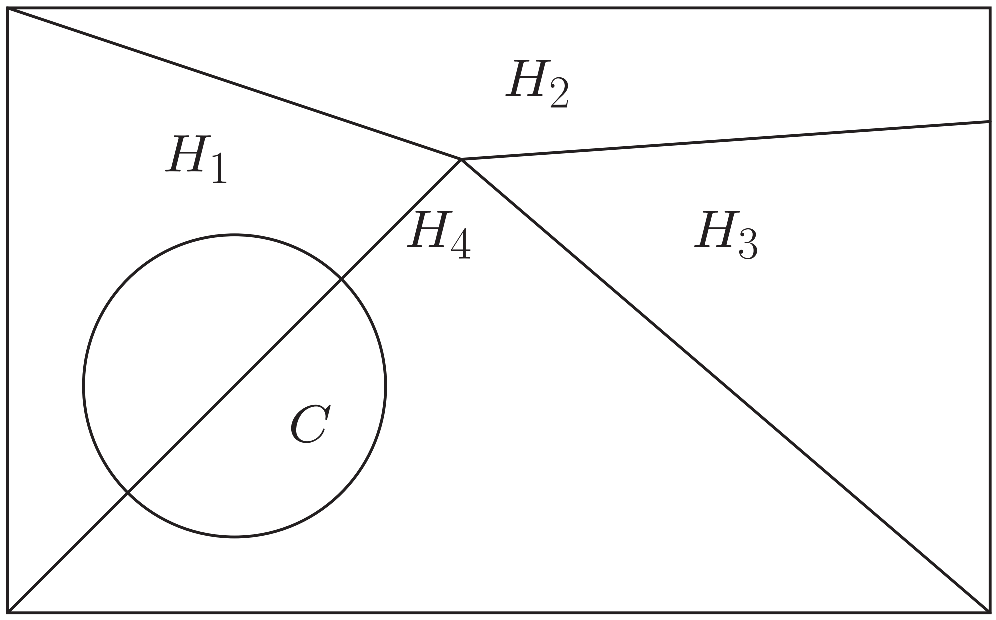
– $H_1, \ldots, H_k$ neslučitelné
– sjednocení $H_1, \ldots, H_k$ – jev jistý

vzorec **pro úplnou pst** $\boxed{\mathsf{P}(C) = \sum_{j=1}^{k} \mathsf{P}(C|H_j)\mathsf{P}(H_j)}$

Bayesův vzorec

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

$H_1, \ldots, H_k$ – 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



$$\mathsf{P}(H_1|C) = \mathsf{P}(H_4|C) = \frac{1}{2}$$

$$\mathsf{P}(H_2|C) = 0$$

$$\mathsf{P}(H_3|C) = 0$$

příklad **děti** C – otitida

H_j – výskyt zánětu HCD

H_j	$\mathsf{P}(H_j)$	$\mathsf{P}(C\mid H_j)$	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

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

pst opakovaného zánětu HCD u otitid $\mathsf{P}(H_3\mid C) = 0{,}728$

pst opakovaného zánětu HCD u všech $\mathsf{P}(H_3) = 0{,}555$

pst opak. zánětu HCD u NEotitid $\mathsf{P}(H_3\mid \overline{C}) = 0{,}485$

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)

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

náhodná veličina

- číselně vyjádřený výsledek náhodného pokusu
- každému elementárnímu jevu přiřadíme reálné číslo
- **diskrétní rozdělení**
  - model pro počty případů (četnosti)
  - možné hodnoty x^*
  - psti hodnot $\mathsf{P}(x_j^*)$ (pstní funkce)
- **spojité rozdělení**
  - model pro spojitou veličiny (délka, váha, koncentrace . . .)
  - interval možných hodnot
  - hustota $f(x)$

Příklad **rodina**: náhodná veličina – počet děvčat

rozdělení X dáno hodnotami x_j^* a pstmi těchto hodnot

ω_i	x_i	x_j^*
(m,m,m)	0	0
(m,m,f)	1	
(m,f,m)	1	1
(f,m,m)	1	
(f,f,m)	2	
(f,m,f)	2	2
(m,f,f)	2	
(f,f,f)	3	3

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

$$m = \sum_{j=1}^{4} m_j = 8$$

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 $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$ zřejmě pak: $f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$
- vlastnosti distribuční funkce

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

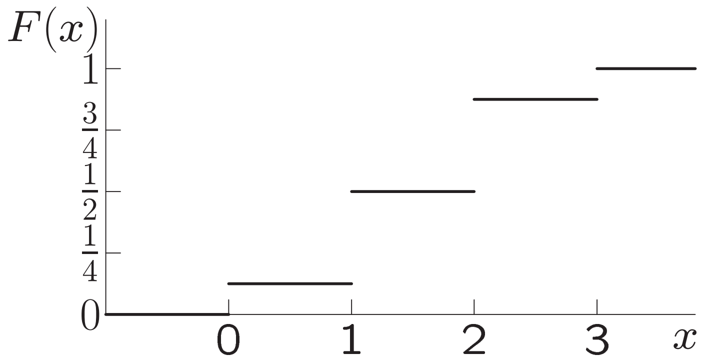
neklesající: $x_1 < x_2 \Rightarrow F(x_2) \ge F(x_1)$

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

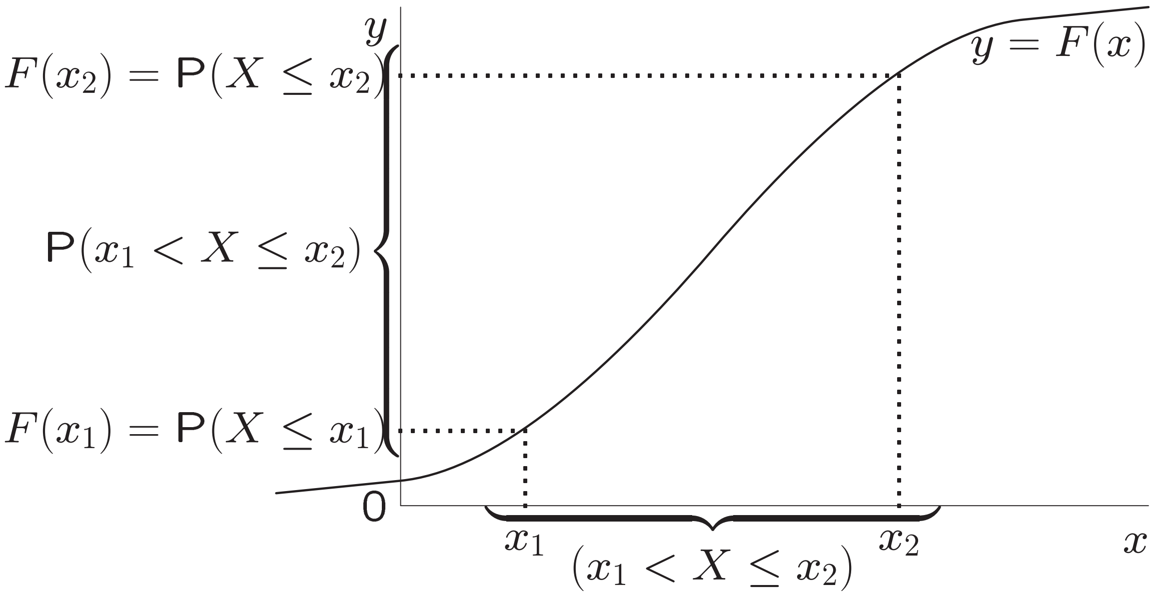
$$\mathsf{P}(X \le x_2) = \mathsf{P}(X \le x_1) + \mathsf{P}(x_1 < X \le x_2)$$
$$F(x_2) = F(x_1) + \mathsf{P}(x_1 < X \le x_2)$$

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	



geometrický význam **distribuční funkce**

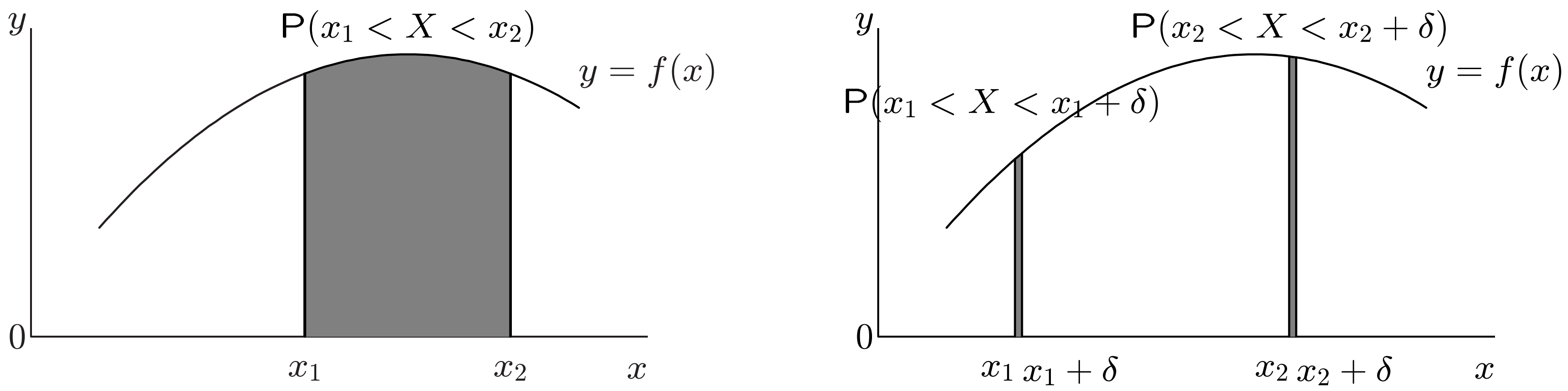


$$\begin{aligned}\mathsf{P}(X \leq x_2) &= \mathsf{P}(X \leq x_1) + \mathsf{P}(x_1 < X \leq x_2)\\ F(x_2) &= F(x_1) + \mathsf{P}(x_1 < X \leq x_2)\end{aligned}$$

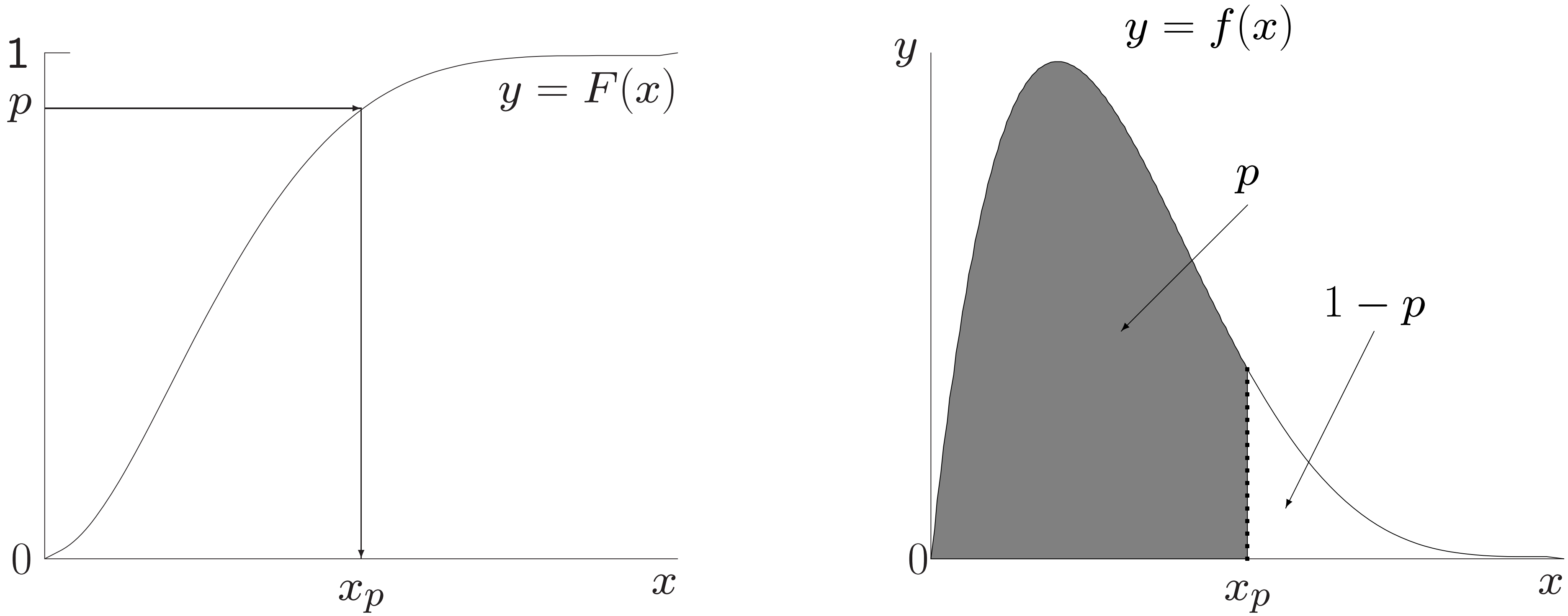
$$\boxed{\mathsf{P}(x_1 < X \leq x_2) = F(x_2) - F(x_1)}$$

význam **hustoty spojitého** rozdělení:

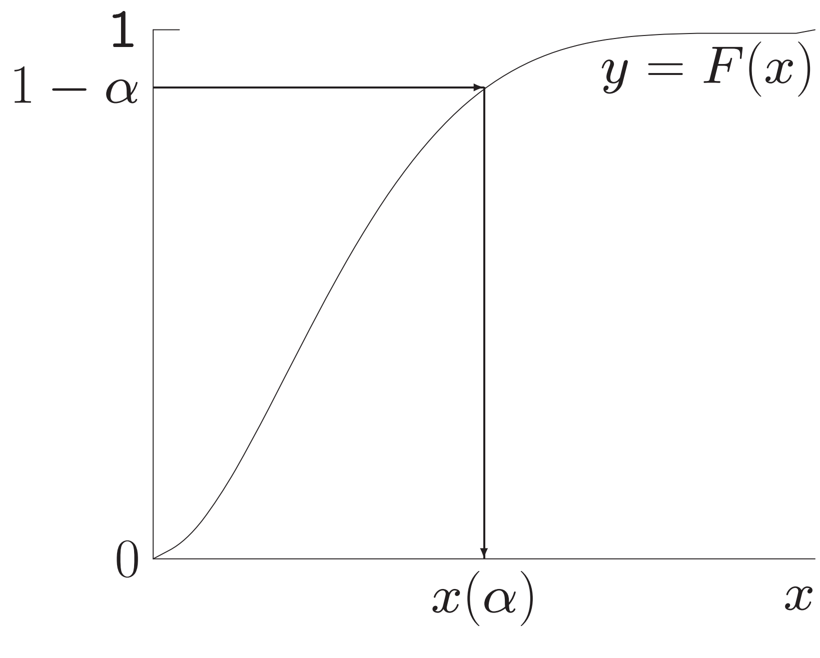
$$f(x) \geq 0, \qquad \int_{-\infty}^{\infty} f(x)\mathsf{d}x = 1$$

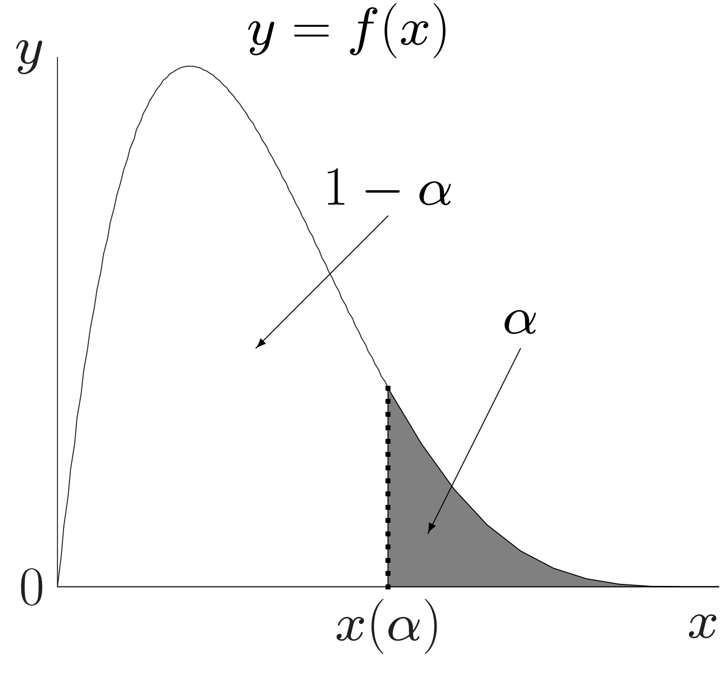


p-**kvantil** x_p je určen požadavkem $\boxed{\mathsf{P}(X \leq x_p) = p}$



kritická hodnota $x(\alpha)$ je určena požadavkem $\boxed{\mathsf{P}(X \geq x(\alpha)) = \alpha}$





$$x_{1-\alpha} = x(\alpha), \qquad x_p = x(1-p)$$

střední hodnota μ

- míra polohy, **populační průměr**
- metoda výpočtu se značí $\mathsf{E}\,X$
- vypočtená hodnota se značí μ
- vážený průměr možných hodnot
- diskrétní (vahami jsou pravděpodobnosti)

$$\boxed{\mu_X = \sum_j x_j^* \mathsf{P}(X = x_j^*)}$$

- spojité (místo vah je hustota)

$$\mu_X = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

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 (**směr. odchylka** σ)

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

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

- diskrétní $\boxed{\sigma_X^2 = \sum_j \left(x_j^* - \mu_X\right)^2 \mathsf{P}\left(X = x_j^*\right)}$
- spojité $\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) \mathrm{d}x$

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	0,28125
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	0,28125
$\sum$		1,000	0,0		0,75000

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

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

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

sdružené rozdělení:

popisuje **společné chování** X, Y 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

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

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

Příklad **rodina**

X počet děvčat v rodině s třemi dětmi

Y počet děvčat mezi dvěma staršími 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

	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

kovariance vyjadřuje závislost náh. veličin:

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

$$\boxed{\sigma_{X,Y} = \sum_i \sum_j (x_i^* - \mu_X)(y_j^* - \mu_Y)\mathsf{P}(X = x_i^*, Y = y_j^*)}$$

označení metody výpočtu: $\mathrm{cov}(X,Y)$

zřejmě platí $\boxed{\mathrm{cov}(X,X) = \mathrm{var}\,X}$ tj. $\boxed{\sigma_{X,X} = \sigma_X^2}$

pro **nezávislé** náhodné veličiny platí

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

jsou-li X, Y – nezávislé $\Rightarrow$ $\boxed{\sigma_{X,Y} = 0}$ (nikoliv obráceně) příklad **děti**: výpočet střední hodnoty, rozptylu a kovariance

x_i^*	y_j^* 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

$$\begin{aligned}
\mu_X &= 0 \cdot 0{,}125 + 1 \cdot 0{,}375 + 2 \cdot 0{,}375 + 3 \cdot 0{,}125 = 1{,}5 \\
\mu_Y &= 0 \cdot 0{,}250 + 1 \cdot 0{,}500 + 2 \cdot 0{,}250 = 1 \\
\sigma_X^2 &= (0 - 1{,}5)^2 \cdot 0{,}125 + \ldots + (3 - 1{,}5)^2 \cdot 0{,}125 = 0{,}75 \\
\sigma_Y^2 &= (0 - 1)^2 \cdot 0{,}25 + (1 - 1)^2 \cdot 0{,}5 + (2 - 1)^2 \cdot 0{,}25 = 0{,}5 \\
\sigma_{XY} &= (0 - 1{,}5) \cdot (0 - 1) \cdot 0{,}125 + (1 - 1{,}5) \cdot (1 - 1) \cdot 0{,}250 + \ldots = 0{,}5
\end{aligned}$$

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

vlastnosti populačního průměru a rozptylu
(srovnej s požadavky na míry polohy a míry variability)

$$\mu_{\alpha+X} = \alpha + \mu_X, \qquad \mu_{\beta X} = \beta\mu_X,$$
$$\sigma^2_{\alpha+X} = \sigma^2_X, \qquad \sigma^2_{\beta X} = \beta^2\sigma^2_X,$$
$$\sigma_{\alpha+X} = \sigma_X, \qquad \sigma_{\beta X} = |\beta|\sigma_X,$$

pro součet náhodných veličin $X+Y$ platí

$$\mu_{X+Y} = \mu_X + \mu_Y$$
$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\sigma_{XY}$$
$$\sigma_{X,Y} = 0 \qquad \text{pro nezávislé } X, Y$$
$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y \qquad \text{pro nezávislé } X, Y$$

ukázka důkazu:

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

normování náhodné veličiny X (populační obdoba z-skórů)

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

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

normovaná verze veličiny umožňuje vyšetřovat vlastnosti nezávislé na poloze μ_X a variabilitě (měřítku) σ_X^2:

(populační) **korelační koeficient** (correlation coefficient)

$$\rho_{XY} = \mathsf{cov}\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

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

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

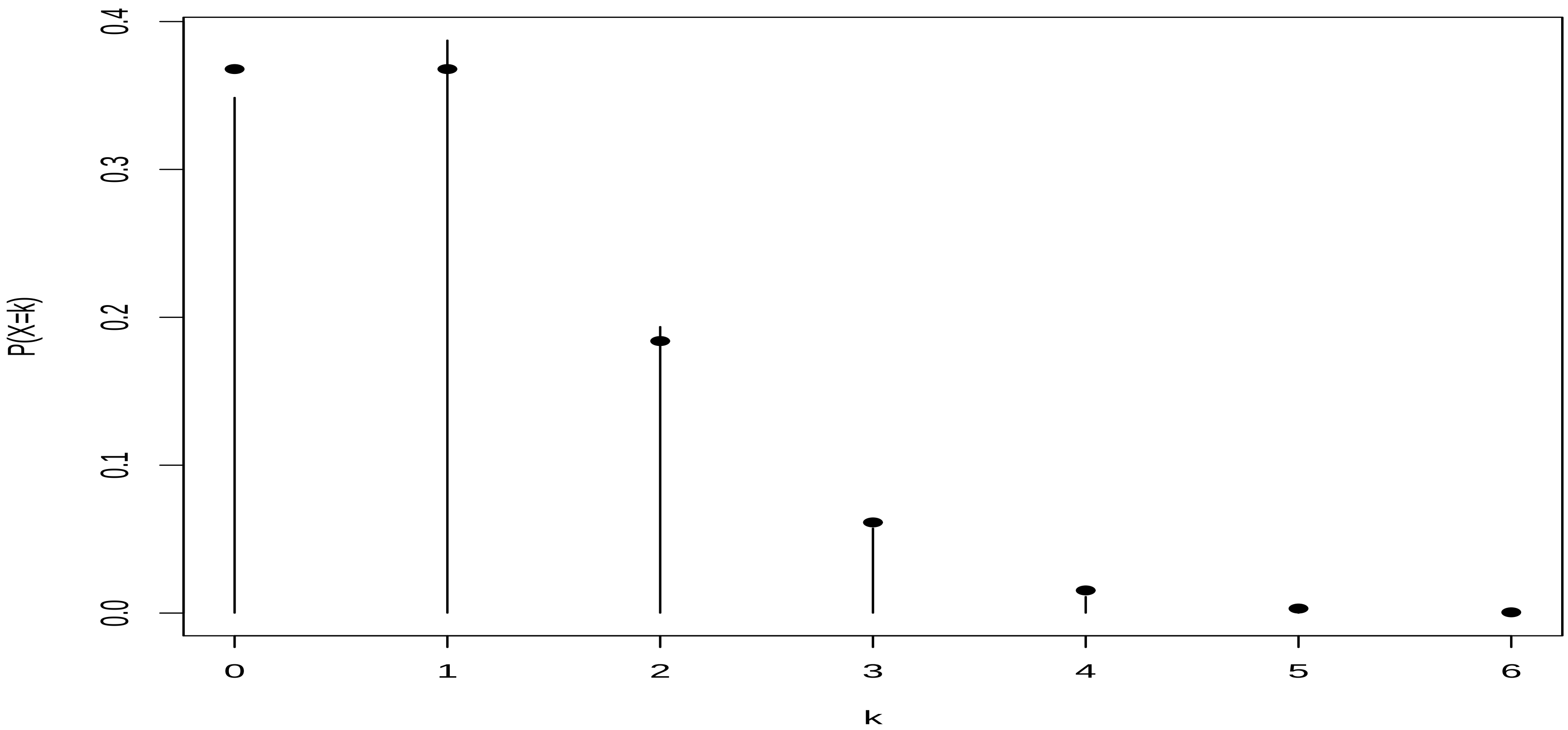
(populační) **špičatost** náhodné veličiny X (kurtosis, někdy se neodečítá 3)

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

Důležitá diskrétní rozdělení

- **alternativní** (nula-jedničkové) rozdělení
  - *zdar* nebo *nezdar* (pouze dvě možné hodnoty)
  - $\mathsf{P}(X = 1) = \pi$, $\mathsf{P}(X = 0) = 1 - \pi$, $(0 < \pi < 1)$
  - $\mathsf{E}\, X = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi$
  - $\operatorname{var} X = (1 - \pi)^2 \cdot \pi + (0 - \pi)^2 \cdot (1 - \pi) = \pi(1 - \pi)$
- **binomické rozdělení** $Y \sim \mathrm{bi}(n, \pi)$
  - n **nezávislých** pokusů takových, že
  - $\mathsf{P}(\text{zdar}) = \pi$, $\mathsf{P}(\text{nezdar}) = 1 - \pi$, $(0 < \pi < 1)$
  - Y je **počet zdarů** v těchto pokusech
  - $\boxed{\mathsf{P}(Y = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}, \quad k = 0, 1, \ldots, n}$

- vyjádření binomického rozdělení pomocí alternativního
  - $Y = \sum_{i=1}^{n} X_i$, X_i − zda zdar v i-tém pokusu
  - $\mathsf{E}\, Y = \mathsf{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{E}\, X_i = n\pi$
  - $\mathsf{var}\, Y = \mathsf{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{var}\, X_i = n\pi(1-\pi)$ (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 . . .
  - $\boxed{\mathsf{P}(X = k) = \frac{\lambda^k}{k!}\mathsf{e}^{-\lambda}}, \qquad k = 0, 1, \ldots$
  - $\mathsf{E}\, X = \lambda$, $\mathsf{var}\, X = \lambda$
  - pro velké n a malé π lze rozdělení $\mathsf{bi}(n, \pi)$ aproximovat pomocí rozdělení $\mathsf{Po}(n\pi)$



bi$(10, 0{,}1)$ (úsečky) aproximované pomocí Po(1) (tečky)

- **normální** (Gaussovo) rozdělení $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$
  - $\boxed{\mathsf{E}\,X = \mu,\ \mathsf{var}\,X = \sigma^2}$
  - $\mathsf{N}(0,1)$: $\varphi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$, $\Phi(x) = \int_{-\infty}^{x} \varphi(t)\mathrm{d}t$ (hustota, distr. fce)
  - $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$, pak $Z = \frac{X-\mu}{\sigma} \sim \mathsf{N}(0,1)$

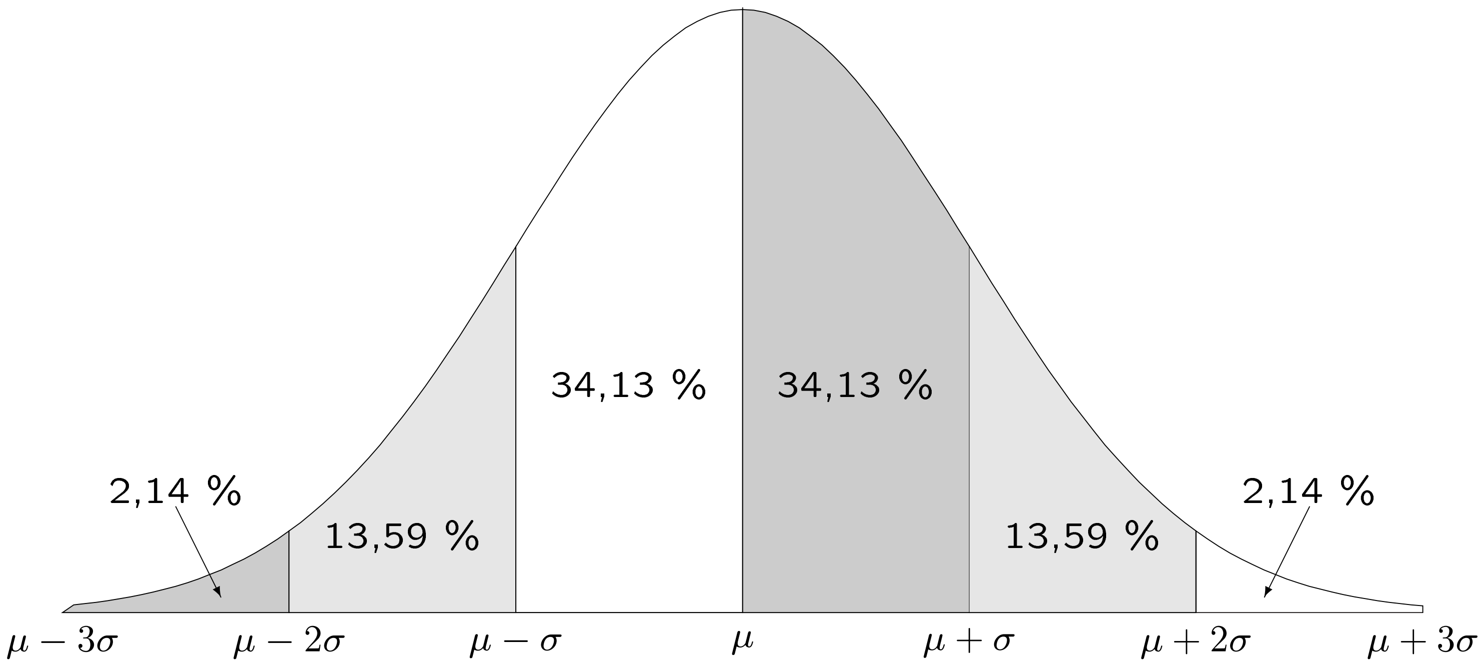
$$\boxed{\mathsf{P}(a < X < b) = \mathsf{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

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

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

  - **aproximace binomického rozdělení** $\mathsf{bi}(n, \pi)$ **normálním** $\mathsf{N}(n\pi, n\pi(1-\pi))$ (použitelné, pokud $n\pi(1-\pi) > 9$)

hustota N$\left(\mu, \sigma^2\right)$



kritické hodnoty

- **normální rozdělení** $\mathsf{N}(0,1)$

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

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

- **Studentovo t-rozdělení** t_k (podobné normálnímu, ale používá odhad s parametru σ, proto má větší rozptyl)

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

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

- **Fisherov**o **F-rozdělení** $\mathsf{F}_{k,m}$

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

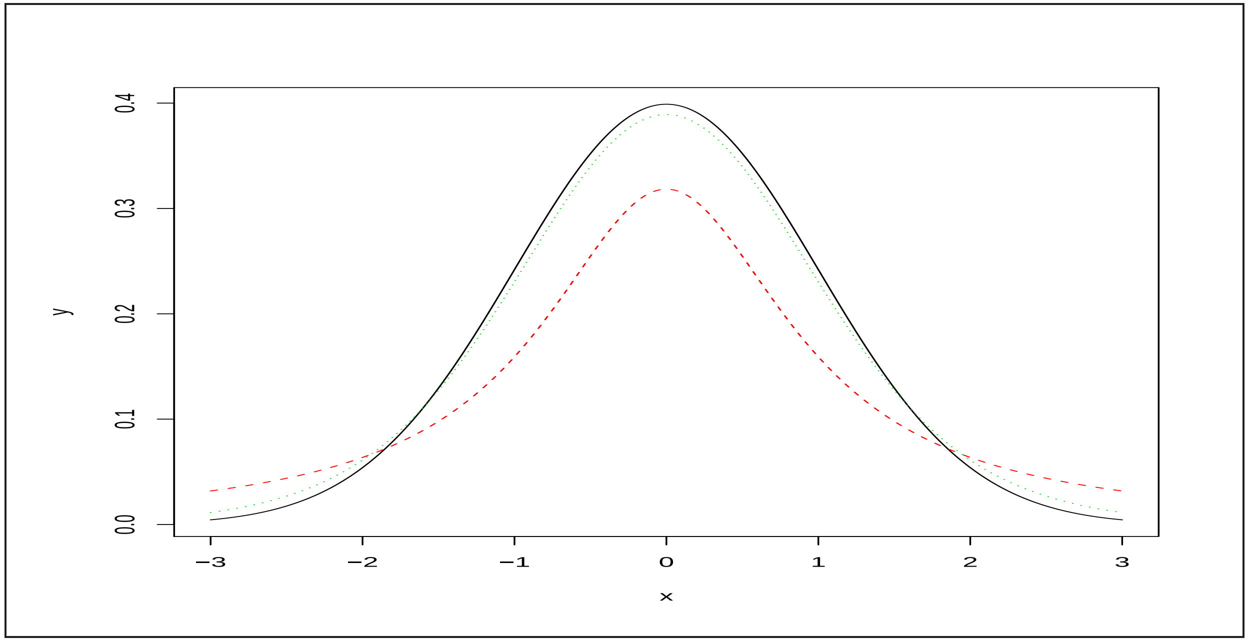
- **rozdělení chí-kvadrát** χ^2_k

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

speciálně platí $\chi^2_1(0{,}05) = 1{,}960^2 = 3{,}841$ (viz $z(0{,}025) = 1{,}960$)

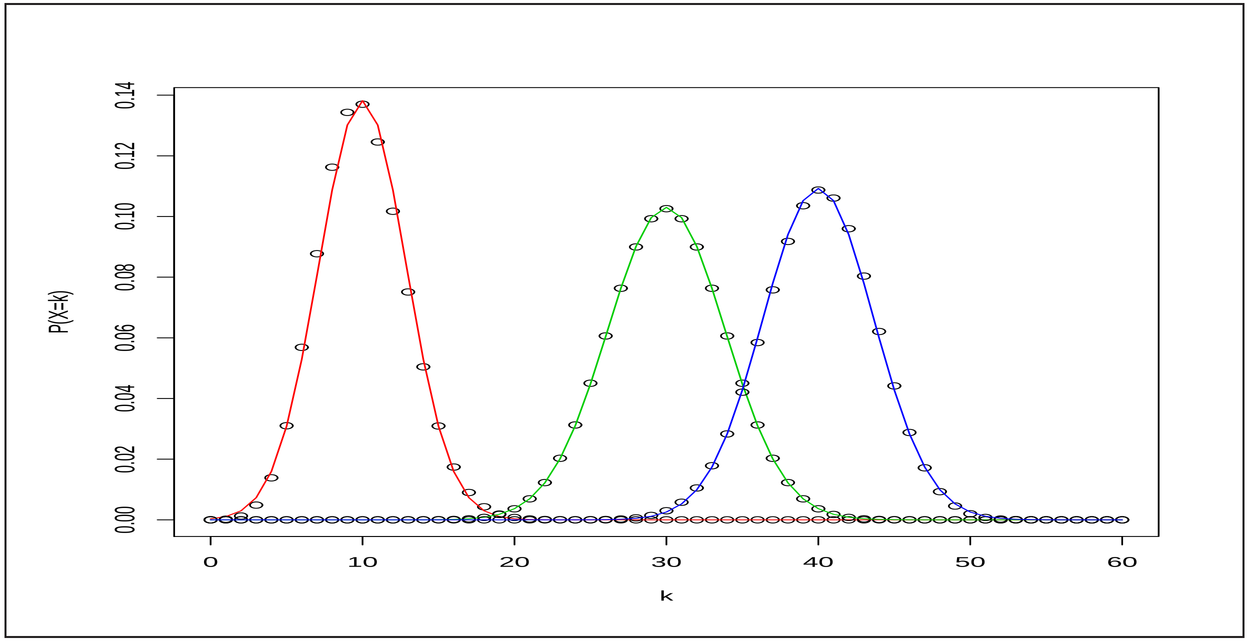
srovnání normálního a Studentova t-rozdělení

čárkovaně t_1, tečkovaně t_{10}, plná čára $\mathrm{N}(0,1)$)



porovnání binomického a normálního rozdělení

bi$(60, 1/6)$, bi$(60, 3/6)$, 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 vyšetřovaná část populace (vzorek)
- **reprezentativní** výběr obráží poměry v populaci (vlastnost)
- **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (naměřené na výběru)
- **parametr** neznámé číslo popisující nějaký rys populace, charakteristika rozdělení náhodné veličiny
- **statistika** funkce náhodného výběru (pozorování)
- **odhad** statistika použitá k odhadu parametru

vlastnosti výběrového průměru

- $X_1, \ldots, X_n$ nezávislé, stejné rozdělení **náhodný výběr**
  $\mathsf{E}\, X_i = \mu$ 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 (unbiased estimator) populačního průměru
- $\boxed{\operatorname{var} \bar{X} = \frac{\sigma^2}{n} = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = (\mathsf{S.E.}(\bar{X}))^2}$ = (**střední chyba průměru**)^2
  (standard error of mean) variabilita průměrů z výběrů rozsahu n je n-krát menší, než u jednoho pozorování, střední chyba průměru je $\sqrt{n}$-krát menší než σ

- pro **normální** rozdělení $X_i \sim \mathsf{N}\Big(\mu, \sigma^2\Big)$

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

- **interval spolehlivosti** pro μ: $\mathsf{P}(|\bar{X} - \mu|/\mathsf{S.E.}(\bar{X}) \leq z(\alpha/2)) = 1 - \alpha$

$$\boxed{(\bar{X} - \mathsf{S.E.}(\bar{X}) \cdot z(\alpha/2); \bar{X} + \mathsf{S.E.}(\bar{X}) \cdot z(\alpha/2))}$$

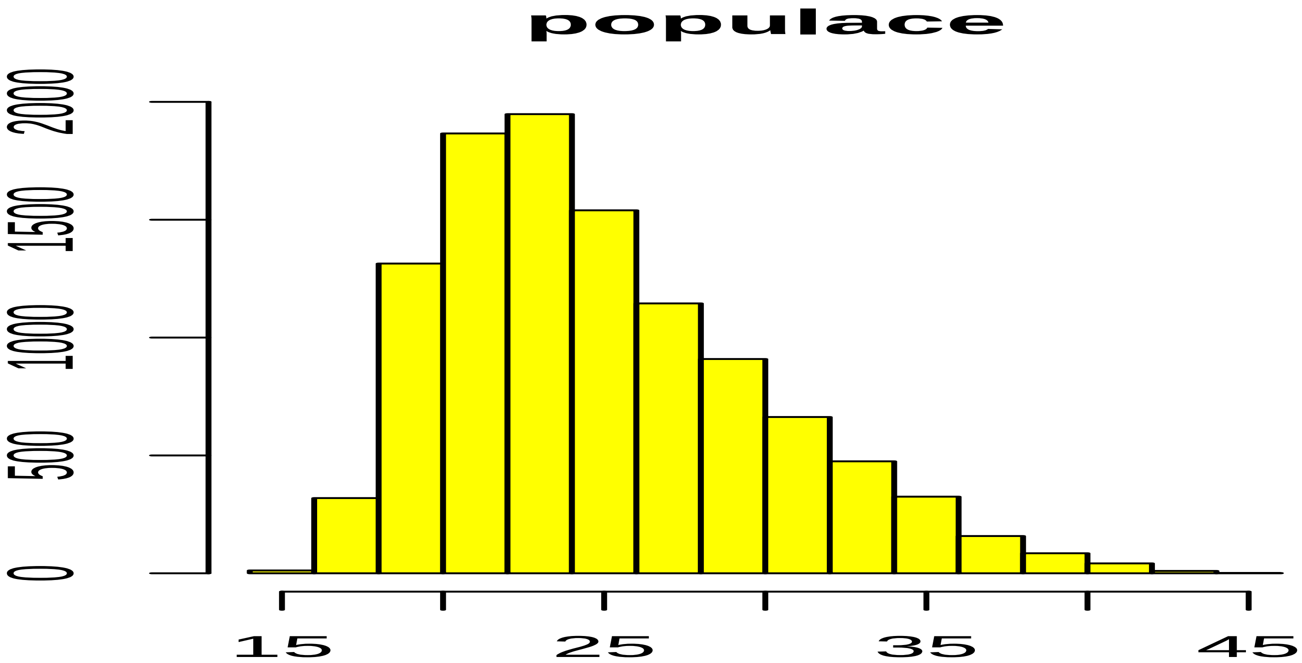
$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2); \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right)$$

- požadujeme int. spolehlivosti šířky $2c$; k tomu stačí

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

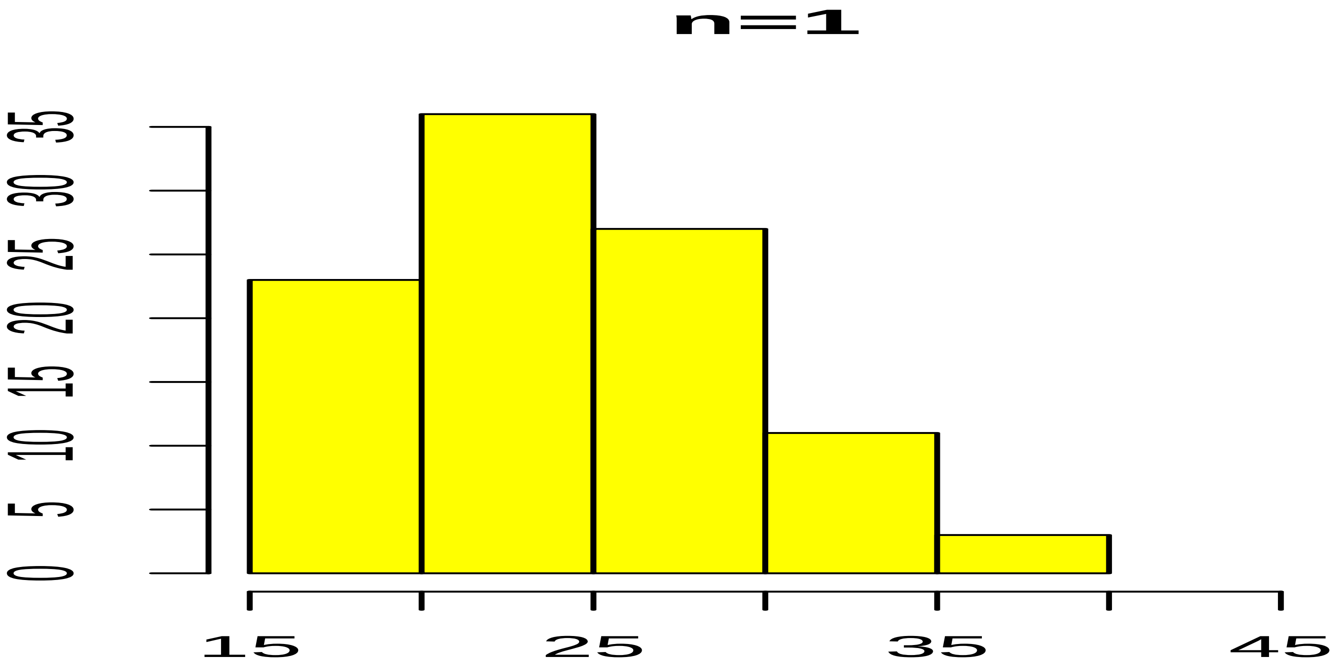
příklad: věk matek

- velká populace rodičů (11 tisíc)



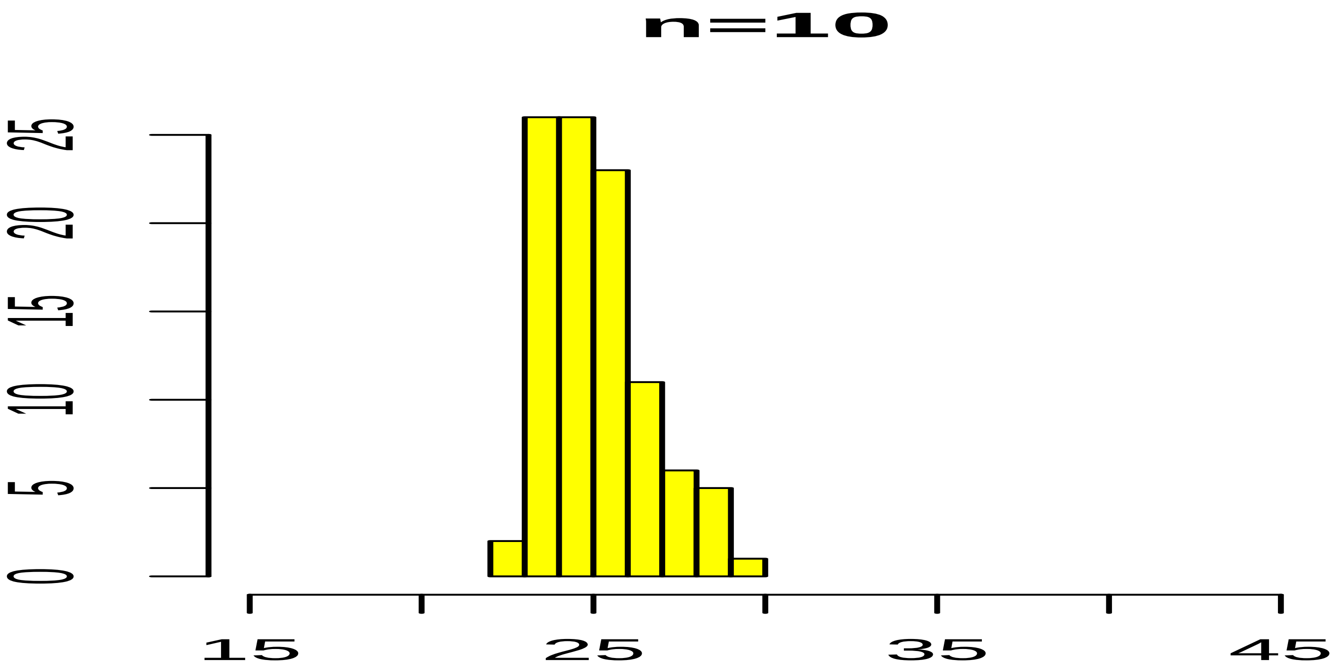
příklad: věk matek

- náhodně vybráno 100 matek (průměry rozsahu $n = 1$)



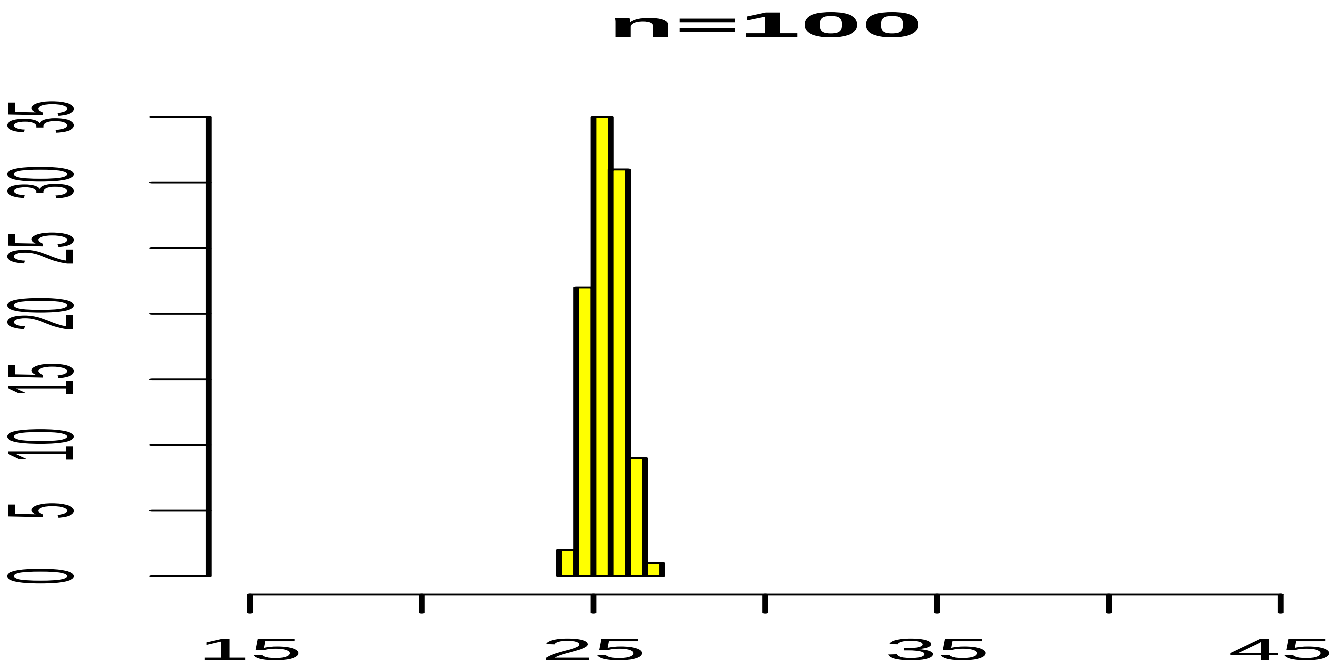
příklad: věk matek

- náhodně vybráno 100 krát po $n = 10$ matkách, průměry:



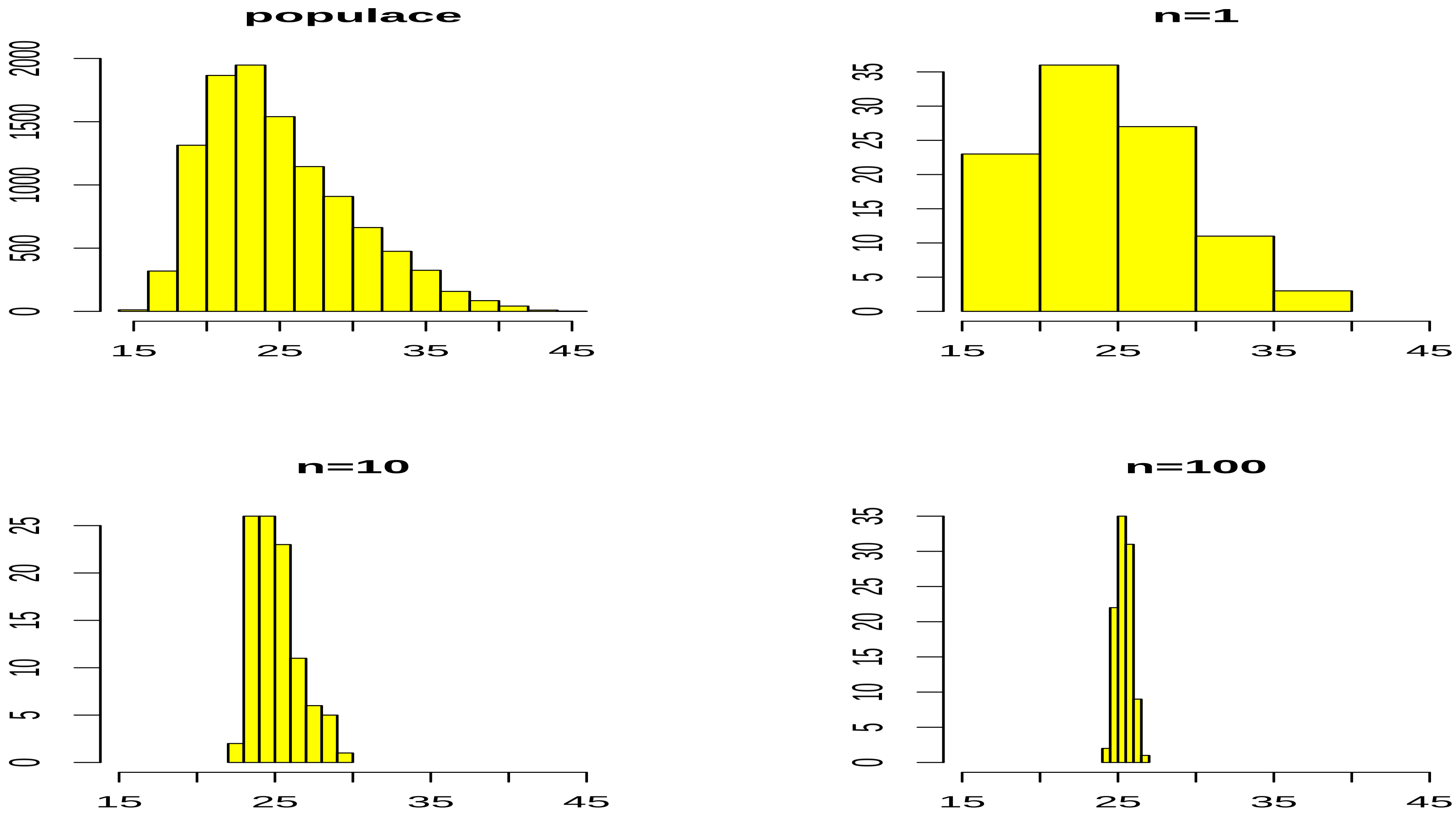
příklad: věk matek

- náhodně vybráno 100 krát po $n = 100$ matkách, průměry:



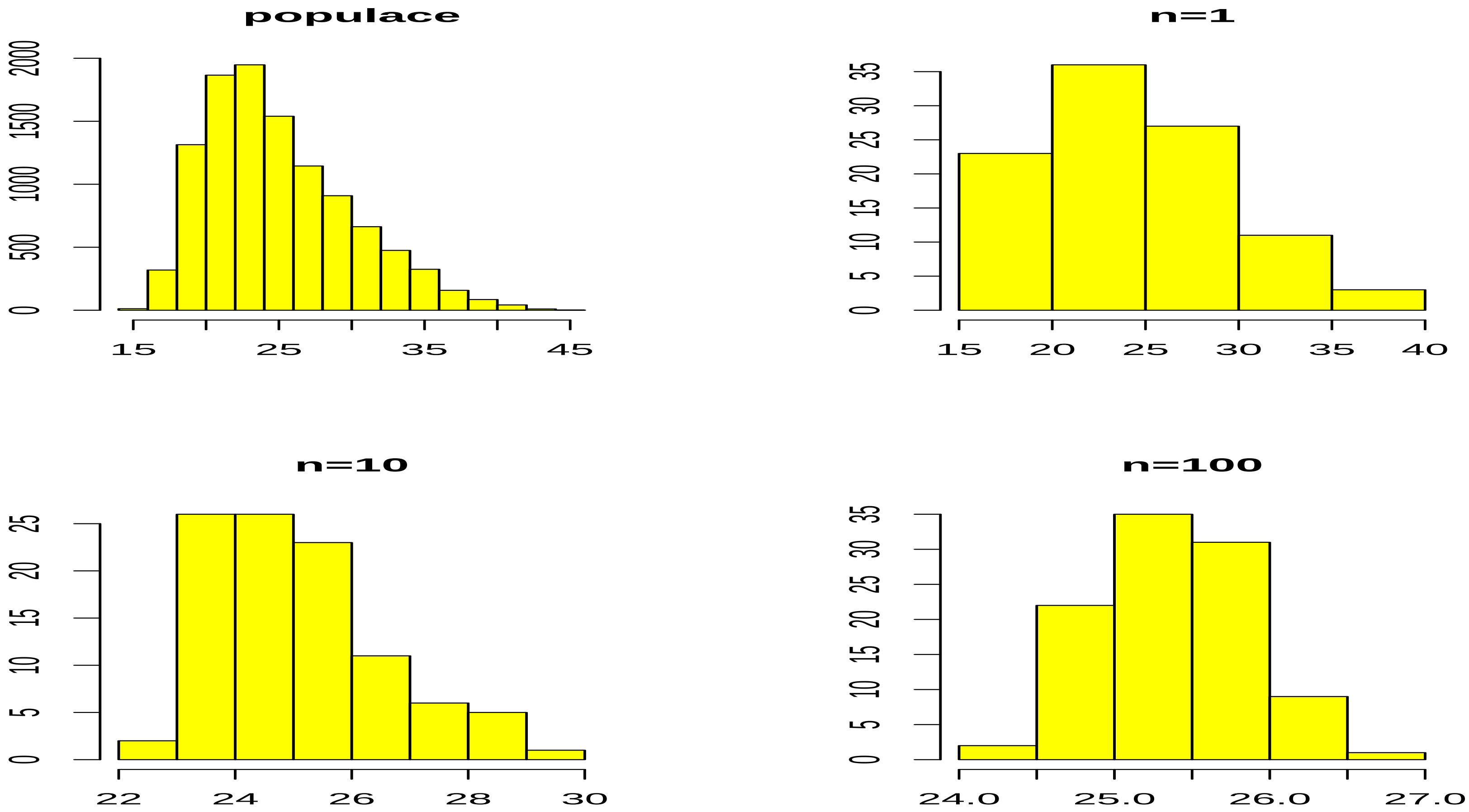
příklad: věk matek

- velká populace rodičů (11 tisíc)
- náhodně vybráno 100 matek (vlastně průměry výběrů rozsahu $n = 1$), nakreslen histogram
- 100 krát náhodně vybráno vždy $n = 10$ matek, spočítán průměr, nakreslen histogram průměrů
- 100 krát náhodně vybráno vždy $n = 100$ matek, spočítán průměr, nakreslen histogram průměrů
- podle teorie by každý další rozptyl ze 100 průměrů měl být 10 krát menší
- skutečnost (odhady ze 100 realizací): 23,5; 2,20; 0,21


populace
n=1
n=10
n=100
0
500
1000
1500
2000
5
10
15
20
25
30
35
45

centrální limitní věta

- Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$. (Nemusí být normální rozdělení.) Potom **pro velké** n má průměr z nich rozdělení N$\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení N$\left(n\mu, n\sigma^2\right)$.
- prakticky: pro dost velká n má **průměr normální rozdělení** s rozptylem n-krát menším než jednotlivá pozorování
- příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení


populace
0
500
1000
1500
2000
15
25
35
45
n=1
0
5
10
15
20
25
30
35
15
20
25
30
35
40
n=10
0
5
10
15
20
25
22
24
26
28
30
n=100
0
5
10
15
20
25
30
35
24.0
25.0
26.0
27.0

průměrný věk matek v opakovaných výběrech:

rozsah výběru n	průměr průměrů	směr. odch. průměrů	šikmost průměrů	špičatost průměrů
1	24,74	4,848	0,682	-0,040
10	25,14	1,482	0,743	-0,199
100	25,40	0,455	0,087	-0,076
1000	25,40	0,146	0,156	-0,212
populace	μ =25,41	σ =4,932	γ_1 =0,771	γ_2 =0,189

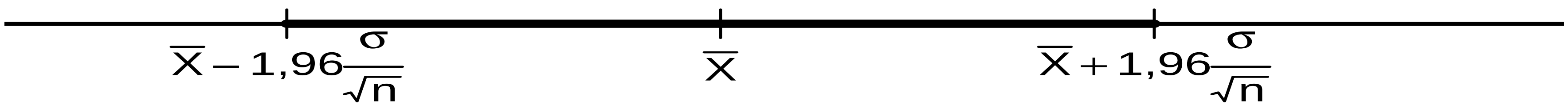
interval spolehlivosti (1)

- protože je $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$, platí $\bar{X} \sim \mathsf{N}\left(\mu, (\sigma/\sqrt{n})^2\right)$

$$\mathsf{P}\left(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| < 1{,}96\right) = 0{,}95$$

  tedy $\mathsf{P}\left(\bar{X} - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}95$

- dostali jsme **95% interval spolehlivosti** pro μ



interval spolehlivosti (2)

- 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé** μ (**odhadovaný parametr**)
- kdybychom postup prováděli opakovaně, pak asi v 95 % případů překryjeme skutečnou hodnotu μ, ve zbylých asi 5 % zůstane skutečné μ mimo interval spolehlivosti
- pro velké n lze neznámé σ nahradit odhadem s_x
- pro obecné α (spolehlivost $1-\alpha$):

$$\mathsf{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

interval spolehlivosti (3)

- pro malé n (asi do 50) a pro X_i s normálním rozdělením lépe použít kritické hodnoty Studentova t-rozdělení (pozor na **jinak značené** kritické hodnoty Studentova t-rozdělení)

$$\mathsf{P}\left(\bar{X} - \frac{s_x}{\sqrt{n}} t_{n-1}(\alpha) \leq \mu \leq \bar{X} + \frac{s_x}{\sqrt{n}} t_{n-1}(\alpha)\right) = 1 - \alpha$$

- interval spolehlivosti lze počítat i pro jiné parametry
- je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

příklad: věk matek

- 95% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

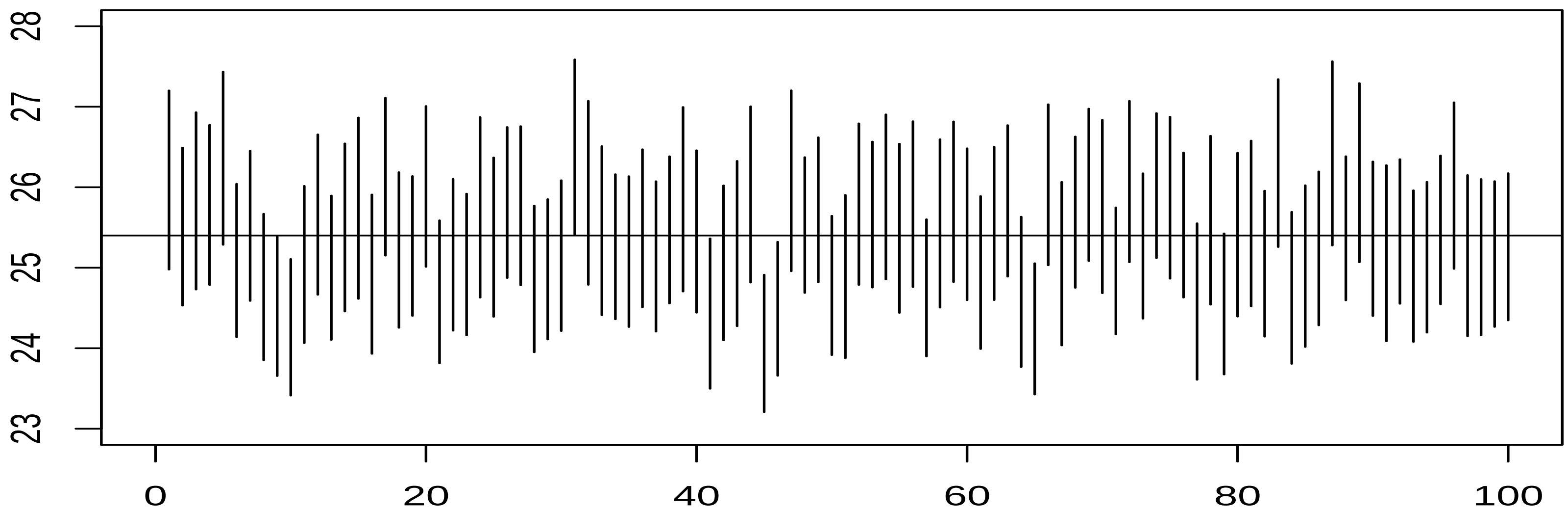
$$\left(25{,}7 - 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}9; 26{,}5)$$

- 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)

$$\left(25{,}7 - 2{,}63 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 2{,}63 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}6; 26{,}8)$$

- větší jistota $\Leftrightarrow$ větší šířka

příklad: simulované výběry pro $n = 100$



celkem 100 95% intervalů spolehlivosti pro μ (ve skutečnosti mimořádně víme, že $\mu = 25{,}4$), v 7 případech μ nepřekryto

statistické rozhodování

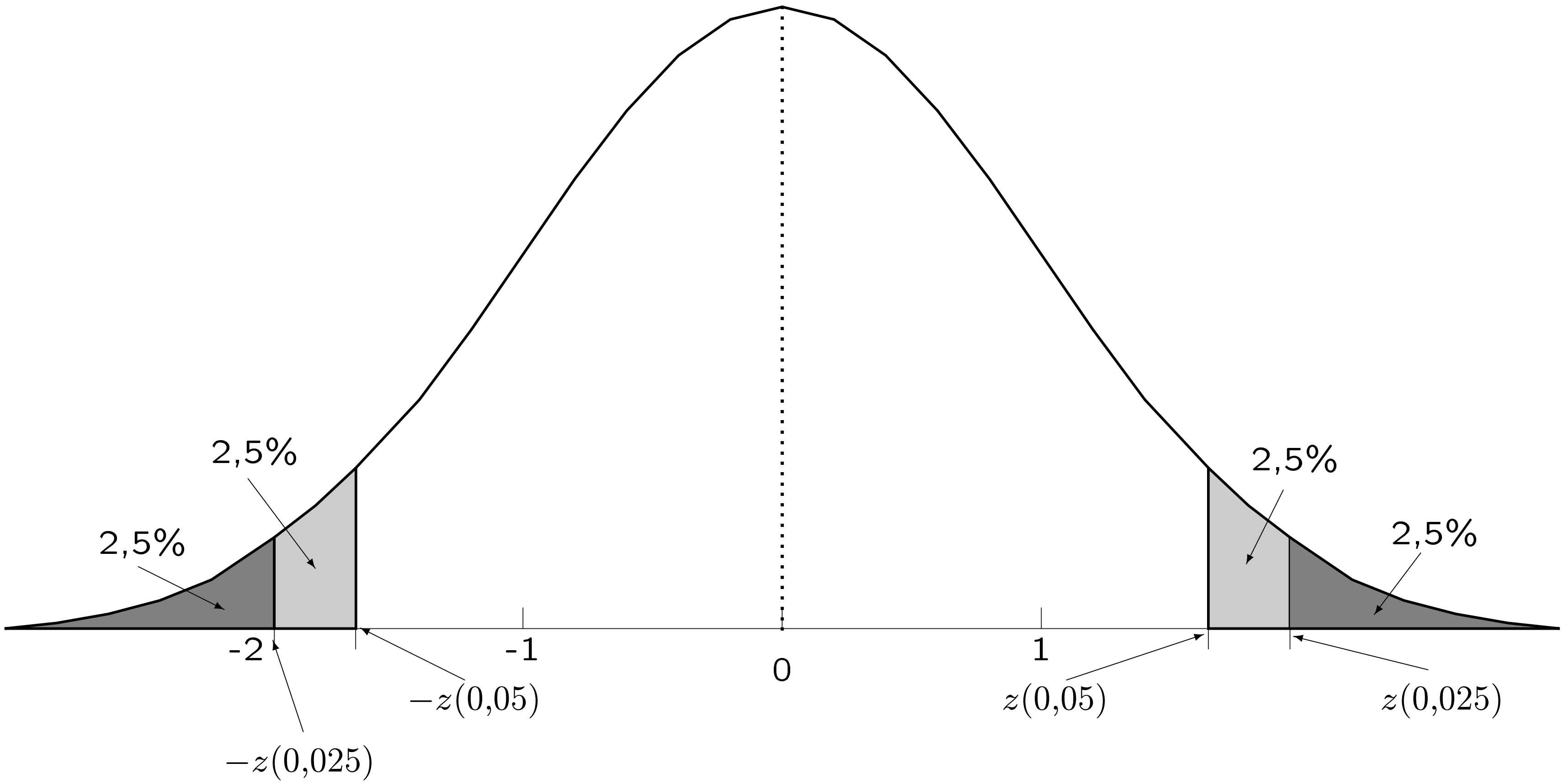
- **nulová hypotéza** H_0 tvrzení o populaci (parametru), o jehož platnosti chceme rozhodnout (zamítnout)
- **alternativní hypotéza** H_1 (alternativa) zbývající možnost (k H_0) často „vědecká hypotéza“
- **kritický obor** možné výsledky pokusu, kdy H_0 zamítáme; zpravidla popsán pomocí statistiky (např. $|T| \geq t_{n-1}(\alpha)$)
- **obor přijetí** možné výsledky pokusu, kdy H_0 nezamítáme
- **chyba prvního druhu** rozhodnutí zamítnout H_0, když platí H_0 falešně prokázat „vědeckou hypotézu“
- **chyba druhého druhu** rozhodnutí nezamítnout H_0, když platí H_1
- **hladina testu** α (zpravidla 5 %, 1 %) maximální dovolená pst chyby prvního druhu; volí před pokusem, nezávisle na výsledku

- **síla testu** $1-\beta$ pravděpodobnost zamítnutí neplatné H_0; prokážeme platnou „vědeckou hypotézu“
- **dosažená hladina testu** p (p-**hodnota**) 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 T
- H_0 se **zamítá**, když $\boxed{p \leq \alpha}$

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)

rozhodování o populačním průměru normálního rozdělení, σ známé

- $X_1, \ldots, X_n \sim \mathsf{N}\big(\mu, \sigma^2\big)$ **nezávislé**; $\sigma > 0$ známe
- $\bar{X} \sim \mathsf{N}\big(\mu, \sigma^2/n\big)$, tedy S.E.$(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- $\mathsf{H}_0 : \mu = \mu_0$ (dané číslo)
- platí-li H_0, pak $\boxed{Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim \mathsf{N}(0,1)}$
- $\mathsf{H}_1 : \mu \neq \mu_0 \Rightarrow$ kritický obor: $|Z|$ velké, tj. $\boxed{|Z| \geq z(\alpha/2)}$
- $\mathsf{H}_1 : \mu > \mu_0$: zamítnout pro $\boxed{Z \geq z(\alpha)}$
- $\mathsf{H}_1 : \mu < \mu_0$: zamítnout pro $\boxed{Z \leq -z(\alpha)}$



Hustota Z za platnosti H_0

příklad **výšky** desetiletých hochů ([cm]) měřené v roce 1961
$\sigma = 6{,}4$ cm (známo z dřívějška), $\alpha = 0{,}05$,
$\mathsf{H}_0 : \mu = 136{,}1$ cm (před 10 lety, v roce 1951), $\mathsf{H}_1 : \mu \neq 136{,}1$ cm

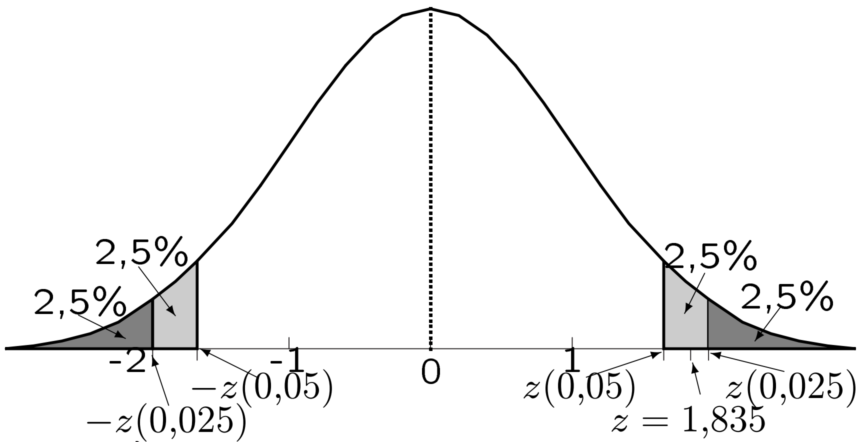
130	140	136	141	139	133	149	151
138	142	127	139	147	139	136	

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

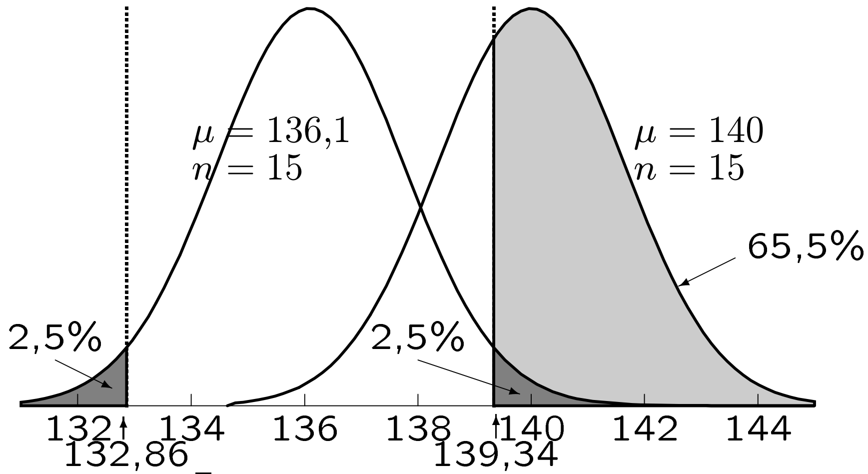
$|z| = 1{,}835 < z(0{,}05/2) = 1{,}960 \Rightarrow \mathsf{H}_0$ nelze na 5% hladině zamítnout
ale $|z| \geq z(0{,}10/2) = 1{,}645 \Rightarrow \mathsf{H}_0$ se na 10% hladině zamítá
$\Rightarrow$ p-hodnota mezi 5 % a 10 %

$p = \mathsf{P}(|Z| \geq 1{,}835) = 2 \cdot \mathsf{P}(Z > 1{,}835) = 0{,}067$, tedy $p = 6{,}7$ %
v naší úloze je rozumnější předem zvolená **jednostranná alternativa** (víme totiž, že po nějakou dobu platí, že každá následující generace je větší než ta předcházející; pokoušíme se totéž dokázat i z těchto dat):
$\mathsf{H}_1 : \mu > 136{,}1$: $z = 1{,}835 \geq 1{,}645$ na 5 % zamítnout
$p = \mathsf{P}(Z \geq 1{,}835) = 0{,}033$ ($< 0{,}05$)

výšky desetiletých hochů (opět uvažujeme oboustrannou alternativu)



a) Hustota Z za platnosti H_0



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

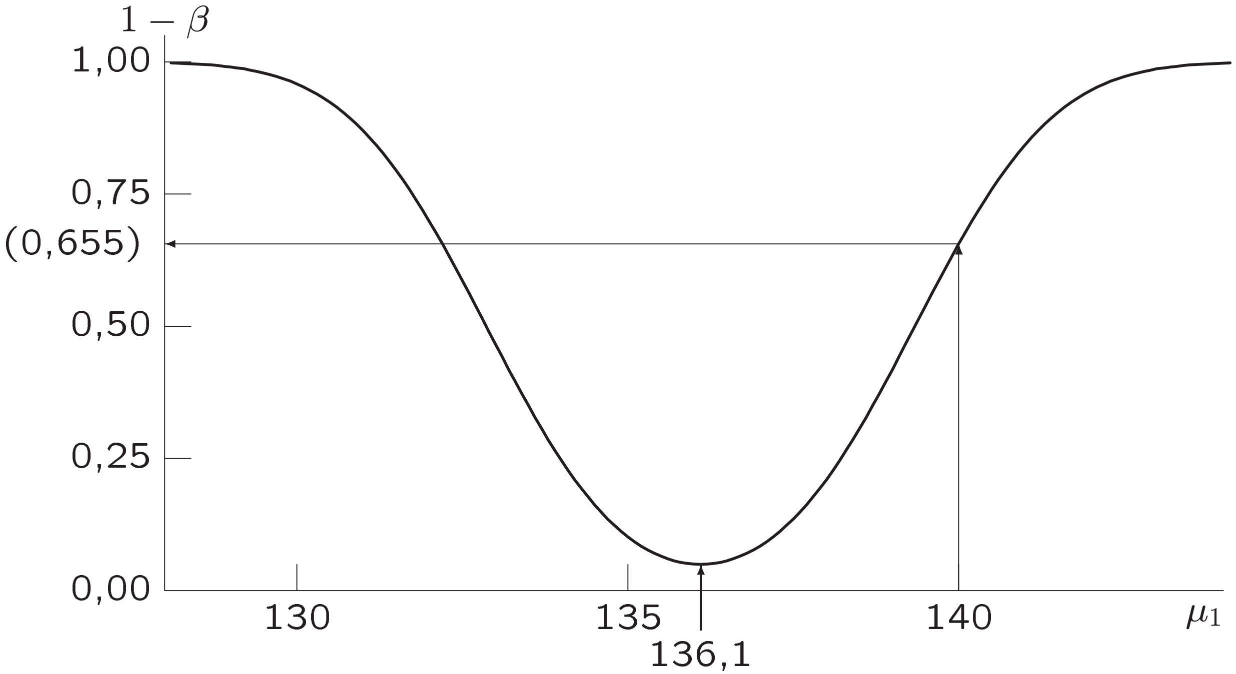
$$\text{S.E.}(\bar{X}) = \sqrt{\frac{6{,}4^2}{15}} = 1{,}6525 \Rightarrow 136{,}1 - 1{,}6525 \cdot 1{,}96 = 132{,}86$$

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

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 zvolenou hodnotu μ_1 požadujeme sílu $1-\beta$ (volíme pravděpodobnost, s jakou odhalíme neplatnost H$_0$, je-li skutečnost $\mu = \mu_1$):

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

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

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

(místo 15 pozorování jich potřebujeme aspoň 29)

jednovýběrový t-test

- n nezávislých pozorování $X_1, \ldots, X_n$ z normálního rozdělení $\mathsf{N}\left(\mu, \sigma^2\right)$
- $\mathsf{H}_0 : \mu = \mu_0$ (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 (místo σ S) $\boxed{T = \frac{\bar{X} - \mu_0}{S}\sqrt{n} = \frac{\bar{X} - \mu_0}{\mathsf{S.E.}(\bar{X})}}$
- $\mathsf{H}_1 : \mu \neq \mu_0$ 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

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

příklad **výšky hochů** pro případ neznámého rozptylu ($\mathrm{H}_1 : \mu \neq 136{,}1$)

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 = 130^2 + \ldots + 147^2 - 15 \cdot 139{,}133^2 = 601{,}733$$

$$s^2 = \frac{601{,}733}{15-1} = 42{,}981 = 6{,}556^2$$

$$t = \frac{139{,}133 - 136{,}1}{6{,}556}\sqrt{15} = 1{,}792 < 2{,}145 = t_{14}(0{,}05)$$

$$p = \mathsf{P}(|T| \geq 1{,}792) = 0{,}0948 \qquad (\text{ tj. } 9{,}48\ \%)$$

H_0 se na 5% hladině **nezamítá**

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

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

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

interval obsahuje $\mu_0 = 136{,}1 \Rightarrow \mathsf{H}_0$ se na 5% hladině **nezamítá**

jednostranná alternativa $\mathsf{H}_1 : \mu > 136{,}1$:

$$\begin{array}{lll}
t \geq t_{14}(2 \cdot 0{,}05) = 1{,}761 & \Rightarrow & \text{zamítnout } \mathsf{H}_0\,(\alpha = 5\%) \\
t < t_{14}(2 \cdot 0{,}01) = 2{,}624 & \Rightarrow & \text{nezamítnout } \mathsf{H}_0\,(\alpha = 1\%) \\
p = \mathsf{P}(T > t) = 0{,}0474 & \Rightarrow & (4{,}74\ \%)
\end{array}$$

párové testy

- $(U_1, V_1), \ldots, (U_n, V_n)$ **nezávislé** dvojice (závislých) náhodných veličin (párová pozorování)
- výhodná je těsná závislost uvnitř dvojic
- $X_i = U_i - V_i$ (označení rozdílů)
  $X_1, \ldots, X_n$ mají **stejné** rozdělení
- dvojice měření na stejných jedincích (hodnota před ošetřením a po něm)
- věk otce a věk matky
- výška otce a výška syna

- **párový t-test**
  - **normální** rozdělení: $X_i = U_i - V_i \sim \mathrm{N}\Big(\mu, \sigma^2\Big)$ **nezávislé**
  - $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
  - $T = \dfrac{\bar{X}}{\text{S.E.}(\bar{X})} = \dfrac{\bar{X}}{S}\sqrt{n} = \dfrac{\bar{U} - \bar{V}}{\text{S.E.}(\bar{U} - \bar{V})}$
  - $\mathrm{H}_0 : \mu = 0$ (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
- $\alpha = 0{,}05, \mathsf{H}_0 : \mu_U - 10 = \mu_V$ resp. $\mu_U - \mu_V = 10$
- $n = 99, \bar{u} = 179{,}267, \bar{v} = 166{,}970$
- $\bar{x} = \bar{u} - \bar{v} - 10 = 2{,}293, s_X = s_{U-10-V} = s_{U-V} = 8{,}144$
- $t = \frac{2{,}293}{8{,}144}\sqrt{99} = 2{,}801$
- $t_{98}(0{,}05) = 1{,}9845 \qquad \Rightarrow$ zamítnout
- $p = \mathsf{P}(|T| \geq t) = 0{,}0061 \qquad (0{,}61\ \%)$
- 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 $X_i = U_i - V_i$
  - pozorování s $U_i = V_i$ (tj. $X_i = 0$) se zpravidla vynechají
  - Y – počet kladných znamének $X_i = U_i - V_i$
  - H_0 : rozdělení U a V jsou stejná, pak je nutně $Y \sim \mathrm{bi}(n, 1/2)$
  - H_0 zamítáme pro velká nebo malá Y:

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

  - H_0 zamítáme pro velká nebo malá Y (Yatesova korekce):

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

- **párový Wilcoxonův test**
  - nutné **symetrické** rozdělení $X_i = U_i - V_i$
  - vyloučíme případy $U_i = V_i$ (tj. $X_i = 0$)
  - určíme pořadí R_i^+ hodnot $|X_i| = |U_i - V_i|$
  - W součet pořadí, kde $U_i > V_i$ (tj. $X_i > 0$)

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

  - někdy pod odmocninou ještě oprava na výskyt shodných pořadí
  - někdy (nové verze NCSS) poněkud jinak s nulovými X_i

příklad rozdíly dvou metod učení nazpaměť:

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

- H_0 : populační medián rozdílů =0
- znaménkový test: $y = 5$; $n = 8$ $z = \frac{|5-8/2|-0{,}5}{\sqrt{8/4}} = 0{,}3536$; $p = 0{,}7237$
- 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

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

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

$$p = 0{,}1614$$

centrální limitní věta pro četnosti

- Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$. Potom pro velké n má průměr z nich rozdělení $\mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $\mathsf{N}\left(n\mu, n\sigma^2\right)$.
- absolutní četnost Y
  - Y – součet veličin s alternativním rozdělením
  - $Y \sim \mathrm{bi}(n, \pi)$, proto přibližně $Y \sim \mathsf{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $\hat{\pi} = Y/n$
  - $\hat{\pi}$ – průměr veličin s alternativním rozdělením
  - $\hat{\pi} \sim \mathsf{N}(\pi, \pi(1-\pi)/n)$

interval spolehlivosti pro podíl (1)

- populace: **podíl** π prvků s danou vlastností
- π – **pravděpodobnost**, že vlastnost má náhodně vybraný prvek
- výběr: **relativní četnost** $\hat{\pi} = \frac{Y}{n}$ ve výběru
- relativní četnost je průměr nula-jedničkové veličiny – pro velké n má přibližně normální rozdělení
- nula-jedničková veličina má rozptyl $\pi(1 - \pi)$
- relativní četnost (=průměr) má rozptyl $\frac{\pi(1-\pi)}{n}$

interval spolehlivosti pro podíl (2)

- střední chyba relativní četnosti = směrodatná odchylka relativní četnosti = odmocnina z rozptylu je tedy $\sqrt{\frac{\pi(1-\pi)}{n}}$
- pravděpodobnost π neznáme, odhadneme ji pomocí relativní četnosti $\hat{\pi} = \frac{Y}{n}$
- odtud je $100(1-\alpha)$% interval spolehlivosti pro π

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

- existuje přesnější (pracnější) postup

příklad: hody s hrací kostkou

- odhadujeme pravděpodobnost šestky, $\alpha = 0{,}05$
- kostka A: $n = 100, n_A = 17, \hat{\pi}_A = 0{,}17$

$$\left(0{,}17 - 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}; 0{,}17 + 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}\right) = (0{,}10; 0{,}24)$$

- kostka B: $n = 100, n_B = 41,_B = 0{,}41$

$$\left(0{,}41 - 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}; 0{,}41 + 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}\right) = (0{,}31; 0{,}51)$$

- důležitý rozdíl: u kostky A patří $1/6 = 0{,}167$ do 95% intervalu spolehlivosti; u kostky B nikoliv

pst výskytu jevu (binomické rozdělení) $Y \sim \text{bi}(n, \pi)$

- $\text{H}_0 : \pi = \pi_0$: $Z = \dfrac{Y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \dfrac{\hat{\pi} - \pi_0}{\text{S.E.}(\hat{\pi})} \sim \text{N}(0,1)$
- někdy s opravou na spojitost

$$Z = \frac{|Y - n\pi_0| - 0{,}5}{\sqrt{n\pi_0(1-\pi_0)}} \operatorname{sign}(Y - n\pi_0) \sim \text{N}(0,1)$$

- $\text{H}_1 : \pi \neq \pi_0$: zamítnout pokud $|Z| \geq z(\alpha/2)$
- $\text{H}_1 : \pi > \pi_0$: zamítnout pokud $Z \geq z(\alpha)$
- $\text{H}_1 : \pi < \pi_0$: zamítnout pokud $Z \leq -z(\alpha)$
- existují přesný postup, bez použití aproximace

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

kritický obor: velká hodnota Y (tj. velké $\hat{\pi}$ resp. velké Z)

$$z = \frac{33 - 50 \cdot 0{,}5}{\sqrt{50 \cdot 0{,}5 \cdot 0{,}5}} = 2{,}263 \qquad p = \mathsf{P}(Z \geq 2{,}263) = 0{,}0118$$

s opravou na spojitost (NCSS):

$$z = \frac{33 - 50 \cdot 0{,}5 - 0{,}5}{\sqrt{50 \cdot 0{,}5 \cdot 0{,}5}} = 2{,}121 \qquad p = \mathsf{P}(Z \geq 2{,}121) = 0{,}0169$$

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 (vážený průměr odhadů z jedn. výběrů)

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

- statistika (pro test hypotézy, že rozdělení X a Y jsou stejná)

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

- $\mathsf{H}_0 : \mu_X = \mu_Y$
  zamítnout ve prospěch alternativy
  - $\mathsf{H}_1 : \mu_X \neq \mu_Y$ když $|T| \geq t_{n_X+n_Y-2}(\alpha)$
  - $\mathsf{H}_1 : \mu_X > \mu_Y$ když $T > t_{n_X+n_Y-2}(2\alpha)$
  - $\mathsf{H}_1 : \mu_X < \mu_Y$ když $T < -t_{n_X+n_Y-2}(2\alpha)$
- Welchův test při pochybách o shodě rozptylů (modifikace T)
- shodu rozptylů lze ověřit např. F-testem nebo testem Leveneho

příklad **výšky** dětí (opět v [cm])

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

H_0: shodné populační průměry, H_1: neshodné

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

odhad S.E.$(\bar{X} - \bar{Y})$: $\sqrt{38{,}936\,\frac{15+12}{15\cdot 12}} = \sqrt{5{,}8404} = 2{,}4167$

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

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

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

- test **Mannův-Whitneyův** (dvouvýběrový Wilcoxonův)
  - nahraď pozorování jejich pořadími
  - dva nezávislé výběry rozsahu n_X, n_Y
  - spojitá rozdělení
  - H$_0$: rozdělení jsou stejná
  - za H$_0$ 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, méně vůči nestejné variabilitě

hoši			dívky				poř.
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

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

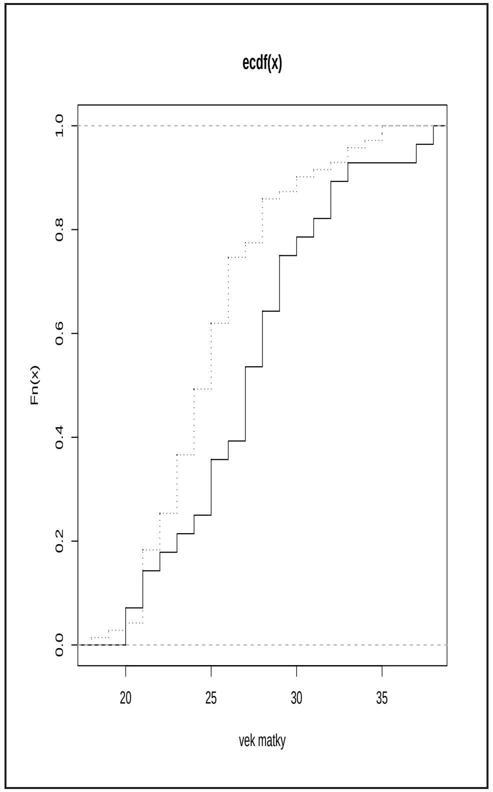
$w_y = 3 + 4 + 6 + 4 \cdot 16 + 19{,}5 + 21 + 2 \cdot 22{,}5 + 26{,}5 = 189$

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

$p = 0{,}3055$

přesně: $p = 0{,}3149$

- test **Kolmogorovův-Smirnovův**
  - porovná empirické distribuční funkce
  - citlivý vůči všem neshodám
  - věk matky, zda kojí v 24. týdnu
    $D = 0{,}354, \qquad p = 0{,}95$ %(NCSS)



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 4 takové, že rozdíl průměrů nejvýše –7,45 (3,17 %)
- při oboustranné alternativě další dvě kombinace, kdy rozdíl aspoň 7,45, což je 1,59 % permutací
- celkem $p = \dfrac{4+2}{126} = \dfrac{6}{126} = 0{,}0476$ (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
	...					...				
	*			*			*	*	–0,25	20
	...					...				
			*	*		*	*		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:

$$5, -1, 2, 3, -1, 4, 3, -3 \quad \text{průměr} = 1{,}5$$

- pokud jsou obě metody ekvivalentní, pak mají rozdíly náhodná znaménka; případnou nulu nutno předem vyloučit
- pro znaménka celkem $2^8 = 256$ možností
- ideál pro průměr 0
- v 27 případech průměr aspoň 1,5, v 27 případech nejvýše -1,5
-

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

	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**
				...						
23	−5	−1	−2	−3	1	4	−3	−3	**−1,50**	8,5
24	−5	1	−2	3	1	−4	−3	−3	**−1,50**	**8**
25	−5	1	−2	−3	−1	4	−3	−3	**−1,50**	8,5
26	−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**

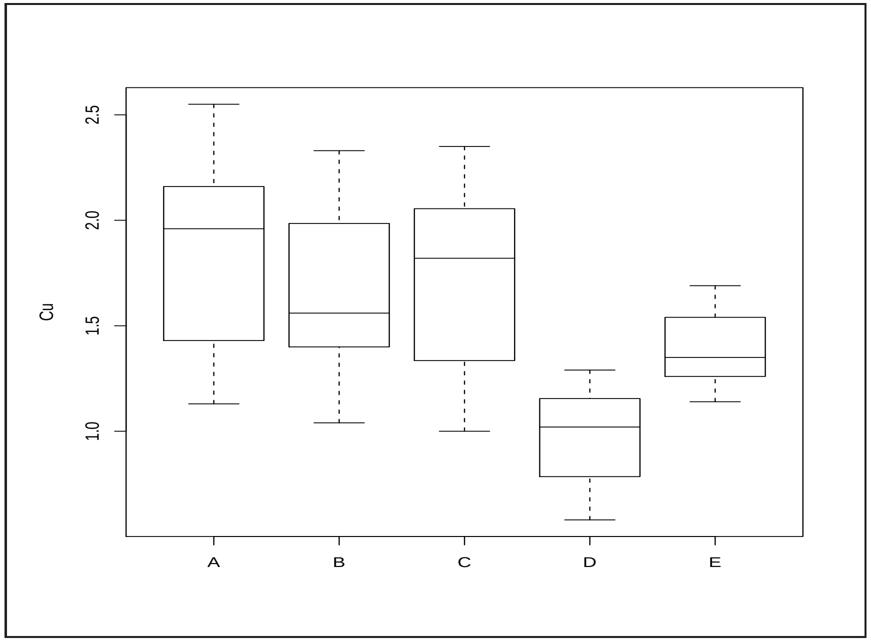
$$p_{\text{perm}} = \frac{27+27}{256} = 0{,}2109$$

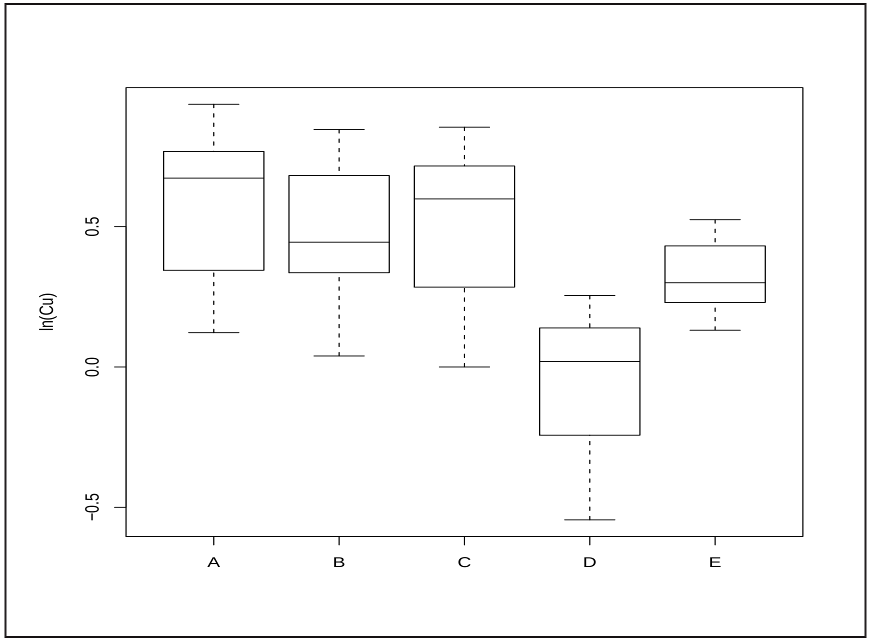
$$p_W = \frac{25+25}{256} = 0{,}1953$$

$$t = 1{,}5$$

$$p = 0{,}1773$$

příklad: játra pět míst na řece, vždy vyloveno po 7 rybách, zjišťována koncentrace mědi v játrech; liší se tato místa svým znečištěním? (logaritmování na pravé straně stabilizuje rozptyl)





analýza rozptylu jednoduchého třídění (ANOVA)

- $Y_{11}, \ldots, Y_{1n_1} \sim \mathsf{N}\Big(\mu_1, \sigma^2\Big)$
  $Y_{21}, \ldots, Y_{2n_2} \sim \mathsf{N}\Big(\mu_2, \sigma^2\Big)$
  ...
  $Y_{k1}, \ldots, Y_{kn_k} \sim \mathsf{N}\Big(\mu_k, \sigma^2\Big)$
- **nezávislé** výběry (shodné rozptyly, normální rozdělení)
- $\mathsf{H}_0 : \mu_1 = \ldots = \mu_k \quad (= \mu) \qquad \mathsf{H}_1$: neplatí H_0
- rozklad součtu čtverců

$$\sum\sum(Y_{it} - \bar{Y}_{\bullet\bullet})^2 = \sum n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum\sum(Y_{it} - \bar{Y}_{i\bullet})^2$$

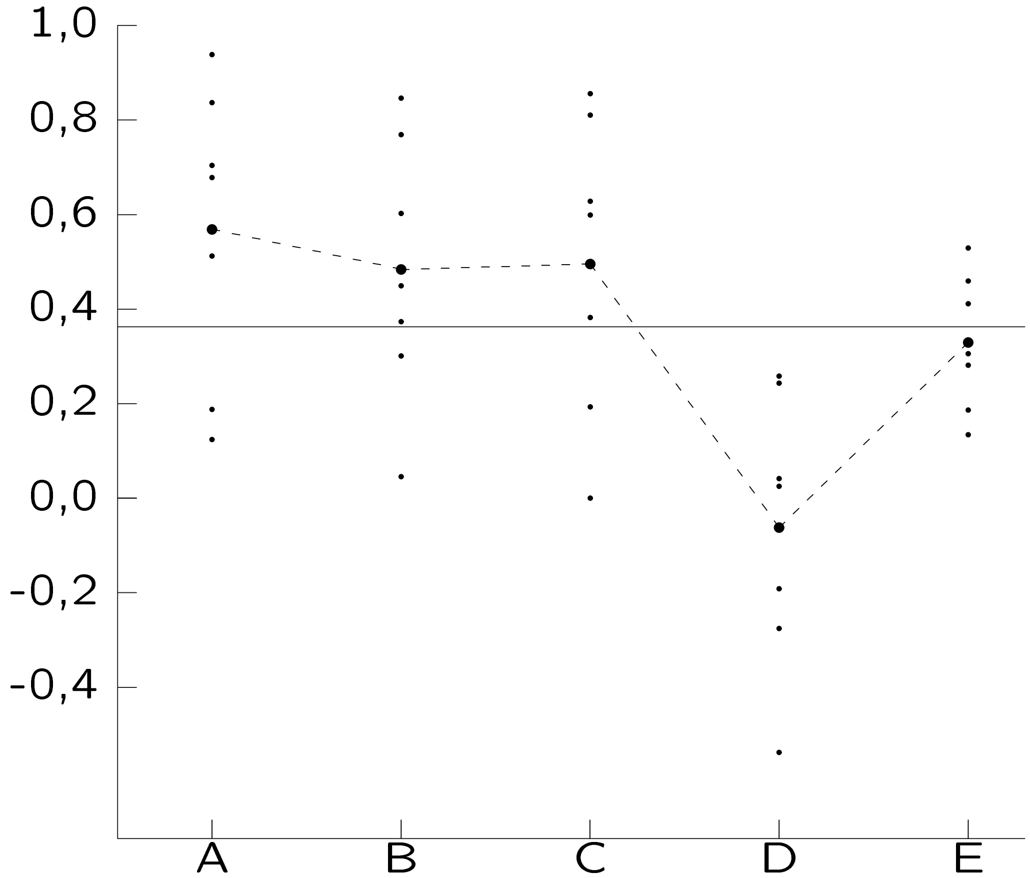
(celková variabilita) = (variabilita mezi) + (variabilita uvnitř)

$$S_T = S_A + S_e$$

$$f_T = f_A + f_e$$

$$(n-1) = (k-1) + (n-k)$$

- H_0 zamítnout, je-li $F_A = \dfrac{S_A/f_A}{S_e/f_e} \geq F_{f_A,f_e}(\alpha)$



tabulka analýzy rozptylu

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

příklad **játra**

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

$F = 5{,}862 > F_{4,30}(0{,}05) = 2{,}690$ na **5%** hladině jsme **prokázali** rozdíl

- **model** (měření = úroveň + chyba)

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

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

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

- $\mathsf{H}_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k$ (totéž, jako $\mu_1 = \mu_2 = \ldots = \mu_k$)
- pro $k = 2$ je $F = T^2$
- zobecnění úlohy dvouvýběrového t-testu na několik nezávislých výběrů

- **mnohonásobná srovnání** (které dvojice μ_i (resp. α_i) se liší?) (Tukeyův test, Kramerova verze)

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

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

  (nutnost zachovat zvolenou hladinu testu)

- **ověření** shody rozptylů
  - Leveneův test (vlastně ANOVA s $|Y_{it} - \text{med}_t Y_{it}|$)
  - Bartlettův test (citlivý na splnění předpokladu o normálním rozdělení)

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

$$q_{5,30}(0{,}05)\sqrt{\frac{0{,}0762}{2}\left(\frac{1}{7}+\frac{1}{7}\right)} = 4{,}10 \cdot 0{,}104 = 0{,}428$$

$-0{,}063 + 0{,}428 = 0{,}365 \Rightarrow$ na 5% hladině se místa D s nejmenším průměrem liší všechna místa s průměry aspoň 0,365, tedy místa A, B, C, nikoliv E

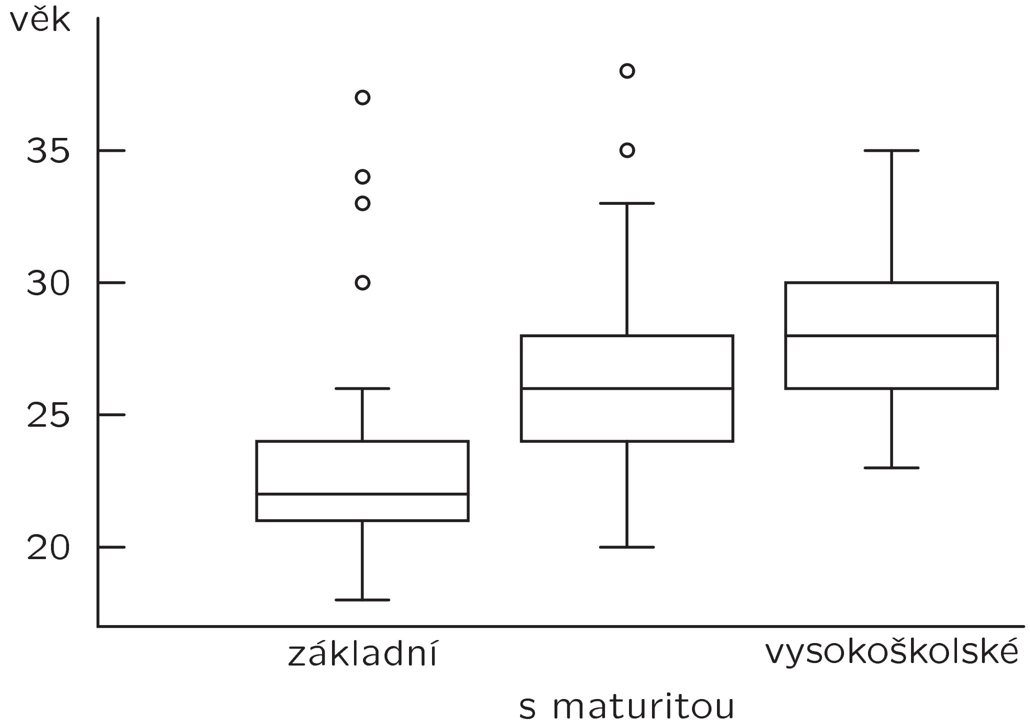
Kruskalův-Wallisův text

- 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)



vzdělání	n_i	prům. věk	stř. chyba	součet pořadí	prům. pořadí
základní	34	23,412	0,638	1025	30,15
maturita	47	26,278	0,543	2618	55,70
VŠ	18	28,500	0,877	1307	72,61
celk.	99	25,697		4 950	50

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

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

náhodné bloky

- zobecnění párových testů na r-tice
- **náhodný blok**
  - homogenní skupina objektů
  - počet objektů ve skupině = počet ošetření (nebo jeho násobek)
  - ošetření se přiřadí uvnitř bloku **náhodně** (každému ošetření stejný počet objektů)
- bloky – náhodné efekty $A_i \sim \mathsf{N}\left(0, \sigma_A^2\right)$
  ošetření – pevné efekty $\beta_j \qquad (\sum \beta_j = 0)$

$$Y_{ij} = \mu + A_i + \beta_j + E_{ij} \qquad E_{ij} \sim \mathsf{N}\left(0, \sigma^2\right)$$

aditivní vliv, symbolicky $A + B$

náhodné bloky

- testované hypotézy
  - H$_A : \sigma_A^2 = 0$ (nulová variabilita mezi bloky)
  - H$_B : \beta_1 = \ldots = \beta_r = 0$ (B nemá vliv)
- rozklad variability

$$S_T = S_A + S_B + S_e$$

- vliv dvou **faktorů**
  (A – náhodný, B – pevný)

příklad **diety**

- váhové přírůstky za danou dobu
  - $r = 4$ ošetření (pevné efekty)
  - $k = 5$ vrhů (náhodné efekty)

	dieta				
vrh	A	B	C	D	prům.
1	6,6	5,2	7,4	9,1	7,075
2	10,1	11,4	13,0	12,6	11,775
3	5,8	4,2	9,5	8,8	7,075
4	12,1	10,7	11,9	13,0	11,925
5	8,2	8,8	9,6	9,4	9,000
prům.	8,56	8,06	10,28	10,58	9,370

- tabulka ANOVA

variabilita	S	f	S/f	F	p
vrhy	91,932	4	22,983	22,26	<0,0001
dieta	23,332	3	7,774	7,53	0,0043
reziduální	12,388	12	1,032	-	-
celk.	127,642	19	-	-	-

- nesprávně jednoduché třídění ANOVA

  kdybychom zapomněli na závislost některých pozorování způsobenou náhodnými bloky (vrhy):

  $S_e = 91{,}932 + 12{,}388 = 104{,}320$, $f_e = 4 + 12 = 16$

$$F = \frac{23{,}332/3}{104{,}320/16} = 1{,}193, \qquad p = 0{,}344$$

Friedmanův test

- model $Y_{ij} = \mu + A_i + \beta_j + E_{ij}$ (náhodný řádkový efekt) nebo $Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$ (pevný řádkový efekt)
- E_{ij} nezávislé, spojité rozdělení
- $\mathrm{H}_0 : \beta_1 = \ldots = \beta_r$ (nezávisí na ošetření)
- urči pořadí v rámci každého bloku (řádku)
- za hypotézy je v každém řádku náhodná permutace čísel 1,. . . ,r, tedy součty ve sloupcích (pro ošetření) podobné
- $$Q = \frac{12}{kr(r+1)} \sum_{j=1}^{r} \left(\sum_{i=1}^{k} R_{ij} \right)^2 - 3k(r+1)$$
- zamítat H_0 : pro $Q \geq \chi^2_{r-1}(\alpha)$

příklad **diety**

vrh	dieta A	B	C	D	prům.
1	6,6	5,2	7,4	9,1	7,075
2	10,1	11,4	13,0	12,6	11,775
3	5,8	4,2	9,5	8,8	7,075
4	12,1	10,7	11,9	13,0	11,925
5	8,2	8,8	9,6	9,4	9,000
prům.	8,56	8,06	10,28	10,58	9,370

vrh	dieta A	B	C	D
1	2	1	3	4
2	1	2	4	3
3	2	1	4	3
4	3	1	2	4
5	1	2	4	3
součet	9	7	17	17

- $k = 5, r = 4$
- $Q = \frac{12}{5 \cdot 4 \cdot 5}\left(9^2 + 7^2 + 17^2 + 17^2\right) - 3 \cdot 5 \cdot 6 = 9{,}96$
- $Q > \chi^2_3(0{,}05) = 7{,}8147$, $p = 0{,}0189$

dvojné třídění s interakcemi

- vliv dvou faktorů nemusí být aditivní

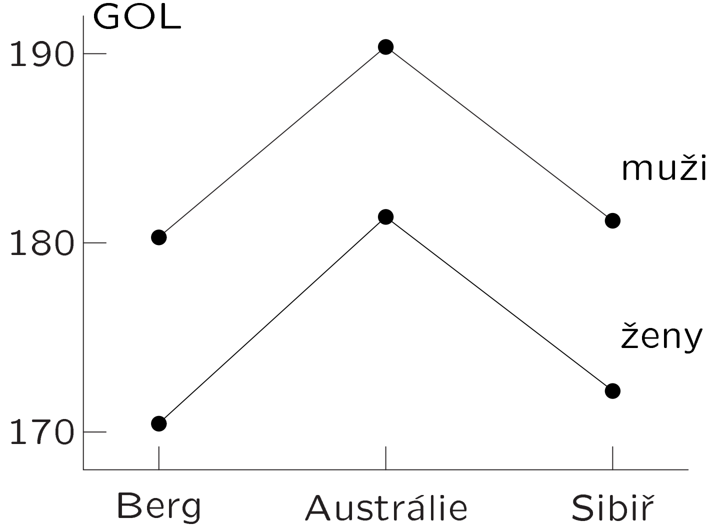
$$Y_{ijt} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijt}$$
$$E_{ijt} \sim \mathsf{N}\Big(0, \sigma^2\Big)$$

  - symbolicky $A + B + AB$
  - $\sum_i \alpha_i = 0$ **efekty** faktoru A odpovídající jeho k úrovním
  - $\sum_j \beta_j = 0$ **efekty** faktoru B odpovídající jeho r úrovním
  - $\sum_i \gamma_{ij} = 0, \quad \sum_j \gamma_{ij} = 0$ **interakce** vyjadřují neaditivitu obou faktorů (vliv A závisí na úrovni B, vliv B závisí na úrovni A)

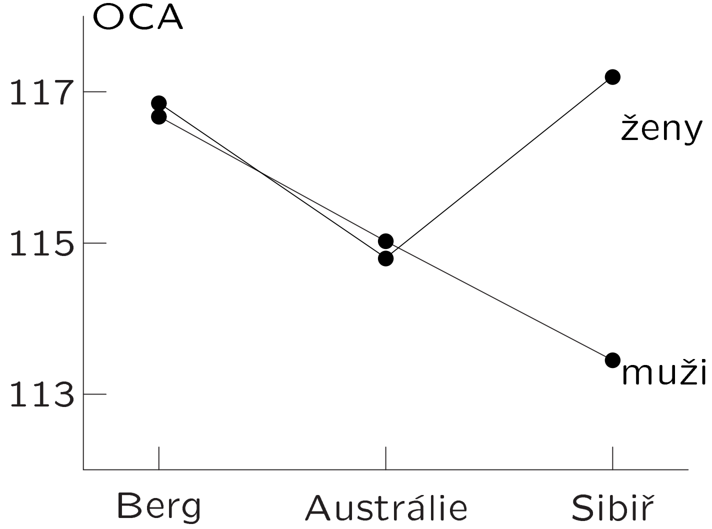
- testy
  - $\mathsf{H}_{AB}: \gamma_{ij} = 0$ (aditivita)
  - $\mathsf{H}_{A}: \alpha_{i} = 0$ (faktor A nemá vliv)
  - $\mathsf{H}_{B}: \beta_{j} = 0$ (faktor B nemá vliv)
  - pokud zamítneme H_{AB}, nemá smysl testovat $\mathsf{H}_{A}, \mathsf{H}_{B}$, neboť prostřednictvím interakcí oba faktory vliv mají; pak je lépr přejít k modelu jednoduchého třídění s kombinovanými úrovněmi

příklad **Howells**:
lebky exhumované na třech místech (A)
rozlišované podle pohlaví (B)

- největší délka mozkovny $(p_{AB} = 0{,}8872)$



- týlní úhel $(p_{AB} = 0{,}0222)$



příklad **Howells** největší délka mozkovny (GOL)

pohlaví	místo	n_{ij}	$\bar{y}_{ij}$	s_{ij}
M	Berg	40	180,300	7,293
F	Berg	40	170,450	6,641
M	Austrálie	40	190,375	5,555
F	Austrálie	40	181,375	6,632
M	Sibiř	40	181,175	6,468
F	Sibiř	40	172,175	5,228

tabulka ANOVA

var.	S	f	S/f	F	p
místa	5242,1	2	2621,1	65,2	$<$0,0001
pohl.	5170,8	1	5170,8	128,6	$<$0,0001
inter.	9,6	2	4,8	0,1	0,8872
rezid.	9410,6	234	40,2		
celk.	19833,2	239			

příklad **Howells** týlní úhel (OCA)

pohlaví	místo	n_{ij}	$\bar{y}_{ij}$	s_{ij}
M	Berg	40	116,675	5,567
F	Berg	40	116,850	5,682
M	Austrálie	40	115,025	4,382
F	Austrálie	40	114,800	4,286
M	Sibiř	40	113,450	4,782
F	Sibiř	40	117,200	4,973

tabulka ANOVA

var.	S	f	S/f	F	p
místa	150,908	2	75,454	3,05	0,0493
pohl.	91,267	1	91,267	3,69	0,0560
inter.	191,608	2	95,804	3,87	0,0222
rezid.	5789,550	234	24,742		
celk.	6223,333	239			

- **pevné** efekty
  - úrovně faktoru volí experimentátor
  - při opakovaném pokusu je lze zvolit stejně
  - vypovídáme o konkrétních úrovních faktoru
  - H_0: nulové efekty
- **náhodné** efekty
  - úrovně faktoru volí příroda
  - při opakovaném pokusu jsou jiné
  - vypovídáme o populaci možných úrovní faktoru
  - H_0: nulová variabilita efektu
- testy obecně závisí na charakteru efektu
- doporučují se **vyvážené** modely

porovnání populačních měr polohy (přehled)

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 jedn. 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} (zaveden na obr. 66)
  - $|\rho_{XY}| \leq 1$; pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy}

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

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_X} \right) \left(\frac{Y_i - \bar{Y}}{S_Y} \right) \qquad (z\text{-skóry})$$

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

- k prokázání závislosti nutno **normální** rozdělení X, Y
- $\mathrm{H}_0 : \rho_{XY} = 0$ se na hladině α zamítá:

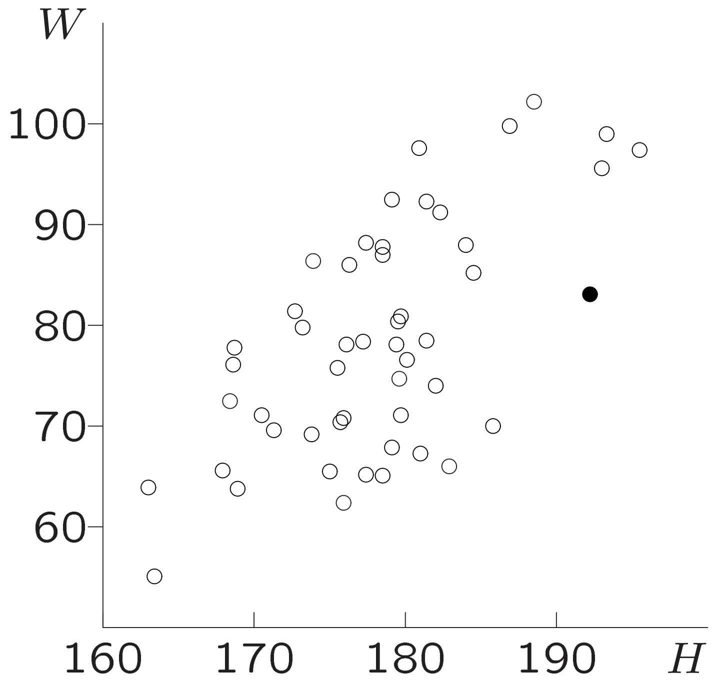
$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}, \qquad |T| \geq t_{n-2}(\alpha)$$

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

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

  - H_0: (nezávislost) se zamítá, je-li $|r_{XY}^{(S)}\sqrt{n-1}| \geq z(\alpha/2)$

příklad **tuk** (v závorce po vynechání označeného pozorování)



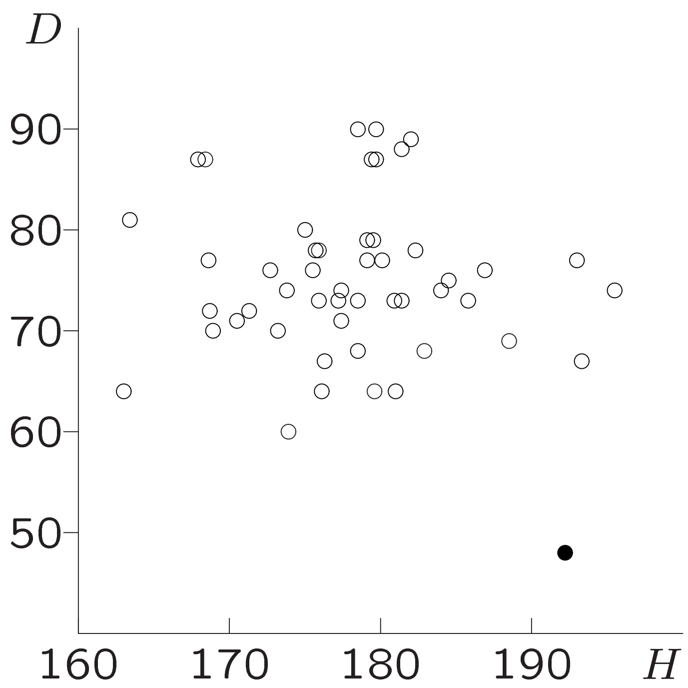
výška vers. hmotnost

$$r = 0{,}643 \qquad (0{,}654)$$

$$t = 5{,}814 \qquad (5{,}921)$$

$$p < 0{,}0001 \qquad (p < 0{,}0001)$$

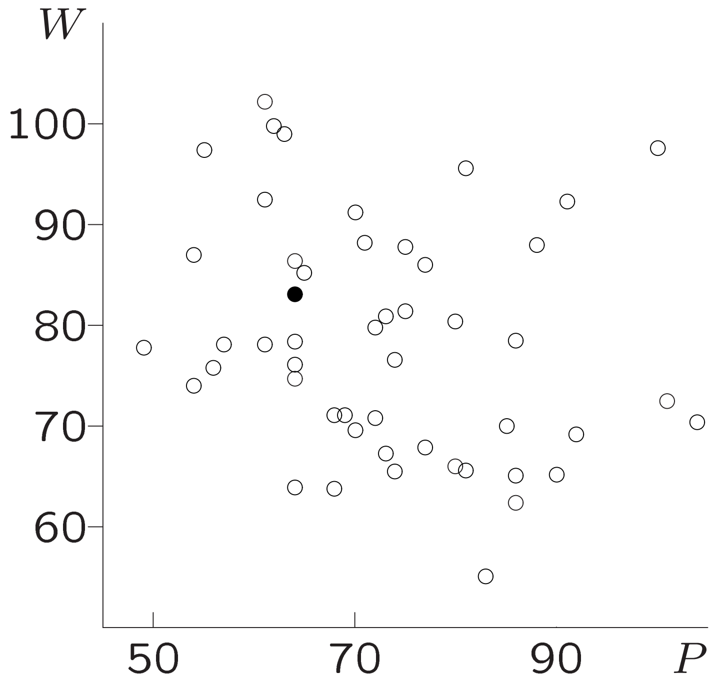
příklad **tuk** (v závorce po vynechání označeného pozorování)



výška vers. diast. tlak

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

příklad **tuk** (v závorce po vynechání označeného pozorování)



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 (přibližší chování r normálnímu rozdělení)

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

- příklad **děti**: porodní délka, hmotnost
  - dívky: $r_1 = 0{,}5687$, $n_1 = 51$, $z_1 = \dfrac{1}{2}\ln\dfrac{1+0{,}5687}{1-0{,}5687} = 0{,}6456$
  - hoši: $r_2 = 0{,}5967$, $n_2 = 49$, $z_2 = 0{,}6880$
  - test shody (odhady r_1, r_2 jsou **nazávislé!**)

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

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

– 95% interval spolehlivosti pro $\zeta = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ u dívek

$$\left(0{,}6456 - \frac{1{,}960}{\sqrt{51-3}}\,;\,0{,}6456 + \frac{1{,}960}{\sqrt{51-3}}\right) = (0{,}363; 0{,}929)$$

– 95% interval spolehlivosti pro ρ_1 inverzní transformací

$$\rho = \frac{\mathrm{e}^{2\zeta}-1}{\mathrm{e}^{2\zeta}+1}$$

tedy

$$\left(\frac{\mathrm{e}^{2\cdot 0{,}363}-1}{\mathrm{e}^{2\cdot 0{,}363}+1}\,;\,\frac{\mathrm{e}^{2\cdot 0{,}929}-1}{\mathrm{e}^{2\cdot 0{,}929}+1}\right) = (0{,}348; 0{,}730)$$

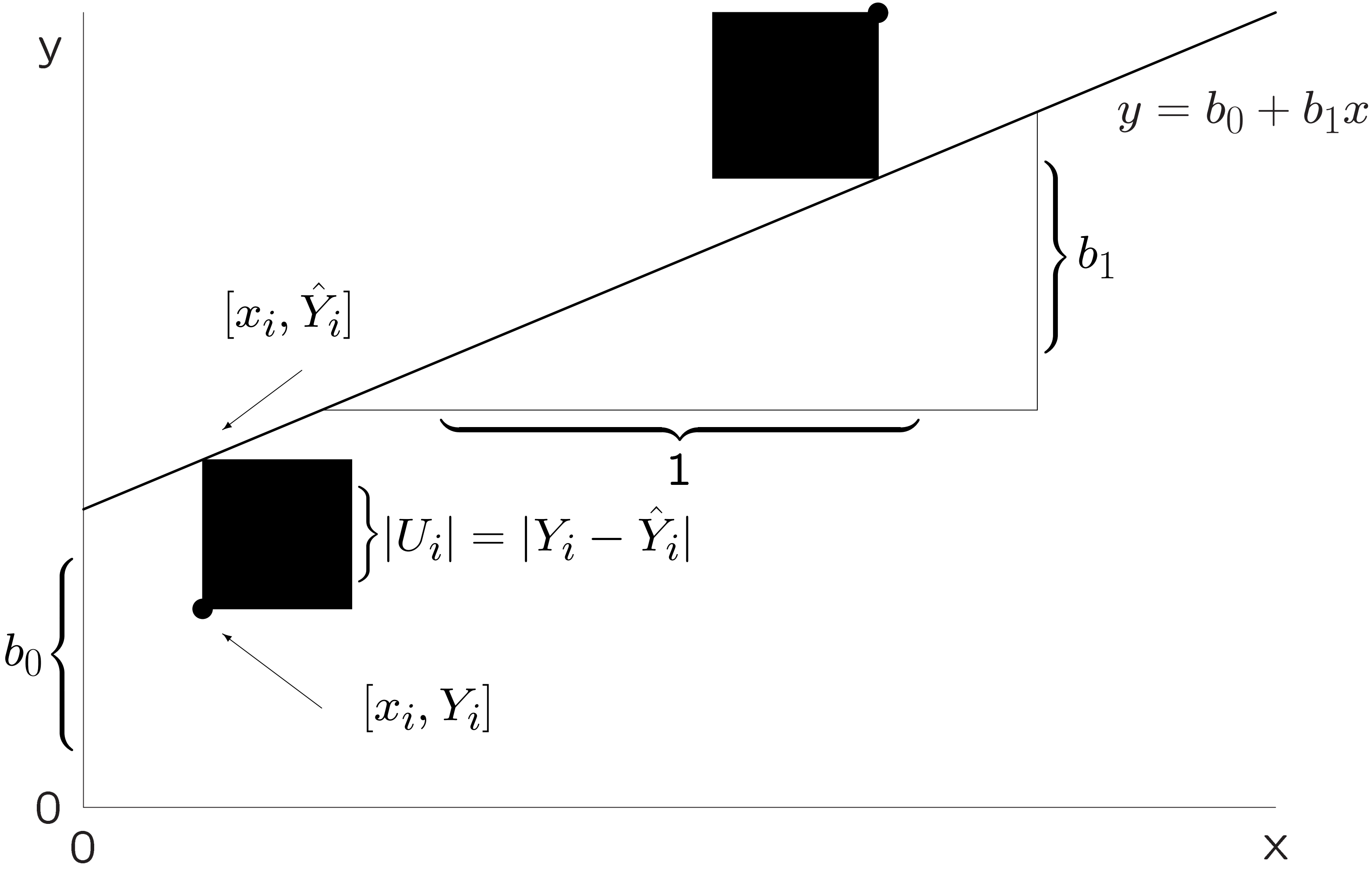
regrese (původ pojmu)

- tendence (návrat) k průměrnosti (F. Galton (1886) vyšetřoval dědičnost výšky postavy)
- 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 $$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$


y
$y = b_0 + b_1 x$
b_1
$[x_i, \hat{Y}_i]$
1
$|U_i| = |Y_i - \hat{Y}_i|$
b_0
$[x_i, Y_i]$
0
0
x

- 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, zda má smysl předpovídat pomocí regrese)

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

- nezávislost $\mathsf{E}\,Y$ na x znamená $\mathsf{H}_0 : \beta_1 = 0$ (zda je závislost průkazná)

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

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 $\widehat{\text{FAT}} = -53{,}870 + 0{,}379 \cdot \text{HEIGHT}$

(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)

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

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

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

- $\mathsf{H}_0 : \beta_1 = \beta_2 = 0$ (nezávisí 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

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

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

χ^2-testy

- pro znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale bez ohledu na uspořádání
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním

- **multinomické** rozdělení
  - v dílčím pokusu k možných výsledků $A_1, \ldots, A_k$ (neslučitelné, sjednocení všech je jev jistý)
  - π_j je pst, že vyjde A_j $\qquad (\sum \pi_j = 1)$
  - n **nezávislých** dílčích pokusů
  - N_j – počet dílčích pokusů, kdy nastalo A_j
  - $(N_1, \ldots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$
  - samotné N_j má binomické rozdělení, tj. $\boxed{N_j \sim \mathrm{bi}\big(n, \pi_j\big)}$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

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

- 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 χ^2_{k-1}
- **test shody** $\mathrm{H}_0 : \pi_1 = \pi_1^0, \dots, \pi_k = \pi_k^0$
  (pravděpodobnosti hypotézou dány **jednoznačně**)
  - platí-li H_0, očekáváme četnosti blízké hodnotám $\mathsf{E}\, N_j = n\pi_j^0$:

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

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

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

složená nulová hypotéza

- hypotéza, která určuje vztahy mezi pravděpodobnostmi (hypotéza o struktuře), některé parametry zůstávají volné,je třeba je odhadnout
- příklad antigen: souvisí výskyt antigenu A s jistou nemocí?
- četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$
- model pro fenotypy AA, Aa, aa (neurčený parametr θ))

$$\mathsf{P}(AA) \equiv \pi_1(\theta) = \theta^2$$
$$\mathsf{P}(Aa) \equiv \pi_2(\theta) = 2\theta(1-\theta)$$
$$\mathsf{P}(aa) \equiv \pi_3(\theta) = (1-\theta)^2$$

- odhad θ minimalizací *logaritmické věrohodnostní funkce*

$$\begin{aligned}
\ell(\theta) &= \ln(\mathsf{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3)) \\
&= \ln\left(c_1 \left(\theta^2\right)^{n_1} (2\theta(1-\theta))^{n_2} \left((1-\theta)^2\right)^{n_3}\right) \\
&= c_2 + (2n_1 + n_2)\ln\theta + (n_2 + 2n_3)\ln(1-\theta) \\
\hat{\theta} &= \frac{1}{2} + \frac{N_1 - N_3}{2n} \qquad \left(= 0{,}5 + \frac{18-6}{82} = 0{,}646\right)
\end{aligned}$$

- zamítat pokud (q počet složek θ)

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi(\hat{\theta}))^2}{n\pi(\hat{\theta})} \geq \chi^2_{k-1-q}(\alpha)$$

- antigen: $\chi^2 = 0{,}355, \quad p = 0{,}551$

kontingenční tabulka

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

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

- **nezávislost** znaků: pro všechny dvojice i, j platí

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

- teoretické četnosti (protějšek N_{ij})

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

- 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 $c = 4$, $n = 6800$
- $o_{11} = 2811 \cdot 2829/6800 = 1169, \quad o_{12} = 2811 \cdot 2632/6800 = 1088 \ldots$
- $o_{34} = 116 \cdot 857/6800 = 14{,}62 \geq 5$

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

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 první měření B_i a druhé měření 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 musí být 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

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

- $\chi^2_3(0{,}05) = 7{,}8147$, $\quad \boldsymbol{p} = 0{,}3597$
- 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 $X^2 \geq \chi_1^2(\alpha) = z(\alpha/2)^2$

- **Yatesova korekce** $X_Y^2 = \dfrac{n(|ad - bc| - n/2)^2}{(a+c)(b+d)(a+b)(c+d)}$
- **Fisherův faktoriálový (exaktní) test**
  - počítá přímo dosaženou hladinu p
  - malé četnosti nevadí

příklad **hraboš**

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

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

$$\chi^2 = \frac{515(4 \cdot 473 - 11 \cdot 27)^2}{15 \cdot 500 \cdot 31 \cdot 484} = 11{,}643, \qquad p = 0{,}0006$$

- **ale**: $15 \cdot 31/515 = 0{,}9 < 5$
- **Yates**: $\chi^2 = 8{,}187 \qquad p = 0{,}0042$

- **Fisherův test**: $p = 0{,}0092$
- na 5% hladině závislost **prokázána**
- (zcela jiná otázka) **vyskytují se cizopasníci se stejnou pstí?** McNemarův test:

$$\chi^2 = \frac{(11-27)^2}{11+27} = 6{,}7368, \qquad p = 0{,}0094$$

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
- zvol rozsah výběru (přesnost, délka int. spolehlivosti, síla testu)
- 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 korel. koeficient)
  - tvar rozdělení (šikmost, špičatost)
- **prokázat vliv ošetření**
  - změna polohy (t-testy, analýza rozptylu)
  - změna variability (Levene, F-test, Bartlettův test)
  - jiná změna (Kolmogorov-Smirnov)

jaké úlohy řešíme

- **prokázat závislost**
  - obě spojité (korelační koeficient)
  - spojitá na kvalitativními (ANOVA)
  - obě kvalitativní (kontingenční tabulka)
- **popsat závislost**
  spojité veličiny na spojitých či kvalitativních – 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 (rozptylový 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 abychom prokázali svoji vědeckou hypotézu
- 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

některé další modely a metody

- **diskriminační analýza**
  - na každém objektu měříme několik spojitých veličin
  - známe příslušnost objektů ke skupinám
  - DA dá rozhodovací pravidlo pro přiřazování dalších objektů do skupin
  - například podle kosterních nálezů určovat pohlaví
- **shluková analýza**
  - na každém objektu měříme několik spojitých veličin
  - konstruujeme skupiny navzájem blízkých (podobných) objektů
  - vzniklé skupiny se snažíme interpretovat

příklad z **archeologie** (Thurzo 1979)

- trojí pohřebiště (avarsko-slovanská, slovanská, maďarská)
- měříme šířku tváře (zy-zy) a míru 8a (sagitální průměr středu diafýzy tibie), uvažujeme ženy
- průměry:

pohřebiště	rozsah	šířka	míra 8a
slovanské	39	122,410	25,615
maďarské	27	127,963	30,471

- odhad varianční matice

$$\mathsf{S} = \begin{pmatrix} 25{,}631 & -0{,}724 \\ -0{,}724 & 6{,}937 \end{pmatrix}$$

- korelační koeficient $r = -0{,}054$
- t-testy: $t_1 = -4{,}381$, $t_2 = -7{,}380$

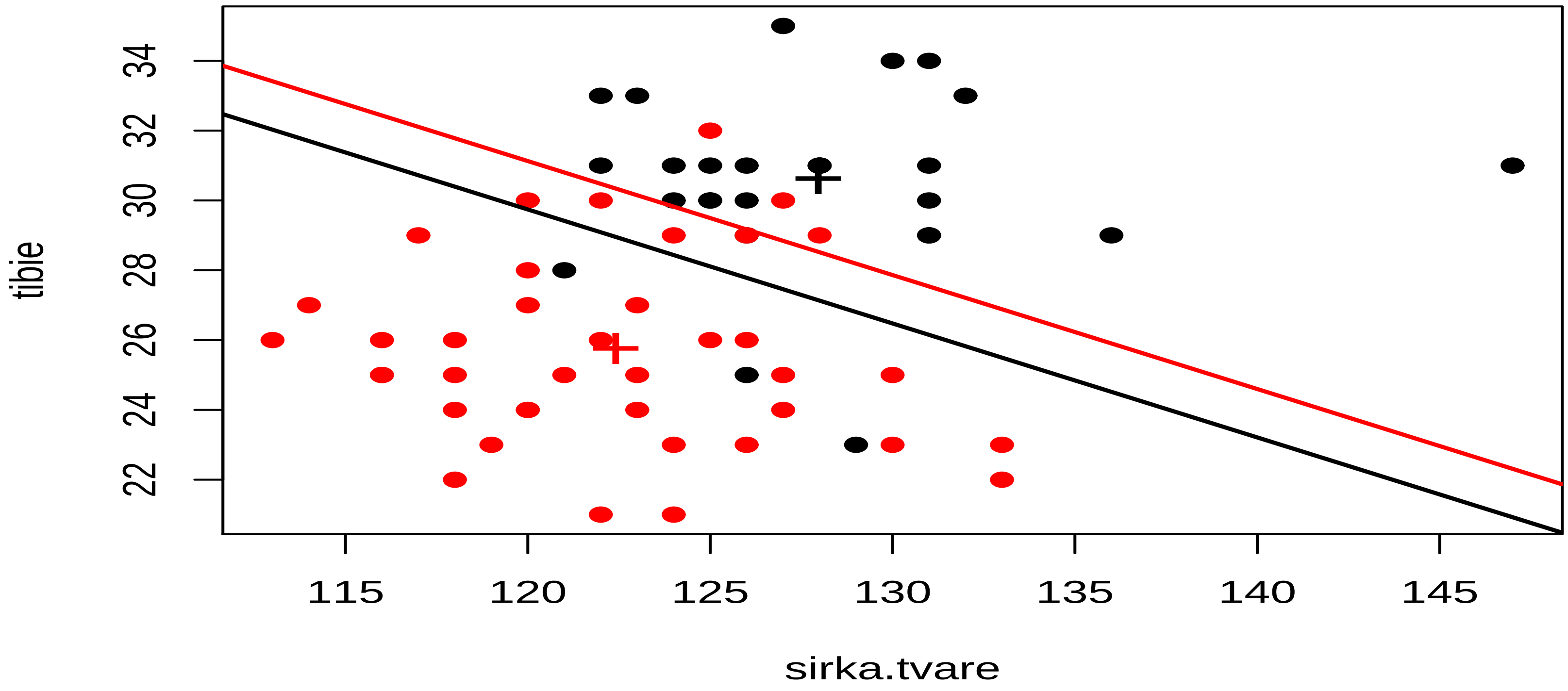
rozhodovací pravidlo (DA)

- rozhodujeme mezi dvěma pohřebišti
- stejné psti obou populací
- ke slovanským přiřaď, když $\boxed{0{,}237 \text{ šířka} + 0{,}726 \text{ míra 8a} < 50{,}060}$
- k maďarským když $\boxed{0{,}237 \text{ šířka} + 0{,}726 \text{ míra 8a} > 50{,}060}$
- špatně zařazeno:
  - pouze 7 z 39 slovanských (17,9 %)
  - pouze 3 z 27 maďarských (11,1 %)
- při očekávaném poměru 4:1 ve prospěch slovanské populace bude ke slovanským pohřebištím přiřazena žena, když

$$0{,}237 \text{ šířka} + 0{,}726 \text{ míra 8a} < 50{,}060 + \ln\left(\frac{4}{1}\right) = 51{,}446$$

slov, mad, + značí těžiště

poměr 1:1, poměr 4:1

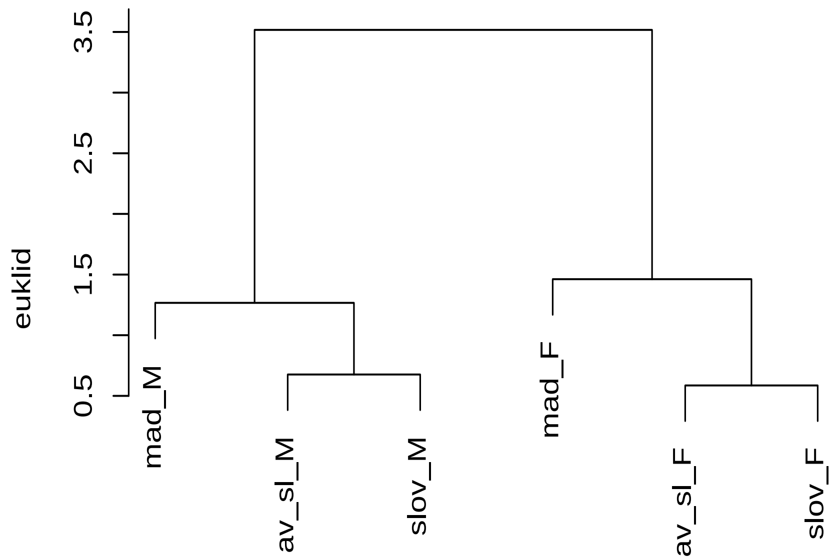


rozlišení pohřebišť (shluky)

- každé pohřebiště a pohlaví charakterizujeme průměrnou hodnotou čtyř veličin (ještě výška a délka lebky (g-op)
- pro těchto šest čtveřic se spočítá **vzdálenost**
- postupně se vytvářejí skupinky nejbližších, pak jejich vzdálenost
- grafické znázornění – **dendrogram**
- vzdálenost (nepodobnost)
  - euklidovská
  - Mahalanobisova (uváží závislosti)
  - 1-korelační koeficient
- vzdálenost skupin: těžiště / nejbližší prvky / nejvzdálenější prvky

vzdálenost skupin = vzdálenost nejvzdálenějších prvků

Cluster Dendrogram



Cluster Dendrogram