Základy biostatistiky

(MD710P09)
ak. rok 2008/2009

Karel Zvára

karel.zvara@mff.cuni.cz
http://www.karlin.mff.cuni.cz/∼zvara
katedra pravděpodobnosti a matematické statistiky MFF UK

(naposledy upraveno 26. května 2009)

- **cvičení na počítačích v B5**
  - od úterka 24. února ve Viničné 7, 1. patro B5
  - nutno **zapsat se do paralelky** prostřednictvím SIS
  - zápočet za aktivní účast (+ odevzdávání souborů/písemky)
  - nutno mít aktivní účet v učebnách, znát svoje heslo
  - volně šiřitelný program R (http://cran.r-project.org/)
- **zkouška v B5**
  - jen se zápočtem, přihlašování prostřednictvím SIS
  - kombinace písemného a ústního zkoušení
  - řešení úloh na počítači
  - základy teorie (pojmy, metody a jejich volba, interpretace)
- **literatura**
  - K. Zvára: Biostatistika. Karolinum 1998,. . . , 2008
  - internetová stránka http://www.karlin.mff.cuni.cz/∼zvara
- **konzultace** pondělí 13:15-14:00 v pracovně, II. patro K234, Sokolovská 83, Karlín (případně po předchozí dohodě jindy, možno také v ÚAMVT, Albertov 6)

tři části přednášky

- popisná statistika
  - několika čísly vystihnout důležitou vlastnost
  - jednoduchým grafem vyjádřit důležitou vlastnost
  - porovnat soubory dat
- abstraktní pohled (teorie)
  - pravděpodobnost, Bayesův vzorec, náhodná veličina, distribuční funkce, nezávislost
  - populace a výběr, populační parametr a jeho odhad
  - interval spolehlivosti pro parametr
  - test statistické hypotézy
- některé metody (modely)
  - testy o jednom, dvou či několika výběrech
  - rozhodování o závislosti kvantitativních či kvalitativních znaků

statistika

- **statistika**
  - **popisná** (deskriptivní):
    data stručně popsat, něco z dat „vydolovat“
    tvrdit něco o daných datech, nezobecňovat
  - **induktivní** (konfirmatorní):
    tvrdit něco nového, zobecnit na větší soubor,
    důležitá je interpretace
- **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 **statistických jednotkách** (osoba, obec, stát, pokusné pole, rostlinka pšenice, třetí list rostlinky pšenice, . . .)
- 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 (vyšetřování závislosti)
- měříme na skupinách jednotek – **souborech**
- zajímají nás **hromadné** vlastnosti, které charakterizují celou velkou skupinu (**populaci**)
- hodnoty znaků zjišťujeme u jedinců, nechceme vypovídat pouze o jednotlivcích
- kolik procent mužů kouří, ne, zda kouří Karel Zvára

měřítka

- **nula-jedničkové**
  pouze dvě možné hodnoty (muž/žena, kouří/nekouří)
- **nominální**
  seznam všech rozlišitelných hodnot, **faktor** (porodnice, pohlaví, odrůda)
- **ordinální**
  hodnoty nominálního měřítka jsou uspořádány, **uspořádaný faktor** (vzdělání matky, stupeň bolesti)
- **intervalové**
  stejné vzdálenosti sousedních hodnot (rok narození)
  „**o kolik** je x menší než y?“ (nikoliv „kolikrát“)
- **poměrové**
  srovnání se zvolenou jednotkou (hmotnost, výška, věk)
  „**kolikrát** je x větší, než y?“

hrubší dělení měřítek
(důležitější, bezprostředně ovlivní volbu metod)

- **kvalitativní**
  nula-jedničkové, nominální, často i ordinální
- u kvalitativních se zpravidla udávají **četnosti** jednotlivých hodnot (kolikrát která hodnota nastala)
- **kvantitativní** (spojité)
  intervalové, poměrové, někdy ordinální (ale není spojité)
- hodnoty kvantitativních – čísla
- pro četnosti hodnot v kvalitativním měřítku se používají zpravidla jiné charakteristiky a metody, než pro hodnoty v kvantitativním měřítku

veličina, statistika

- číselně vyjádřený výsledek měření, pokusu
- možné hodnoty znaků v intervalovém nebo poměrovém měřítku jsou hustě rozmístěné – **spojitá veličina**
- četnosti hodnot znaků v nula-jedničkovém, nominálním (či ordinálním) měřítku – **diskrétní veličina**
- u veličin používáme číselné charakteristiky některých hromadných vlastností (**charakteristiky polohy, charakteristiky variability**, charakteristiky tvaru)
- **statistika** (další význam) – funkce pozorovaných hodnot např. průměrná teplota nebo nejvyšší teplota v roce; číselně charakterizuje důležitou vlastnost veličiny (veličin)

označení

rozlišujte n, n_i m, x_i, x_i^*

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

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

$$\boxed{\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_m}{n} \quad \text{- \textbf{relativní} četnosti}}$$

$$\boxed{N_j = \sum_{i=1}^{j} n_i \quad \text{\textbf{kumulativní} četnosti}}$$

pro kumulativní četnosti nutno aspoň ordinální měřítko

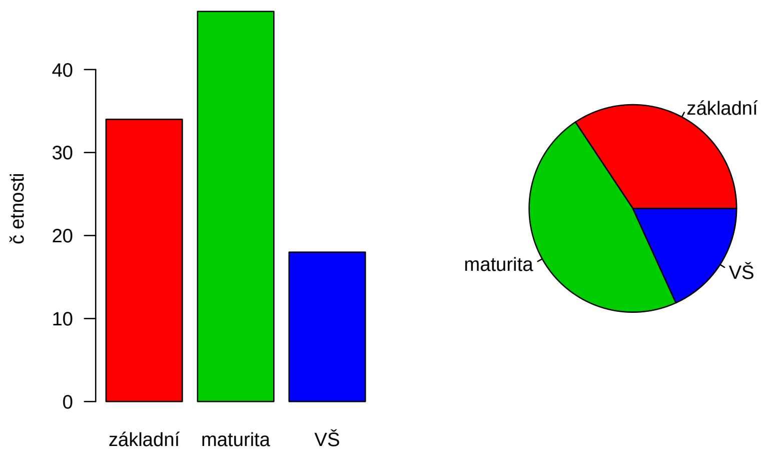
příklad **kojení** (vzdělání 99 matek)

ordinální měřítko se třemi hodnotami

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.

histogram (barplot u kvalitativní veličiny)

- **histogram**
  grafické znázornění intervalových četností spojité veličiny
- **barplot**
  grafické znázorněné četností (počtů hodnot) kvalitativního znaku
- plocha (výška) obdélníku úměrná četnosti
- relativní četnosti mají jen jiné měřítko svislé osy
- **výsečový diagram** pro relativní četnosti kvalitativního znaku (podíly nějakého celku)



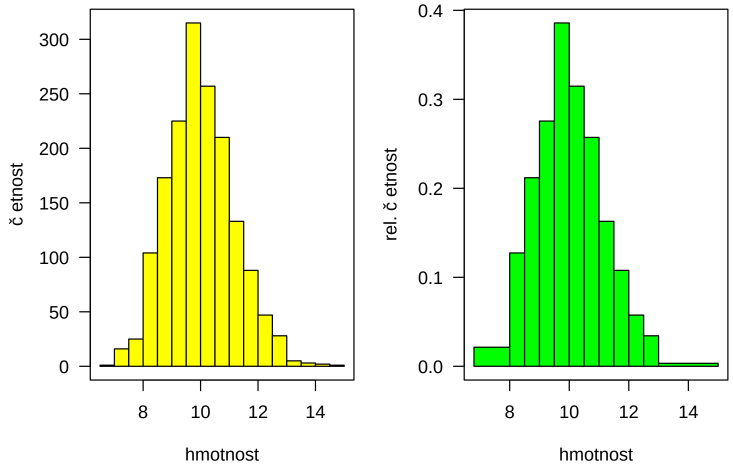
histogram u **spojité** veličiny

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 ve 12. měsíci (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

histogram pro hmotnost v jednom roce

histogram napravo podle třídních četností udává relativní četnosti



variační řada, pořadí

nutno rozlišovat x_i a $x_{(i)}$

- původní hodnoty spojité veličiny (kvantitativní znak)

  $x_1, x_2, \ldots, x_n$ např. 7, 4, 5, 4, 2

- **variační řada** [sort(x)]

  $x_{(1)} \le x_{(2)} \ldots \le x_{(n)}$ např. 2, 4, 4, 5, 7

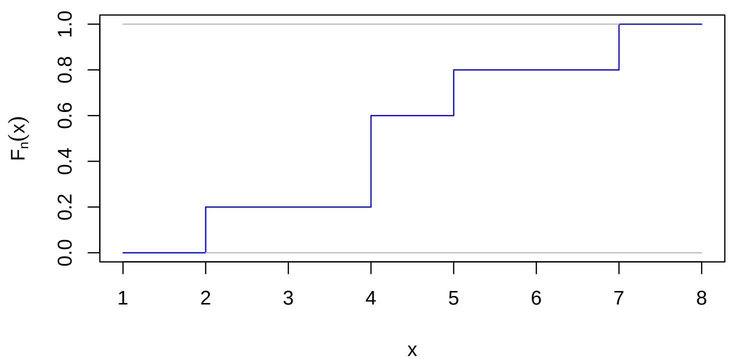
- **pořadí**: [rank(x)]
  na které místo ve variační řadě se dostane daná hodnota
  nejmenší dostane pořadí 1, druhá nejmenší dostane 2, ...
- je-li několik hodnot stejných, dostanou průměr z odpovídajících pořadí
- pořadí hodnot 7, 4, 5, 4, 2 jsou po řadě 5, 2,5, 4, 2,5, 1

empirická distribuční funkce

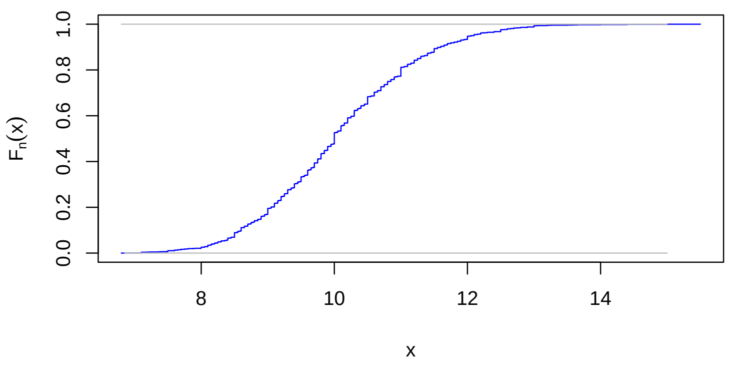
[empirical distribution function]

relativní četnost hodnot, které jsou nejvýše x $\boxed{F_n(x) = \frac{\#(x_i \le x)}{n}}$

naše variační řada: 2, 4, 4, 5, 7



empirická distribuční funkce



- příklad: váha dětí v jednom roce
- připomíná hladkou neklesající funkci

průměry

- **průměr** [mean(x)]

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

- **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^*

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

[weighted.mean(x, w)]

příklad: vážený průměr známek vážený kredity

jaký je nevážený průměr?

známka	kreditů	součin
x_j^*	w_j	$x_j^* \cdot w_j$
1	6	6
2	4	8
2	2	4
3	4	12
celkem	16	30

$$\bar{x} = \frac{6 \cdot 1 + 4 \cdot 2 + 2 \cdot 2 + 4 \cdot 3}{6 + 4 + 2 + 4} = \frac{30}{16} = 1{,}875$$

[weighted.mean(x=c(1,2,2,3),w=c(6,4,2,4))]

další míry polohy

opět jsou důležité závorky kolem indexů

- **medián** (**pro**střední hodnota, NIKOLIV *střední hodnota*)

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

[median(x)]

- **minimum, maximum**

$$x_{\min} = x_{(1)} \qquad \text{[min(x)]}$$

$$x_{\max} = x_{(n)} \qquad \text{[max(x)]}$$

[range(x)] spočítá dvojici $(x_{\min}, x_{\max})$

- **variační průměr** [mean(range(x))]

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

kvartily, decily

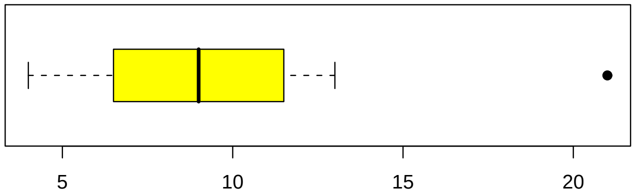
- **medián** $\tilde{x}$ je číslo, které dělí data na dvě poloviny: hodnot menších nebo stejných jako medián – hodnot větších nebo stejných jako medián [median(x)] [quantile(x,probs=1/2)]
- **dolní kvartil** Q_1 je číslo, které oddělí čtvrtinu hodnot (menších či stejných jako Q_1) od tří čtvrtin hodnot (větších či stejných jako Q_1) [quantile(x,probs=1/4)]
- **horní kvartil** Q_3 je číslo, které oddělí tři čtvrtiny hodnot (menších či stejných jako Q_3) od čtvrtiny hodnot (větších či stejných jako Q_3) [quantile(x,probs=3/4)]
- **první decil** je číslo, které oddělí desetinu nejmenších hodnot od ostatních hodnot [quantile(x,probs=1/10)]
- **percentil** x_p je číslo, které oddělí $100p$ % nejmenších hodnot od ostatních hodnot [quantile(x,probs=p)]
- několik percentilů současně [quantile(x,probs=(0:4)/4)]

výpočet percentilu x_p

jeden z nejčastěji užívaných postupů, též v R

- najde se celé číslo k splňující
$$\frac{k-1}{n-1} \le p < \frac{k}{n-1}$$
- tedy $k = \lfloor 1 + (n-1) \cdot p \rfloor$
($\lfloor x \rfloor$ znamená celou část z x, zaokrouhlí dolů)
- provede se lineární interpolace mezi $x_{(k)}$ a $x_{(k+1)}$
($\{x\}$ znamená zlomkovou část x, o kolik přesahuje celé číslo)
$$q = \{1 + (n-1) \cdot p\} = (1 + (n-1) \cdot p) - k$$
$$x_p = (1-q) \cdot x_{(k)} + q \cdot x_{(k+1)}$$
- např. pro $n = 99, p = 0{,}25$ bude
$$k = \lfloor 1 + (99-1) \cdot 0{,}25 \rfloor = \lfloor 25{,}5 \rfloor = 25, \quad q = 25{,}5 - 25 = 0{,}5$$
$$Q_1 = x_{0{,}25} = 0{,}5 \cdot x_{(25)} + 0{,}5 \cdot x_{(26)}$$

krabicový diagram



[boxplot(c(4,5,8,9,10,13,21),horizontal=TRUE,col=7,pch=16)]
znázorněna řada statistik pro data: 4, 5, 8, 9, 10, 13, 21

- medián ($\tilde{x} = 9$) – příčka obdélníka
- kvartily ($Q_1 = 6{,}5$, $Q_3 = 11{,}5$) – kratší strany obdélníka
- tykadla od kvartilu k minimu (maximu), pokud není odlehlé
- odlehlé pozorování – je dál, než $3/2 \cdot (Q_3 - Q_1)$ ($= 7{,}5$) od bližšího kvartilu

vlastnosti míry polohy

- přičteme-li ke každé hodnotě x stejnou konstantu a, musíme tutéž konstantu a přičíst k průměru (mediánu, kvartilu, ...)
- vynásobíme-li každou hodnotu x stejnou kladnou konstantou b, musíme průměr (medián, kvartil, ...) vynásobit totéž konstantou b
- pro dobrou míru polohy $\mu(X)$ platí:
$$\mu(a + X) = a + \mu(X)$$
$$\mu(b \cdot X) = b \cdot \mu(X) \qquad (b > 0)$$
- dobrá míra polohy je citlivá vůči posunutí i vůči změně měřítka, zachová je

míry variability

- míra variability $\sigma(x)$ číselně charakterizuje jinou vlastnost, než míry polohy, proto na poloze nesmí záviset
- ukazuje nakolik jsou zjištěné hodnoty nestejné, velikost jejich kolísání, jejich **variabilitu**
- pro dobrou míru variability $\sigma(X)$ platí:

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

- přičtení konstanty a míru variability nezmění, na vynásobení kladnou konstantou b reaguje

směrodatná odchylka, rozptyl

- **rozptyl** (variance, $s^2_{b \cdot x} = b^2 s^2_x$) $\boxed{s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

  [var(x)]
- např. pro data: 4, 5, 8, 9, 10, 13, 21 dostaneme $\bar{x} = 10$, tedy

$$s_x^2 = \frac{1}{7-1}\left((4-10)^2 + (5-10)^2 + \ldots + (21-10)^2\right) = \frac{196}{6}$$

- **směrodatná odchylka** [standard deviation]

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

další míry variability

- **rozpětí** $R = x_{\max} - x_{\min}$ [range]
- **kvartilové rozpětí** $\boxed{R_Q = Q_3 - Q_1}$ [interquartile range]
- **variační koeficient** (nesplňuje ani jeden požadavek) slouží k porovnání variability při různých úrovních [coefficient of variation]

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

- **entropie** (pro nominální, požadavky nemají smysl, nezávisí na označení hodnot, jen na jejich relativních četnostech)

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

příklad ICHS: vztah mužů ke kouření

	vztah ke kouření							
vzděl.	nekuřák/bývalý		střední		silný		celk.	H
zákl.	25	21,4 %	14	12,0 %	78	66,7 %	117	0,854
odb.	83	28,0 %	24	8,1 %	189	63,9 %	296	0,847
stř.	99	33,2 %	24	8,1 %	175	63, %	298	0,882
VŠ	115	48,3 %	17	7,1 %	106	44,5 %	238	0,900

muži se základním vzděláním:

$$H = -\left(\frac{25}{117}\ln\frac{25}{117} + \frac{14}{117}\ln\frac{14}{117} + \frac{78}{117}\ln\frac{78}{117}\right) = 0{,}854123$$

větší vyrovnanost $\Rightarrow$ větší entropie

z-skóry

- z-**skóry** (normovaná veličina)

$$\boxed{z_i = \frac{x_i - \bar{x}}{s_x}, \qquad i = 1, 2, \ldots, n}$$ [(x-mean(x))/sd(x)]

- hodnoty $z_1, z_2, \ldots, z_n$ „ztratily“ informaci o poloze a variabilitě, vždy platí $\bar{z} = 0, \quad s_z = 1$
- přičtení konstanty ani násobení konstantou z-skóry nezmění
- hodnocení vlastností nezávislých na poloze a variabilitě
- pro data: 4, 5, 8, 9, 10, 13, 21 platí $\bar{x} = 10, s_x = 5{,}715$
- proto dostaneme

$$z_1 = \frac{4 - 10}{5{,}715} = -1{,}050, \ldots, z_7 = \frac{21 - 10}{5{,}715} = 1{,}925$$

šikmost, špičatost

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

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

[mean(((x-mean(x))/sd(x))^3)]

- **špičatost** (průměr 4. mocnin z-skórů zmenšený o 3)

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

[mean(((x-mean(x))/sd(x))^4)-3]

- g_1, g_2 se používají k posouzení normality
- pro data: 4, 5, 8, 9, 10, 13, 21 dostaneme

$$g_1 = 0{,}771 \qquad g_2 = -0{,}770$$

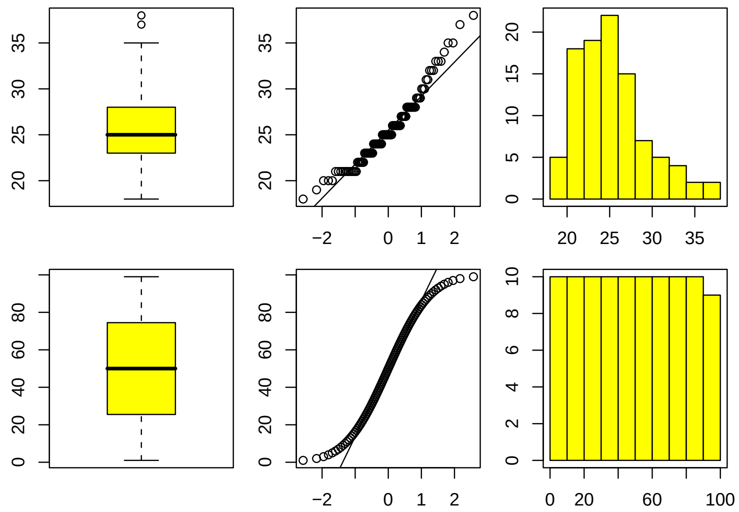
normální diagram

[(normal) probability plot], [quantile-comparison plot]

- k ověřování předpokladu **normálního** rozdělení
- porovnává skutečnou variační řadu s ideální řadou normálního (Gaussova) rozdělení
- v ideálním případě body téměř na přímce
- systematická odchylka ukazuje na rozdělení, které není normální
- konvexní či konkávní průběh – nesymetrie (nenulová šikmost)
- esovitý průběh – nenulová špičatost
- [qqnorm(x)]
- přímku vloží [qqline(x)]

příklad: věk matky, čísla 1 až 99

věk matek: $g_1 = 0{,}741, \quad g_2 = 0{,}220$ čísla 1 až 99: $g_1 = 0, \quad g_2 = -1{,}236$



závislost dvojice znaků

- možnost zkoumání závislosti dvou znaků
- způsob znázornění (prokazování) závisí na měřítcích znaků
- **kvantitativní – kvantitativní**
  rozptylový (bodový) diagram [scatter plot]
  korelace, regrese [correlation, regression]
- **kvantitativní - kvalitativní**
  krabicový diagram [box-plot]
  t-test, ANOVA
- **kvalitativní - kvalitativní**
  kontingenční tabulka [contingency table]
  chí-kvadrát test, Fisherův exaktní test

kvantitativní – kvantitativní

- pokud záleží na směru závislosti, pak vysvětlovanou (**závisle proměnnou**) veličinu umístíme na svislou osu y
- **korelační koeficient** vyjadřuje sílu a směr **vzájemné** závislosti

$$\boxed{r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}}, \quad \text{kde} \quad s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- [cor(x,y)] [correlation coefficient]
- s_{xy} – výběrová **kovariance** [covariance]
- pomocí z-skórů (nezávislost na poloze a měřítku)

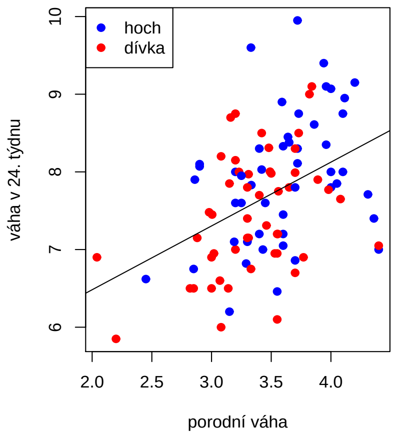
$$\boxed{r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}$$

- pro $r_{xy} > 0$ s rostoucím x v průměru roste y
  pro $r_{xy} < 0$ s rostoucím x v průměru klesá y

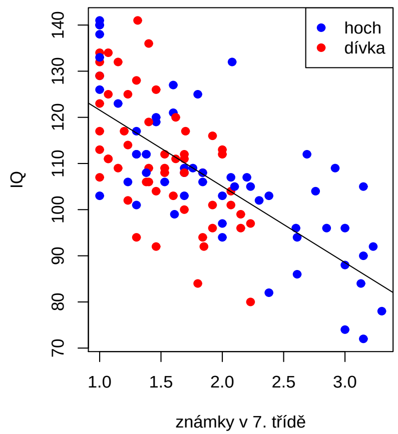
$$\boxed{-1 \leq r_{xy} \leq 1}$$

kvantitativní – kvantitativní, příklady

vlevo – závislost váhy v 24. týdnu na porodní váze s rozlišením pohlaví (data: Kojení)
vpravo – závislost IQ na průměrné známce v 7. třídě (data: Iq3)



$r = 0{,}429$



$r = -0{,}689$

kvalitativní – kvalitativní

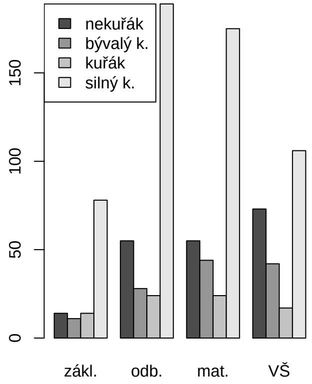
- **kontingenční tabulka** [contingency table]
  obsahuje přehledně zapsané úplné údaje
- sdružené četnosti jednotlivých kombinací hodnot dvou znaků
- marginální četnosti:
  - **řádkové** marginální četnosti: součty sdružených četností v jednotlivých řádcích (pro jednotlivé hodnoty řádkového znaku)
  - **sloupcové** marginální četnosti: součty sdružených četností v jednotlivých sloupcích (pro jednotlivé hodnoty sloupcového znaku)
- [table(F,G)] nebo [xtabs(∼ F + G)]
  resp. [xtabs(∼ F + G , data=DataFrame)]
  kde F a G jsou v R faktory, DataFrame je databáze

příklad: kouření u mužů

data: Ichs

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

v grafu znázorněny **absolutní** četnosti (sdružené, marginální četnosti)
[barplot(t,beside=TRUE)]



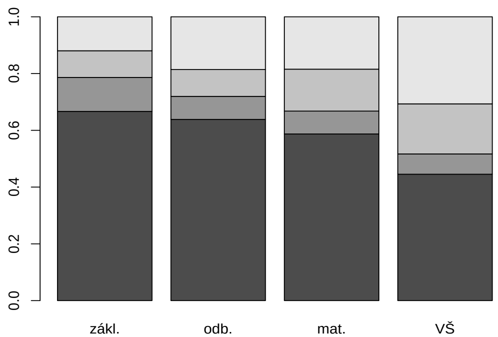
relativní četnosti v kontingenční tabulce

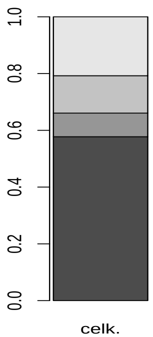
- **řádková procenta** (relativní četnosti v daném řádku)
  - podíl jednotlivých hodnot sloupcového znaku pro danou hodnotu řádkového znaku
  - **podmíněné rozdělení** hodnot sloupcového znaku pro danou hodnotu řádkového znaku
- **sloupcová procenta** (relativní četnosti v daném sloupci)
  - podíl jednotlivých hodnot řádkového znaku pro danou hodnotu sloupcového znaku
  - **podmíněné rozdělení** hodnot řádkového znaku pro danou hodnotu sloupcového znaku
- **nezávislosti** obou znaků odpovídá situace, kdy jsou např. sloupcová procenta pro všechny hodnoty sloupcového znaku podobné

příklad: kouření u mužů

podmíněné relativní četnosti marginální relativní četnosti

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

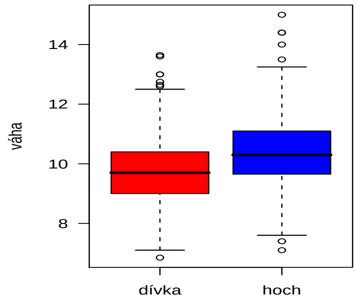


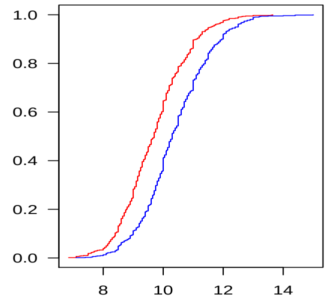


kvantitativní – kvalitativní

váha v jednom roce podle pohlaví, data: Deti1633

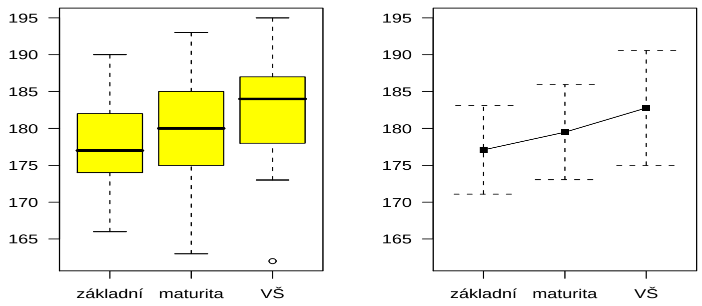
- lze chápat jako závislost spojité veličiny na kvalitativní
- srovnání souborů dat (spojitá veličina)
- krabicové diagramy resp. empirické distribuční funkce
- příklad: hmotnost chlapců a dívek v jednom roce
- **nezávislosti** odpovídá podobné umístění krabic resp. empirických distribučních funkcí





příklad: závislost výšky otce na vzdělání matky
data: Kojení

- porovnáme výšky otců ve skupinách podle vzdělání matky
- napravo znázorníme průměry a směrodatné odchylky
- intervaly kolem průměru mívají i jinou interpretaci (jsou jiné)

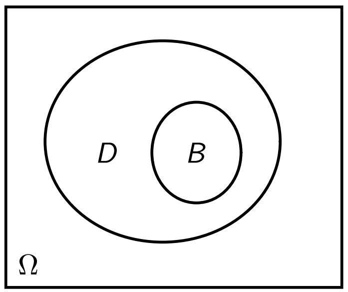


Náhodné jevy

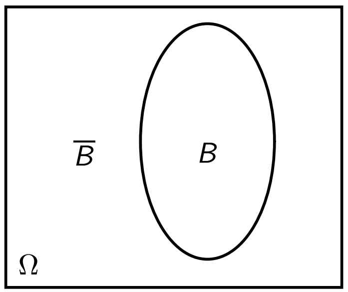
- **náhodný pokus** výsledek předem neurčitý
- předpokládá se **stabilita relativních četností** možných výsledků, která s nezávislými opakováními pokusu roste
- **náhodný jev** tvrzení o výsledku náhodného pokusu (podmnožina množiny Ω)
- **jistý jev** Ω nastává vždy
- **nemožný jev** $\emptyset$ nenastává 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** $B \cap D = \emptyset$

znázornění pomocí Vennova diagramu
celý obdélník – jev jistý

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



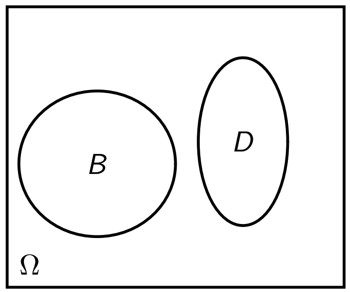
$\mathrm{P}(\overline{B}) = 1 - \mathrm{P}(B)$



velikost plochy odpovídá pravděpodobnosti

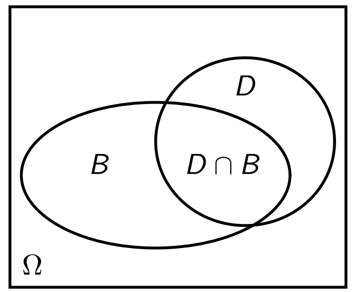
$B \cap D = \emptyset \Rightarrow$

$\mathrm{P}(B \cup D) = \mathrm{P}(B) + \mathrm{P}(D)$



obecně platí

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



velikost plochy odpovídá pravděpodobnosti

pravděpodobnost

- objektivní číselné vyjádření „naděje“, že nastane jev B
- modelový protějšek relativní četnosti
- pravděpodobnost (pst) by měla mít stejné vlastnosti jako relativní četnost:
  - $\boxed{0 \leq \mathrm{P}(B) \leq 1}$
  - $\boxed{\mathrm{P}(\Omega) = 1, \mathrm{P}(\emptyset) = 0}$
  - $\boxed{B \cap D = \emptyset \Rightarrow \mathrm{P}(B \cup D) = \mathrm{P}(B) + \mathrm{P}(D)}$ **(sčítání pravděpodobností)**
  - $\boxed{\mathrm{P}(B \cup D) = \mathrm{P}(B) + \mathrm{P}(D) - \mathrm{P}(B \cap D)}$
  - $\boxed{B \subset D \Rightarrow \mathrm{P}(B) \leq \mathrm{P}(D)}$
  - $\boxed{\mathrm{P}(\overline{B}) = 1 - \mathrm{P}(B)}$

- **klasická definice psti**
  - m **stejně pravděpodobných** elementárních jevů $\omega_1, \ldots, \omega_m$
  - jsou neslučitelné, sjednocení všech je jistý jev
  - m_B elementárních jevů **příznivých jevu** B (tj. takových ω_i, že $\omega_i \in B$, je právě m_B)

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

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

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

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

hypergeometrické rozdělení

příklad na klasickou pravděpodobnost

- kombinační číslo $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$
- v rybníku je N ryb (zpravidla neznámý počet) A ryb vylovíme, označíme a vypustíme zpět
- po nějaké době vylovíme n ryb, z nich Y je označených
- s jakou pravděpodobností je $Y = k$?
  - celkem $\binom{N}{n}$ možných n-tic vylovených ryb
  - k označených lze vybrat $\binom{A}{k}$ způsoby
  - $n - k$ neoznačených lze vybrat $\binom{N-A}{n-k}$ způsoby

$$\mathrm{P}(Y = k) = \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}}, \qquad \max(0, n+A-N) \leq k \leq \min(A, n)$$

- např. odhad neznámého N: $\hat{N} = n\frac{A}{Y}$ (neboť $Y/n \doteq A/N$)

příklad rodina

(nejmladší, prostřední, nejstarší)

tři sourozenci, celkem 8 elementárních jevů $\omega_1, \ldots, \omega_8$

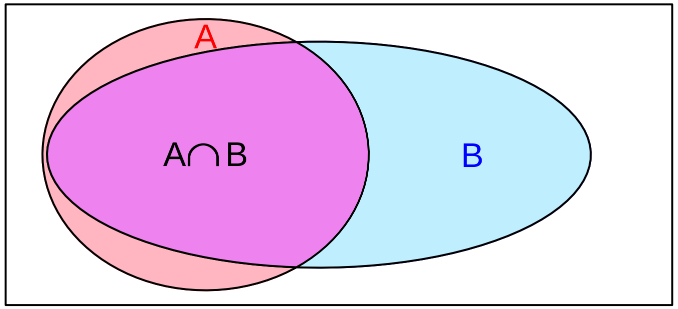
ω_i	D	B	$B \cap D$	$B \cup D$	C
(m, m, m)					+
(f, m, m)	+	+	+	+	+
(m, f, m)		+		+	+
(f, f, m)	+			+	+
(f, f, f)	+			+	
(m, f, f)					
(f, m, f)	+			+	
(m, m, f)		+		+	

D nejmladší je dívka, $\mathrm{P}(D) = 4/8 = 1/2$
B v rodině je jediná dívka, $\mathrm{P}(B) = 3/8$
$B \cap D$ jediná dívka je nejmladší, $\mathrm{P}(B \cap D) = 1/8$
$\mathrm{P}(B \cup D) = \mathrm{P}(B) + \mathrm{P}(D) - \mathrm{P}(B \cap D) = \frac{3}{8} + \frac{4}{8} - \frac{1}{8} = \frac{6}{8}$
C nejstarší je hoch, $\mathrm{P}(C) = 4/8 = 1/2$

podmíněná pravděpodobnost

když víme, že nastalo A (je to jisté, pst A za podmínky A je rovna 1), pak **podmíněná pst** jevu B za podmínky A bude rovna relativní velikosti $B \cap A$ vzhledem k velikosti A

$$\boxed{\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}}$$



příklad rodina

B – jediná dívka, D nejmladší je dívka

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

- jaká je pst jevu B (v rodině je jediná dívka) **když víme**, že platí D (nejmladší je dívka)?

$$\mathrm{P}(B|D) = \frac{m_{B\cap D}}{m_D} = \frac{1/8}{4/8} = \frac{1}{4} = \frac{2}{8}$$

- jaká je pst jevu B (v rodině je jediná dívka) **když víme**, že platí $\overline{D}$ (nejml. **není** dívka)?

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

$$\mathrm{P}(B|D) = \frac{2}{8} < \mathrm{P}(B) = \frac{3}{8} < \mathrm{P}(B|\overline{D}) = \frac{4}{8}$$

jev B tedy **závisí** na jevu D!

nezávislost náhodných jevů

ω_i	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)		

- když víme, že nejstarší je hoch (C), jaká je pak pst, že nejmladší je dívka (D)?
- dva ze čtyř elementárních jevů, tedy

$$\boxed{\frac{m_{D\cap C}}{m_C} = \frac{2/8}{4/8} = \frac{2}{4} = \frac{4}{8} = \frac{m_D}{m}}$$

- zde pst jevu D **nezávisí** na tom, zda platí C
- **nezávislost**: pst jevu D nezávisí na tom, zda C nastal či nenastal

můžeme upravit na **definici nezávislosti náhodných jevů**

$$\frac{m_{D\cap C}}{m} = \frac{m_D}{m}\frac{m_C}{m}$$

$$\boxed{\mathrm{P}(D \cap C) = \mathrm{P}(D)\mathrm{P}(C)}$$

vlastnosti podmíněné pravděpodobnost a nezávislosti

- pravděpodobnost jevu D za podmínky jevu C

$$\boxed{\mathrm{P}(D|C) = \frac{m_{D\cap C}}{m_C} = \frac{m_{D\cap C}/m}{m_C/m} = \frac{\mathrm{P}(D \cap C)}{\mathrm{P}(C)}}$$

- pravděpodobnost průniku jevů D, C obecně

$$\mathrm{P}(D \cap C) = \mathrm{P}(D|C)\mathrm{P}(C)$$
$$\mathrm{P}(C \cap D) = \mathrm{P}(C|D)\mathrm{P}(D)$$

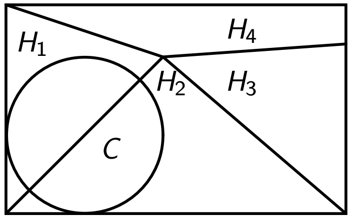
(ale $\mathrm{P}(C \cap D) = \mathrm{P}(D \cap C)$, neboť $D \cap C = D \cap C$)

- pro **nezávislé jevy** platí (**násobení pstí**)

$$\boxed{\mathrm{P}(D \cap C) = \mathrm{P}(D)\mathrm{P}(C)}$$

vzorec pro úplnou pst, Bayesův vzorec

počítáme $P(H_1|C)$, např. C – správná odpověď, H_j – správná známka j



$$P(H_1) = 0{,}231$$
$$P(H_2) = 0{,}375$$
$$P(H_3) = 0{,}219$$
$$P(H_4) = 0{,}175$$
$$P(C|H_1) = 0{,}589$$
$$P(C|H_2) = 0{,}362 \quad (\text{proč je } P(C|H_2) < P(C|H_1)?)$$

$$P(C) = P(C \cap H_1) + P(C \cap H_2)$$
$$P(C \cap H_1) = P(C|H_1)P(H_1), \quad P(C \cap H_2) = P(C|H_2)P(H_2)$$
$$P(H_1 \cap C) = P(H_1|C)P(C)$$
$$P(H_1|C) = \frac{P(H_1 \cap C)}{P(C)} = \frac{P(C|H_1)P(H_1)}{P(C|H_1)P(H_1) + P(C|H_2)P(H_2)} = \frac{1}{2}$$

obecný vzorec pro úplnou pravděpodobnost

(totéž, ale obecně))

- $H_1, \dots, H_k$ neslučitelné (tj. $H_i \cap H_j = \emptyset$ pro $i \neq j$)
- sjednocení $H_1, \dots, H_k$ dá jev jistý (tj. $H_1 \cup \dots \cup H_k = \Omega$)

z definice podmíněné psti plyne $P(C \cap H_j) = P(C|H_j) \cdot P(H_j)$

$$\begin{aligned} P(C) &= P(C \cap \Omega) = P(C \cap (H_1 \cup H_2 \cup \dots \cup H_k)) \\ &= P((C \cap H_1) \cup (C \cap H_2) \cup \dots \cup (C \cap H_k)) \text{ (neslučitelné jevy)} \\ &= P(C \cap H_1) + P(C \cap H_2) + \dots + P(C \cap H_k) \\ &= P(C|H_1)P(H_1) + P(C|H_2)P(H_2) + \dots + P(C|H_k)P(H_k) \end{aligned}$$

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

$P(C)$ je váženým průměrem podmíněných pstí $P(C|H_j)$

Bayesův vzorec [Bayes formula]

stejné předpoklady: H_j neslučitelné, sjednocení všech jistý jev

$$P(H_i|C) = \frac{P(H_i \cap C)}{P(C)}, \qquad P(C|H_i) = \frac{P(C \cap H_i)}{P(H_i)}$$

odtud je pro libovolně zvolené i

$$P(H_i \cap C) = P(C \cap H_i) = P(C|H_i)P(H_i)$$

proto pro každé i, $i = 1, \dots, k$ platí

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

$H_1, \dots, H_k$ – **hypotézy**, $P(H_1|C), \dots, P(H_k|C)$ – **aposteriorní** psti
$P(H_1), \dots, P(H_k)$ – **apriorní** psti (nutně $P(H_1) + \dots + P(H_k) = 1$)

příklad: zkoušení

H_j – student si zaslouží známku j, učitel studenta (tedy j) nezná
C – student správně odpoví na položenou otázku
$P(H_j)$ – apriorní představa učitele o neznámém studentovi
$P(C|H_j)$ – obtížnost otázky, volí učitel

H_j	$P(H_j)$	$P(C\|H_j)$	$P(C\|H_j)P(H_j)$	$P(H_j\|C)$	$P(H_j\|C_2)$	$P(H_j\|C_3)$
1	0,20	1,00	0,2000	0,2694	0,3451	0,4230
2	0,35	0,80	0,2800	0,3771	0,3865	0,3790
3	0,25	0,65	0,1625	0,2189	0,1822	0,1452
4	0,20	0,50	0,1000	0,1347	0,0863	0,0529
Σ	1,00		0,7425	1,0000	1,0000	1,0000

$P(C) = 0{,}7425$
podobně C_2, C_3 správné odpovědi na další stejně obtížné otázky,
když použijeme předchozí aposteriorní psti jako apriorní

senzitivita, specificita, prevalence

- D – subjekt je nemocen, **prevalence** – podíl nemocných v populaci $\mathrm{P}(D)$, zvolme $\mathrm{P}(D) = 0{,}001$
- nemoc je skrytá, vyhledáváme ji pomocí testu s vlastnostmi:
  - $\mathrm{P}(P|D)$ – pravděpodobnost pozitivního výsledku u nemocného (**senzitivita**, pokud možno velká, zvolme $\mathrm{P}(P|D) = 0{,}98$ pozitivně reaguje 98 % nemocných)
  - $\mathrm{P}(\overline{P}|\overline{D})$ – pravděpodobnost negativního výsledku u zdravého (**specificita**, pokud možno velká, zvolme $\mathrm{P}(\overline{P}|\overline{D}) = 0{,}99$ pozitivně reaguje jen 1 % zdravých)

senzitivita, specificita, prevalence

- jaká je pst, že pozitivně reagující je opravdu nemocný?

$$\mathrm{P}(D|P) = \frac{\mathrm{P}(P|D)\mathrm{P}(D)}{\mathrm{P}(P|D)\mathrm{P}(D) + \mathrm{P}(P|\overline{D})\mathrm{P}(\overline{D})} = \frac{0{,}98 \cdot 0{,}001}{0{,}98 \cdot 0{,}001 + 0{,}01 \cdot 0{,}999} \doteq 0{,}089$$

- jaká je pst, že jde o zdravého člověka v případě, že test byl negativní?

$$\mathrm{P}(\overline{D}|\overline{P}) = \frac{\mathrm{P}(\overline{P}|\overline{D})\mathrm{P}(\overline{D})}{\mathrm{P}(\overline{P}|\overline{D})\mathrm{P}(\overline{D}) + \mathrm{P}(\overline{P}|D)\mathrm{P}(D)} = \frac{0{,}99 \cdot 0{,}999}{0{,}99 \cdot 0{,}999 + 0{,}02 \cdot 0{,}001} = 0{,}99998$$

- porovnej s apriorními pstmi: 0,001 resp. 0,999

náhodná veličina
[random variable]

- číselně vyjádřený výsledek náhodného pokusu
- předem nevíme, který výsledek vyjde, známe jen
  - možné hodnoty
  - jejich pravděpodobnosti
- každému elementárnímu jevu přiřadíme reálné číslo
- **diskrétní rozdělení** náhodné veličiny X
  - model pro počty případů (četnosti)
  - možné hodnoty $x_1^*, x_2^*, \ldots$
  - psti hodnot $\mathrm{P}(X = x_1^*), \mathrm{P}(X = x_2^*), \ldots$ (pstní funkce)
- **spojité rozdělení** náhodné veličiny X
  - model pro spojitou veličiny (délka, váha, koncentrace ...)
  - obor (množina) možných hodnot X
  - hustota $f(x)$

příklad: rodina

náhodná veličina X – počet děvčat
rozdělení X dáno hodnotami x_j^* a pstmi těchto hodnot $\mathrm{P}(X = x_j^*)$

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	$\mathrm{P}(X = x_j^*)$
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

protějšek empirické distribuční funkce (str. 16), [(cumulative) distribution function]

- pst, že X nepřekročí x $\boxed{F_X(x) = \mathrm{P}(X \le x)}$
- diskrétní rozdělení: $\boxed{F(x) = \sum_{t \le x} \mathrm{P}(X = t)}$
- spojité rozdělení: $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$, kde $f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$
- vlastnosti distribuční funkce

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

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

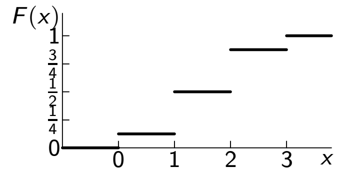
$$\boxed{\mathrm{P}(x_1 < X \le x_2) = F(x_2) - F(x_1)}$$

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

příklad diskrétního rozdělení

rozdělení počtu děvčat X

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



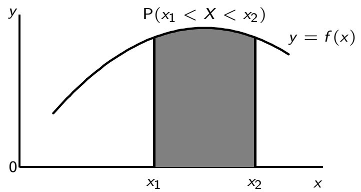
hustota spojitého rozdělení

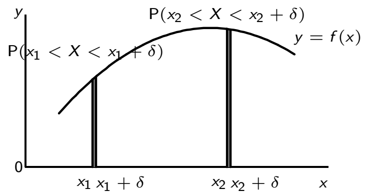
[density function]

- plocha pod celou hustotou je rovna jedné

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

- plocha pod hustotou nad intervalem x_1, x_2 je rovna pravděpodobnosti, že X je mezi x_1, x_2



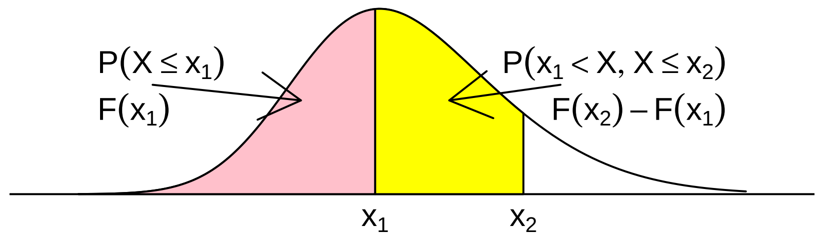


geometrický význam hustoty

$\mathrm{P}(x_1 < X, X <= x_2) = \mathrm{P}(x_1 < X <= x_2)$, vpravo stručnější, používaný zápis

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

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

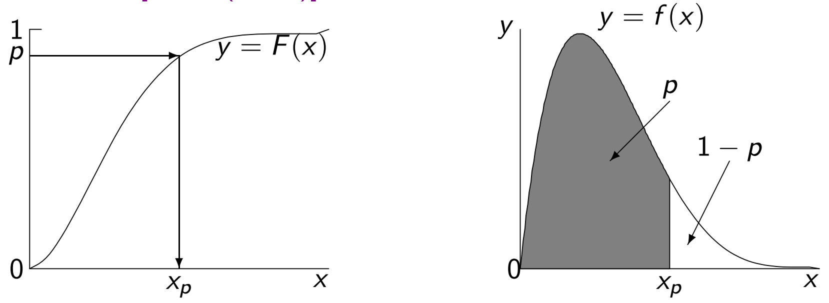


p-kvantil x_p

- x_p je hodnota, pod kterou je $100p$ procent pravděpodobnosti

$$\boxed{\mathrm{P}(X \le x_p) = p}$$

- populační protějšek percentilu
- např. [qnorm(0.975)] dá 1,959964 $\doteq$ 1,96

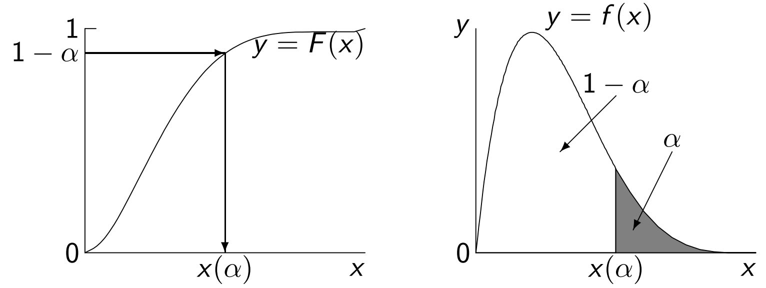


kritická hodnota $x(\alpha)$

kritická hodnota $x(\alpha)$ je překročena s pstí α

$$\boxed{\mathrm{P}(X \ge x(\alpha)) = \alpha}$$

např. [qnorm(1-0.025)] dá 1,959964 $\doteq$ 1,96



střední hodnota

pokračujeme v idealizovaných představách

- míra polohy, **populační průměr**
  [expected value, mean value]
- metoda výpočtu se značí $\mathrm{E}\,X$
- vypočtená hodnota se značí μ nebo úplněji μ_X
- **vážený průměr možných hodnot**
- ideální protějšek výběrového průměru
- diskrétní rozdělení: vahami jsou pravděpodobnosti

$$\boxed{\mu_X = \mathrm{E}\,X = \sum_j x_j^* \mathrm{P}(X = x_j^*)}$$

- spojité rozdělení: místo vah je hustota $f_X(x)$

$$\mu_X = \mathrm{E}\,X = \int_{-\infty}^{\infty} x \, f_X(x) \mathrm{d}x$$

příklad rodina

X – počet děvčat mezi třemi dětmi

j	x_j^*	$\mathrm{P}(X = x_j^*)$	$x_j^* \cdot \mathrm{P}(X = x_j^*)$
1	0	0,125	0,000
2	1	0,375	0,375
3	2	0,375	0,750
4	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, směrodatná odchylka σ

[variance, standard deviation]

- míra variability, **populační rozptyl, popul. směr. odchylka**
- udává velikost kolísání (variabilitu) kolem střední hodnoty
- metoda výpočtu se značí var X
- vypočtená hodnota σ^2, úplněji σ_X^2
- lze vyjádřit pomocí střední hodnoty

$$\sigma_X^2 = \operatorname{var} X = \mathrm{E}\,(X-\mu_X)^2 = \mathrm{E}\,(X^2) - (\mu_X)^2$$

- ideální protějšky výběrového rozptylu, směr. odchylky
- diskrétní rozdělení
$$\boxed{\sigma_X^2 = \operatorname{var} X = \sum_j \left(x_j^* - \mu_X\right)^2 \mathrm{P}\left(X = x_j^*\right)}$$
- spojité rozdělení $\qquad \sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x) \mathrm{d}x$

příklad rodina

X – počet děvčat mezi třemi dětmi, $\mu_X = 1{,}5$

j	x_j^*	$\mathrm{P}(X = x_j^*)$	$x_j^* - \mu_X$	$(x_j^* - \mu_X)^2$	$(x_j^* - \mu_X)^2\,\mathrm{P}(X = x_j^*)$
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			0,75000

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

sdružené rozdělení

- abychom mohli popsat **závislost** náhodných veličin, 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
- zajímá nás rozdělení náhodného vektoru (X, Y)
- proč nemá smysl vyšetřovat **vektor** (X, Z)?
  (protože Z je určeno X jednoznačně: $Z = 3 - X$)

sdružené, marginální a podmíněné rozdělení

sdružené rozdělení – popisuje **společné chování** X, Y

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

marginální rozdělení: chování jedné bez ohledu na hodnotu druhé

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

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

podmíněné rozdělení: chování Y při **dané** hodnotě X

$$\mathrm{P}(Y = y_j^* | X = x_i^*) = \frac{\mathrm{P}(X = x_i^*, Y = y_j^*)}{\mathrm{P}(X = x_i^*)}$$

příklad rodina

X počet děvčat, Y počet děvčat mezi dvěma staršími dětmi

sdružené, marginální a podmíněné rozdělení

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

x_i^*	y_j^* 0	1	2	celkem
0	1/8	0	0	1/8
1	1/8	2/8	0	3/8
2	0	2/8	1/8	3/8
3	0	0	1/8	1/8
	2/8	4/8	2/8	1

x_i^*	y_j^* 0	1	2	celkem
0	1	0	0	1
1	1/3	2/3	0	1
2	0	2/3	1/3	1
3	0	0	1	1

kovariance

protějšek s_{xy}, [covariance]

kovariance vyjadřuje vzájemnou závislost náhodných veličin:

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

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

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

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

náhodné veličiny jsou **nezávislé** právě tehdy, když platí **(ze znalosti hodnoty jedné nic nevíme o druhé)**

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

jsou-li X, Y – nezávislé $\Rightarrow$ $\boxed{\sigma_{X,Y} = 0}$ (nikoliv obrácená implikace)

příklad rodina

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

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

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

shrnutí vlastností populačního průměru a rozptylu

srovnej s požadavky na míry polohy a míry variability

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

pro součet náhodných veličin $X+Y$ dále platí

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

ukázka důkazu

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

normování náhodné veličiny X (populační obdoba z-skórů)

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

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

charakteristiky založené na normované verzi

charakteristiky X nezávislé na μ_X a σ_X, protějšky popisných statistik

- (populační) **korelační koeficient** [correlation coefficient]

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

- (populační) **šikmost** náhodné veličiny X [skewness]

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

- (populační) **špičatost** náhodné veličiny X (někdy bez -3) [kurtosis]

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

alternativní rozdělení

nula-jedničkové, Bernoulliovo

- pouze dvě možné hodnoty: 1 (zdar), 0 (nezdar)
- $\mathrm{P}(X=1)=\pi, \mathrm{P}(X=0)=1-\pi$
- π je jediný parametr, $0<\pi<1$
- X – počet zdarů v jednom pokusu, kde pst zdaru je π
- $X \sim \mathrm{alt}(\pi)$
- $\mu_X = \mathrm{E}\,X = 1\cdot\pi + 0\cdot(1-\pi) = \pi$
- $\sigma_X^2 = \mathrm{var}\,X = (1-\pi)^2\cdot\pi + (0-\pi)^2\cdot(1-\pi) = \pi(1-\pi)$

binomické rozdělení

[binomial distribution]

- $Y \sim \mathrm{bi}(n,\pi)$
- n **nezávislých** pokusů takových, že
- $\mathrm{P}(\text{zdar}) = \pi, \mathrm{P}(\text{nezdar}) = 1-\pi$, $(0<\pi<1)$
- Y je **počet zdarů** v těchto pokusech
- $\boxed{\mathrm{P}(Y=k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}, \quad k=0,1,\ldots,n}$
  [dbinom(k,n,pst)]
- např. ze 7 vajíček se vylíhne Y slepiček, $Y\sim\mathrm{bi}(7,1/2)$
- např. při 60 hodech kostkou padlo Y šestek, $Y\sim\mathrm{bi}(60,1/6)$
- předem nevíme, kolik bude slepiček (šestek), ale v dlouhodobém průměru je relativní četnost blízká 1/2 (1/6)

binomické rozdělení pomocí alternativního

- $Y \sim \mathrm{bi}(n, \pi)$
- Y je celkový počet zdarů v n pokusech, tedy
- $Y = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$, kde X_i je počet zdarů v i-tém pokusu
- z vlastností střední hodnoty (očekávaný počet zdarů)

$$\mu_Y = \mathrm{E}\,Y = \mathrm{E}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathrm{E}\,X_i = \sum_{i=1}^{n} \pi = n\pi$$

- protože jsou pokusy nezávislé

$$\sigma_Y^2 = \mathrm{var}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathrm{var}\,X_i = \sum_{i=1}^{n} \pi(1-\pi) = n\pi(1-\pi)$$

příklad: kuřáci

- mezi dvacetiletými muži je (řekněme) 35 % kuřáků ($\pi = 0{,}35$)
- je-li dvacetiletých 70 tisíc ($m = 70\ 000$), pak je kuřáků asi $m\pi = 70\ 000 \cdot 0{,}35 = 24\ 500$, ale nevíme, kteří to jsou
- vyberme náhodně $n = 60$ dvacetiletých mužů, označme jako Y počet kuřáků mezi nimi, je tedy $Y \sim \mathrm{bi}(60, 0{,}35)$
- střední hodnota (očekávaný počet), rozptyl

$$\mu_Y = 60 \cdot 0{,}35 = 21 \qquad \sigma_Y^2 = 60 \cdot 0{,}35 \cdot 0{,}65 = 13{,}65 \doteq (3{,}7)^2$$

- ukázky pravděpodobností možných hodnot

k	15	17	19	21	23	25
$\mathrm{P}(Y = k)$	0,029	0,062	0,095	0,107	0,091	0,059

- psti počítány pomocí [dbinom(0:60,60,0.35)]

Poissonovo rozdělení
[Poisson distribution]

- $X \sim \mathrm{Po}(\lambda) \qquad (\lambda > 0)$
- zákon vzácných (řídkých) jevů
- kolikrát nastal jev během jednotkového časového intervalu, na jednotkové ploše, v jednotkovém objemu . . .
- $\boxed{\mathrm{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}}, \qquad k = 0, 1, \ldots$
- $\mu_X = \lambda$, $\sigma_X^2 = \lambda$
- pro velké n a malé π lze rozdělení $\mathrm{bi}(n, \pi)$ aproximovat pomocí rozdělení $\mathrm{Po}(n\pi)$
- např. počet kolonií na Petriho misce

příklad
s jakou pstí udělá 5 z 55 stejně připravených studentů zkoušku na výbornou, je-li pst jedničky 0,1?

- binomické rozdělení $Y \sim \mathrm{bi}(55, 0{,}1)$ [dbinom(5,55,0.1)]

$$\mathrm{P}(Y = 5) = \binom{55}{5} \cdot 0{,}1^5 \cdot 0{,}9^{50} = 0{,}179$$

- aproximace Poissonovým rozdělením $Y \sim \mathrm{Po}(55 \cdot 0{,}1) = \mathrm{Po}(5{,}5)$ [dpois(5, 5.5)]

$$\mathrm{P}(Y = 5) = \frac{5{,}5^5}{5!} e^{-5{,}5} = 0{,}171$$

normální (Gaussovo) rozdělení $\mathrm{N}(\mu, \sigma^2)$

[normal (Gaussian) distribution]

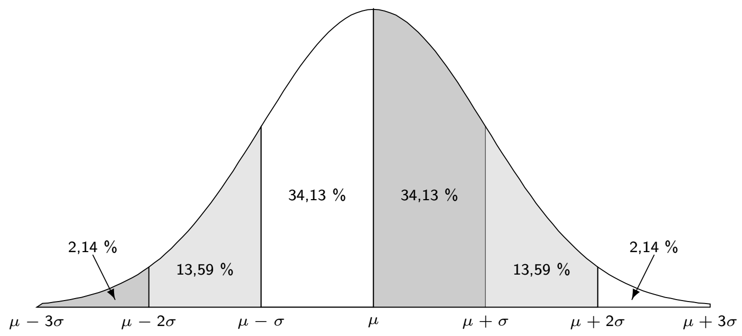
- $\boxed{\mu_X = \mu,\ \sigma_X^2 = \sigma^2}$
- spojité rozdělení, symetrické okolo střední hodnoty μ
- maximální hodnota hustoty přibližně $0{,}4/\sigma$
- $\mathrm{N}(0, 1)$ (normované normální rozdělení):
  $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ (hustota),
  $\Phi(x) = \int_{-\infty}^{x} \varphi(t) dt$ (distr. fce)
- $X \sim \mathrm{N}(\mu, \sigma^2)$, pak $Z = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0, 1)$

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

- model vzniku: součet velkého počtu nepatrných příspěvků
- velmi často modeluje znaky v poměrovém měřítku

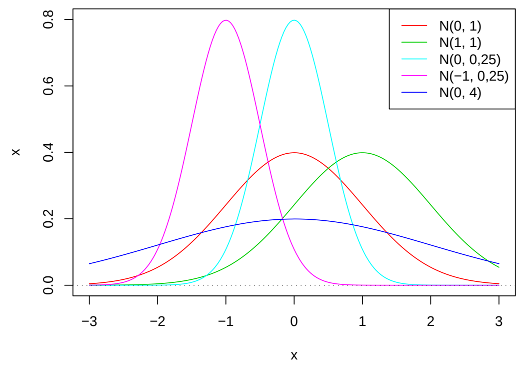
hustota $\mathrm{N}(\mu, \sigma^2)$

[dnorm(x,mu,sigma)]



normální (Gaussovo) rozdělení $\mathrm{N}(\mu, \sigma^2)$

význam parametrů



- spojité rozdělení, symetrické okolo střední hodnoty μ
- maximální hodnota hustoty přibližně $0{,}4/\sigma$
- model vzniku: součet velkého počtu nepatrných příspěvků

výpočet pravděpodobnosti, že $a < X < b$

použije distribuční funkci $\mathrm{N}(0, 1)$

$\mathrm{P}(a < X < b) = F_X(b) - F_X(a)$ platí obecně pro spoj. rozděl.

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

$$\mathrm{P}(X \le x) = \mathrm{P}\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \mathrm{P}\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

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

[pnorm((b-mu)/sigma)-pnorm((a-mu)/sigma)]

v programu R je distribuční funkce $\mathrm{N}(\mu, \sigma^2)$ s obecnými parametry:

[pnorm(b,mu,sigma)-pnorm(a,mu,sigma)]

příklad

- u jakého dílu populace desetiletých hochů naměříme výšku od 135 do 140 cm, když pro výšku desetiletých platí $X \sim \mathrm{N}(136{,}1, 6{,}4^2)$
- předpokládáme zaokrouhlování na celá čísla při měření:

$$\mathrm{P}(134{,}5 < X < 140{,}5) = \Phi\left(\frac{140{,}5 - 136{,}1}{6{,}4}\right) - \Phi\left(\frac{134{,}5 - 136{,}1}{6{,}4}\right)$$
$$= 0{,}754 - 0{,}401 = 0{,}353$$

[pnorm((140.5-136.1)/6.4)-pnorm((134.5-136.1)/6.4)]

- pomocí distribuční fce s obecnými parametry [pnorm(140.5,136.1,6.4)-pnorm(134.5,136.1,6.4)]

kritické hodnoty normálního a Studentova t-rozdělení

[Student distribution]

- **normální rozdělení** $\mathrm{N}(0,1)$ [qnorm(1-alpha)]

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

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

- **Studentovo t-rozdělení** t_k (podobné normálnímu, protože místo σ používá jeho odhad, má větší rozptyl)

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

- jsou to spíše kritické hodnoty $|T|$ [qt(1-alpha/2,k)]

některé kritické hodnoty

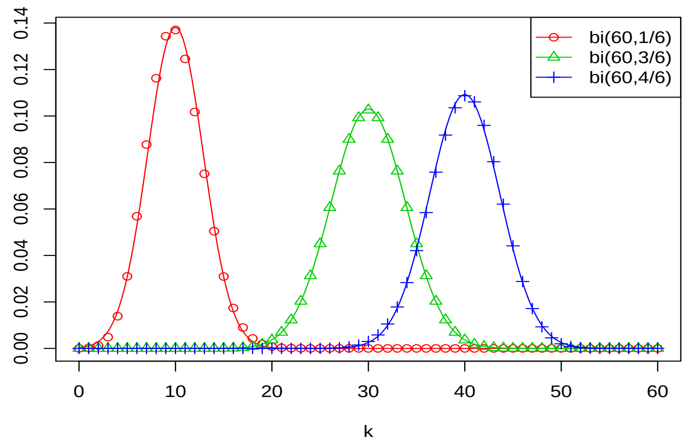
α	0,50	0,25	0,10	**0,05**	0,01
$z(\alpha/2)$	0,674	1,150	1,645	**1,960**	2,576
$t_{100}(\alpha)$	0,677	1,157	1,660	**1,984**	2,626
$t_{20}(\alpha)$	0,687	1,185	1,725	**2,086**	2,845
$t_5(\alpha)$	0,727	1,301	2,015	**2,571**	4,032

- $T \sim \mathrm{t}_k$ má jediný parametr k (počet stupňů volnosti)
- s rostoucím k se chování blíží normálnímu rozdělení
- pro $Z \sim \mathrm{N}(0,1)$ je 95 % hodnot v intervalu $(-1{,}960; 1{,}960)$
- pro $T \sim \mathrm{t}_5$ je 95 % hodnot v intervalu $(-2{,}571; 2{,}571)$
- pro $T \sim \mathrm{t}_{20}$ je 95 % hodnot v intervalu $(-2{,}086; 2{,}086)$
- pro $T \sim \mathrm{t}_{100}$ je 95 % hodnot v intervalu $(-1{,}984; 1{,}984)$

aproximace binomického rozdělení normálním

se stejnou střední hodnotou a stejným rozptylem

rozdělení $\mathrm{bi}(n, \pi)$ lze aproximovat pomocí $\mathrm{N}(n\pi, n\pi(1-\pi))$



další rozdělení související s normálním

[F-distribution, chi-square distribution]

- V má rozdělení (musí být $\mathrm{P}(V > 0) = 1$!!) **logaritmicko-normální**, platí-li $\boxed{\ln V \sim \mathrm{N}(\mu, \sigma^2)}$
- **Fisherovo F-rozdělení** $\mathrm{F}_{k,m}$ [qf(1-alpha,k,m)]

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

- **rozdělení chí-kvadrát** χ^2_k [qchisq(1-alpha,k)]

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

- speciálně platí:
  - $\chi^2_1(0{,}05) = 3{,}841 = 1{,}960^2$
  - $\chi^2_1(\alpha) = z(\alpha/2)^2$
  - $F_{1,m}(\alpha) = (t_m(\alpha))^2$

populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- **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 (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- **statistika** funkce náhodného výběru (pozorování)
- **odhad** statistika použitá k odhadu parametru

průměr z náhodného výběru

- $X_1, \ldots, X_n$ **nezávislé**, stejné rozdělení — **náhodný výběr**
  $\mu_{X_i} = \mathrm{E}\, X_i = \mu$ (stejná střední hodnota) — populační průměr
  $\sigma^2_{X_i} = \mathrm{var}\, X_i = \sigma^2$ (stejný rozptyl) — populační rozptyl
- $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ — výběrový průměr
- $\mu_{\bar{X}} = \mathrm{E}\, \bar{X} = \mu$
  - výběrový průměr $\bar{X}$ je opět náhodná veličina
  - je **nestranným** odhadem [unbiased estimator] parametru μ
  - nestranným odhadem populačního průměru (střední hodnoty)
  - když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

rozptyl průměru z náhodného výběru

$$\boxed{\sigma^2_{\bar{X}} = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}\, X_i = \frac{\sigma^2}{n} = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = (\mathrm{S.E.}(\bar{X}))^2}$$

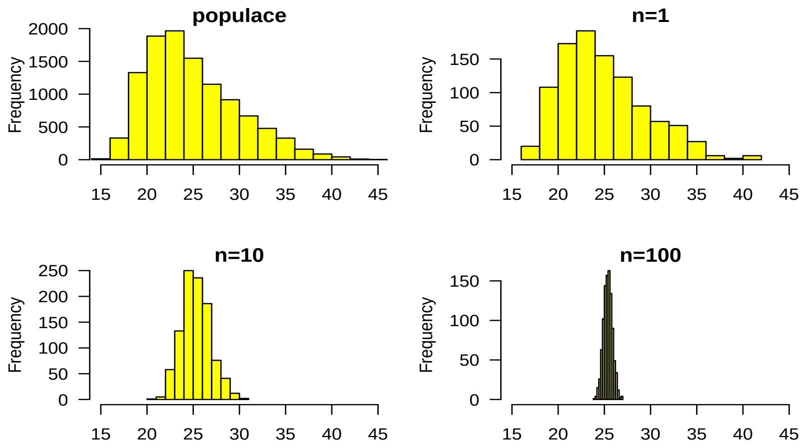
- $\mathrm{S.E.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ – **střední chyba průměru** [standard error of mean]
- variabilita průměrů (měřená rozptylem) z výběrů rozsahu n je n-krát menší, než variabilita jednotlivých pozorování σ^2
- střední chyba průměru je $\sqrt{n}$-krát menší než σ
- čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- speciálně pro normální rozdělení $X_i \sim \mathrm{N}(\mu, \sigma^2)$ nezávislé:

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

(všimněte si závislosti na n)

příklad: věk matek (umělá situace)

populace - 10 916 matek, opakované výběry rozsahu $n = 1, 10, 100$
je patrná variabilita klesající s rostoucím n



příklad: věk matek – shrnutí

- velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- na rozdíl od běžné praxe **známe** populační průměr μ
- náhodně vybráno 1000 matek (vlastně průměry výběrů rozsahu $n = 1$), nakreslen histogram
- 1000 krát náhodně vybráno vždy $n = 10$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- 1000 krát náhodně vybráno vždy $n = 100$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- skutečné rozptyly (odhady z 1000 realizací): 23,5; 2,20; 0,21

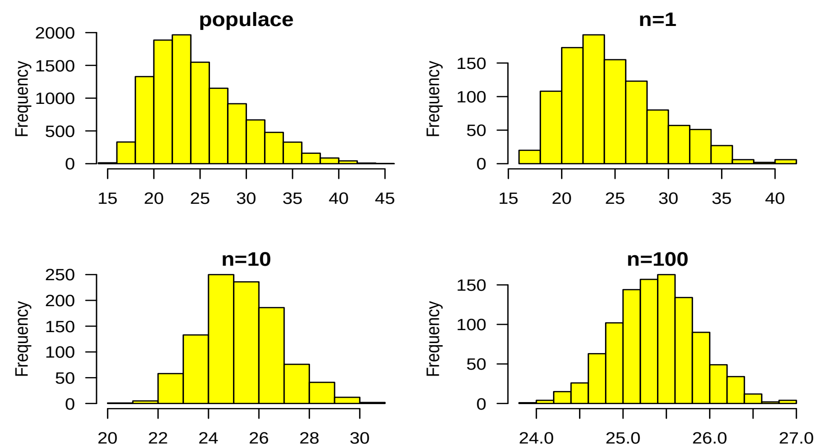
centrální limitní věta (CLV, CLT)

[Central Limit Theorem]

- 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 pocházet z normálního rozdělení). Potom **pro velké** n má průměr $\bar{X}$ přibližně rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, součet $X_1 + \ldots + X_n$ pak rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- prakticky: **průměr** má pro dost velká n **normální rozdělení** s rozptylem n-krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
- CLT je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velikého počtu nahodilých malých vlivů
- příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení

příklad: věk matek (nestejná měřítka!)

populace - 10 916 matek, opakované výběry rozsahu $n = 1, 10, 100$
je patrno, že s rostoucím n se histogram blíží histogramu norm. rozdělení



příklad: věk matek

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
populace	μ =25,41	σ =4,932	γ_1 =0,771	γ_2 =0,189

interval spolehlivosti pro μ (výběr z $\mathrm{N}(\mu, \sigma^2)$)

[confidence interval]

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

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

- což je totéž, jako (μ se od $\bar{X}$ liší nejvýše ...)

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

- tedy (všimněte si zkracování intervalu s rostoucím n)

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

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

interpretace intervalu spolehlivosti

- je to **intervalový** odhad hodnoty μ
- $\bar{X}$ je **bodový** odhad
- **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé** μ (**odhadovaný parametr**)
- kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu μ, ve zbylých asi 5 % zůstane skutečné μ mimo interval spolehlivosti
- pro obecné α (spolehlivost $1-\alpha$):

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

- POZOR na nesprávné interpretace, vypovídá o neznámé **konstantě** μ, nikoliv o **náhodných veličinách** X nebo $\bar{X}$

interval spolehlivosti při neznámém σ

- pro X_i s normálním rozdělením je třeba použít kritické hodnoty Studentova t-rozdělení (pozor na **jinak značené** kritické hodnoty Studentova t-rozdělení)

$$\mathrm{P}\left(\bar{X} - \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha)\right) = 1-\alpha$$

- jako odhad σ^2 se použije výběrový rozptyl

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

- při velkých n ($n \geq 50$) stačí použít $z(\alpha/2)$ místo $t_{n-1}(\alpha)$
- interval spolehlivosti se počítá i při odhadu jiných parametrů
- je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

příklad: věk matek I

normální rozdělení dáno CLT a velkým *n*

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

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

  [confint(lm(vek.m~1,data=Kojeni))]
- 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)
- větší jistota způsobí delší interval spolehlivosti (méně vypovídající tvrzení)

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

  [confint(lm(vek.m~1,data=Kojeni),level=0.99)]

příklad: věk matek II

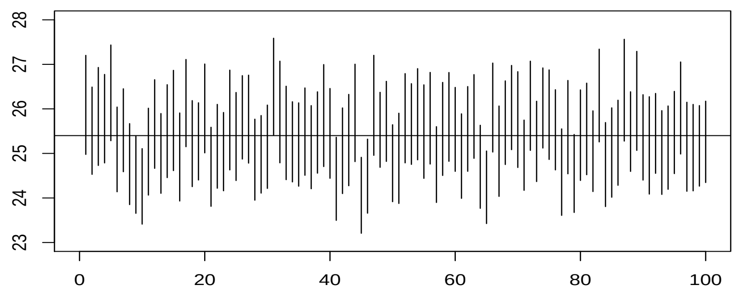
normální rozdělení dáno CLT a velkým *n*

- 90% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left(25{,}7 - 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (25{,}0; 26{,}4)$$

  [confint(lm(vek.m~1,data=Kojeni),level=0.9)]
- příklady **nesprávné interpretace** 90% intervalu spolehlivosti:
  - *90 % žen má věk v intervalu (25,0; 26,4)*
    např. mezi našimi 99 matkami je jen 12 ve věku 25 a 10 ve věku 26 roků, navíc, s rostoucím *n* se interval zužuje
  - *výběrový průměr věku matek je s pravděpodobností 90 % v intervalu (25,0; 26,4)*
    výběrový průměr je uprostřed (tedy uvnitř) intervalu **vždy**

simulované výběry pro $n = 100$ (věk matek)



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

centrální limitní věta pro četnosti

- (CLT obecně:) 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 přibl. rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet přibl. rozdělení $\mathrm{N}(n\mu, n\sigma^2)$.
- $Y \sim \mathrm{bi}(n, \pi)$: Y je absolutní četnost výskytu jevu s pstí π v n nezáv. pokusech
- $Y = \sum_{i=1}^{n} X_i$ je součet nezávislých náhodných veličin X_i s alternativním rozdělením, $X_i \sim \mathrm{alt}(\pi), \mathrm{var}\, X_i = \pi(1-\pi)$
- podle CLT proto přibližně $Y \dot\sim \mathrm{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $Y/n = \bar{X}$ je průměr veličin s alternativním rozdělením, označme $\hat\pi = Y/n$
- podle CLT je přibližně $\hat\pi \dot\sim \mathrm{N}(\pi, \pi(1-\pi)/n)$
- $\hat\pi$ je **nestranný** odhad π

interval spolehlivosti pro pravděpodobnost π

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

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

- existují přesnější (pracnější) postupy
  [prop.test(y,n,correct=FALSE)]
  [binom.test(y,n)]

příklad: hody s hrací kostkou

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

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

- kostka B: $n = 100, y = 41, \hat{\pi}_B = 0{,}41$

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

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

příklad

- v roce 1951 bylo provedeno rozsáhlé měření výšky desetiletých hochů, byla vyšetřena v populaci desetiletých: $\mu = 136{,}1$ cm, $\sigma = 6{,}4$ cm
- na základě výběru pořízeného v roce 1961 máme rozhodnout, zda se po deseti letech výška populace desetiletých zvýšila
- hodnoty zjištěné v roce 1961 [cm]: 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147
- $\bar{x} = 139{,}13$, $s^2 = 6{,}56^2$, vše v cm
- výběr z roku 1961 mohl obsahovat jiných 15 hochů, tedy mohl vést k jinému průměru
- stačí rozdíl $139{,}13 - 136{,}1$, abychom prokázali změnu?

testování statistických hypotéz

[hypothesis testing, null hypothesis, alternative hypothesis, critical (rejection) region, Type I (II) error, significance level]

- **nulová hypotéza** H_0: tvrzení o populaci (parametru), o jehož platnosti rozhodujeme (**ne**ní rozdíl, **ne**závisí, **ne**liší se od ...)
- **alternativní hypotéza** H_1: (alternativa) zbývající možnost (k H_0), často „vědecká hypotéza“, kterou chceme dokázat
- volba mezi H_0, H_1 dána, volíme **o čem** budou hypotézy
- **kritický obor**: možné výsledky pokusu, kdy H_0 zamítáme; zpravidla popsán pomocí statistiky (např. $|Z| \geq z(\alpha/2)$)
- **obor přijetí**: možné výsledky pokusu, kdy H_0 nezamítáme
- **chyba prvního druhu**: (náhodný jev) rozhodnutí zamítnout H_0, když platí H_0, tj. falešně prokázat „vědeckou hypotézu“
- **chyba druhého druhu**: (náhodný jev) rozhodnutí nezamítnout H_0, když platí H_1, tj. nepoznat neplatnost H_0

statistické rozhodování

[significance level, power, p-value]

- **hladina testu** α (zpravidla 5 %, 1 %)
  - maximální dovolená pst chyby prvního druhu
  - volí se před pokusem, nezávisle na jeho výsledku
- **síla testu** $1-\beta$
  - pravděpodobnost zamítnutí neplatné H_0
  - pst, s jakou prokážeme platnou „vědeckou hypotézu"
  - závisí na tom, co opravdu platí
- **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
  - „stupeň důvěry" v platnost nulové hypotézy
- H_0 se **zamítá**, právě když $\boxed{p \le \alpha}$ (**zapamatovat**)

testování statistických hypotéz

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

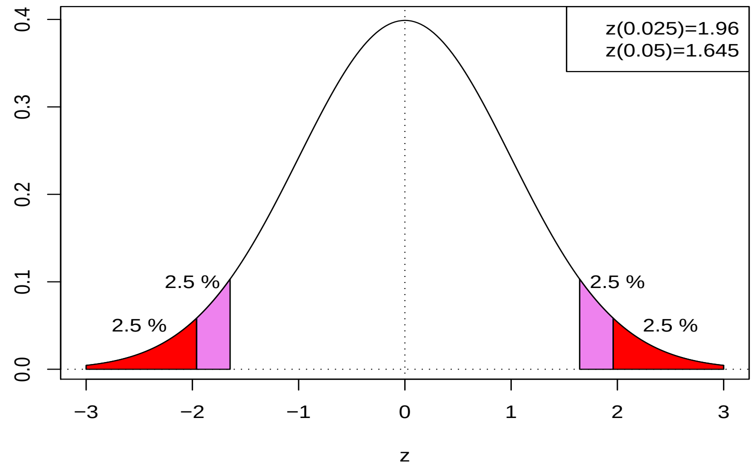
- zamítnutí $\Leftrightarrow$ výsledek pokusu v kritickém oboru
- přijetí $\Leftrightarrow$ výsledek pokusu v oboru přijetí
- nikdy spolehlivě nevíme, zda H_0 platí

rozhodování o populačním průměru normálního rozdělení (σ známé)

- $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ **nezávislé**; $\sigma > 0$ známe
- $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, tedy $\mathrm{S.E.}(\bar{X}) = \sigma/\sqrt{n}$
- $H_0 : \mu = \mu_0$ (dané číslo, jiný zápis $H_0 : \mu - \mu_0 = 0$)
- platí-li H_0, pak $\boxed{Z = \dfrac{\bar{X}-\mu_0}{\mathrm{S.E.}(\bar{X})} = \dfrac{\bar{X}-\mu_0}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)}$
- $H_1 : \mu \ne \mu_0 \Rightarrow$ kritický obor: $|Z|$ velké, tj. $\boxed{|Z| \ge z(\alpha/2)}$
- $H_1 : \mu > \mu_0$: zamítnout pro $\boxed{Z \ge z(\alpha)}$
- $H_1 : \mu < \mu_0$: zamítnout pro $\boxed{Z \le -z(\alpha)}$
- volba jednostranné alternativy jen podle zadání úlohy, nikoliv podle výsledku pokusu

kritický obor pro Z

červeně na 5% hladině, červeně a fialově na 10% hladině



příklad: výška desetiletých chlapců

- zvolíme klasickou hladinu $\alpha = 5\ \%$
- v roce 1951 velký výběr: $\mu_0 = 136{,}1$ cm, $\sigma = 6{,}4$ cm
- v roce 1961 změřeno $n = 15$ náhodně vybraných desetiletých hochů, $\bar{x} = 139{,}13$ cm
- stačí tento vzrůst k důkazu, že nová generace je vyšší?
- vzrostla výška desetiletých ? $H_0 : \mu = \mu_0$ proti $H_1 : \mu > \mu_0$

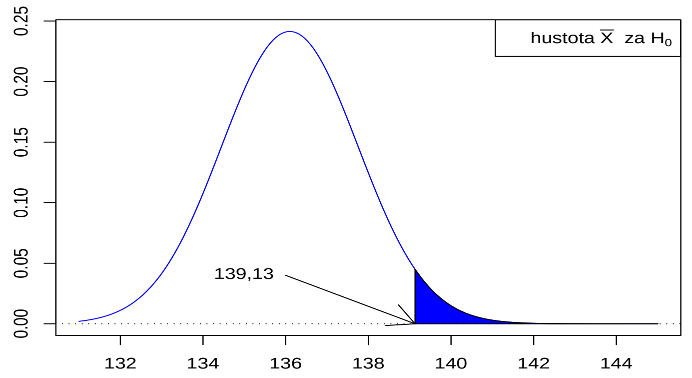
$$z = \frac{139{,}13 - 136{,}1}{6{,}4}\sqrt{15} = 1{,}836$$

- $z(0{,}05) = 1{,}645 < 1{,}836$, tedy H_0 na 5% hladině **zamítáme**
- na 5% hladině jsme prokázali, že nová generace je vyšší
- v případě, že nová generace není vyšší, riskovali jsme jen 5% pravděpodobnost, že budeme nesprávně tvrdit, že vyšší je

výška desetiletých hochů

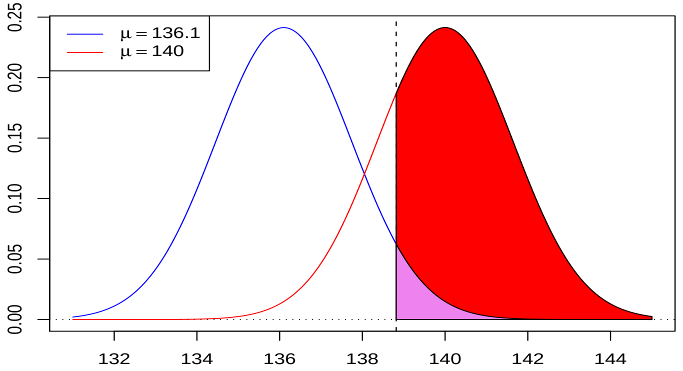
hustota $\bar{X}$ za platnosti hypotézy

- p-hodnota – pst, že za H_0 vyjde $Z > 1{,}836$ tj. $\bar{X} > 139{,}13$ [1-pnorm(1.836)]
- p-hodnota – modrá plocha napravo od 139,13, $p = 3{,}3\ \%$



výška desetiletých chlapců

hustota $\bar{X}$ za hypotézy (modře) a při $\mu = 140$ (červeně)
hladina testu – fialová plocha, síla testu – fialová + červená plocha



$$136{,}1 + 6{,}4/\sqrt{15} \cdot 1{,}645 = 138{,}8$$

volba rozsahu výběru

$H_0 : \mu = \mu_0$ proti $H_1 : \mu \neq \mu_0$

- pro zvolenou hodnotu $\mu_1 \neq \mu_0$ požadujeme sílu $1 - \beta$
- $1 - \beta$ je pravděpodobnost, s jakou odhalíme neplatnost H_0, je-li skutečnost $\mu = \mu_1$

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

- při jednostranné alternativě by bylo $z(\alpha)$ místo $z(\alpha/2)$
- aby pro $\mu_1 = 140$ byla síla 90 % (tj. $1 - \beta = 0{,}9$, $\beta = 0{,}1$, $z(0{,}1) = 1{,}282$), 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

výběr z $\mathrm{N}(\mu, \sigma^2)$, σ neznámé

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

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

- statistika (místo σ použijeme S_x)

$$\boxed{T = \frac{\bar{X} - \mu_0}{\mathrm{S.E.}(\bar{X})} = \frac{\bar{X} - \mu_0}{S_x} \sqrt{n}}$$

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

výšky hochů pro případ neznámého σ

- $\mathrm{H}_0 : \mu = 136{,}1$ proti $\mathrm{H}_1 : \mu > 136{,}1$ ($\alpha = 5$ %)

$$\bar{x} = 139{,}133 \qquad s_x^2 = 6{,}556^2$$
$$t = \frac{139{,}133 - 136{,}1}{6{,}556}\sqrt{15} = 1{,}792 > 1{,}761 = t_{14}(0{,}10)$$
$$p = \mathrm{P}(T \geq 1{,}792) = 0{,}047 \qquad (\text{ tj. } 4{,}7\ \%)$$

- na 5% hladině jsme prokázali zvýšení populačního průměru (H_0 se na 5% hladině **zamítá**)
- [t.test(hosi,mu=136.1,alternative="greater")]

výšky hochů pro případ neznámého σ

(jiné zadání úlohy)

- **kdybychom** předem neměli určenu jednostrannou alternativu, ale zvolili $\mathrm{H}_1 : \mu \neq 136{,}1$, pak

$$|t| = |1{,}792| < 2{,}145 = t_{14}(0{,}05)$$
$$p = \mathrm{P}(|T| \geq 1{,}792) = 0{,}0948 \qquad (\text{tj. } 9{,}48\ \%)$$

- hypotézu na 5% hladině nezamítáme
- [t.test(hosi,mu=136.1,alternative="two.sided")], stačí ale [t.test(hosi,mu=136.1)]

interval spolehlivosti pro μ

souvislost s testem o μ při oboustranné alternativě

- oboustranný interval spolehlivosti pro μ

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

- μ_0 patří do intervalu spolehlivosti, právě když platí

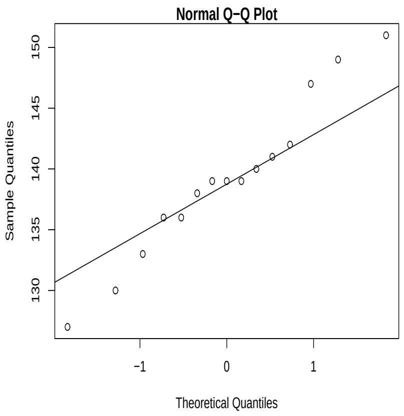
$$|\bar{X} - \mu_0| < \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha)$$

- tedy, právě když se nezamítne hypotéza $\mathrm{H}_0 : \mu = \mu_0$ při oboustranné alternativě $\mathrm{H}_1 : \mu \neq \mu_0$
- interval spolehlivosti obsahuje takové hodnoty μ_0, pro které bychom **nezamítli** hypotézu $\mathrm{H}_0 : \mu = \mu_0$
- podobně u jednostranných intervalů spolehlivosti a jednostranných alternativ

výšky hochů pro případ neznámého σ

- 95% interval spolehlivosti: (135,5; 142,8)
  s 95% pravděpodobností je skutečný populační průměr (střední hodnota μ) v uvedeném intervalu
- je jen 5% riziko, že leží mimo uvedený interval
- 99% interval spolehlivosti (134,1; 144,2)
  [t.test(hosi,mu=136.1,conf.level=0.99)] (vedlejší výsledek)
  [confint(lm(hosi~1),level=0.99)]
- aby byla zajištěna větší spolehlivost intervalu (větší pravděpodobnost, že zachytí skutečnou hodnotu), je nutně 99% interval spolehlivosti delší, než 95% interval spolehlivosti

ověření předpokladu o normálním rozdělení



- **Shapirův-Wilkův test**
- H_0 : normální rozdělení s nějakými (neznámými) parametry
- [shapiro.test(hosi)]
- $W = 0{,}966$, $p = 80$ %
- hodnotí kvalitu přiblížení bodů k přímce na diagramu normality
- [qqnorm(hosi);qqline(hosi)]

pst výskytu jevu

test hypotézy o parametru π binomického rozdělení

- $Y \sim \mathrm{bi}(n, \pi) \quad \mathrm{H}_0 : \pi = \pi_0$:
  $$Z = \frac{Y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\mathrm{S.E.}(\hat{\pi})} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_o)/n}} \dot{\sim} \mathrm{N}(0,1)$$
- někdy s opravou na spojitost (Yates)
  $$Z = \frac{|Y - n\pi_0| - 0{,}5}{\sqrt{n\pi_0(1-\pi_0)}} \mathrm{sign}(Y - n\pi_0) \dot{\sim} \mathrm{N}(0,1)$$
- $\mathrm{H}_1 : \pi \neq \pi_0$: zamítnout pokud $|Z| \geq z(\alpha/2)$
- $\mathrm{H}_1 : \pi > \pi_0$: zamítnout pokud $Z \geq z(\alpha)$
- $\mathrm{H}_1 : \pi < \pi_0$: zamítnout pokud $Z \leq -z(\alpha)$
- existuje přesný postup, bez použití aproximace

příklad kalous

- pokusit se prokázat, že kalous dá přednost infikované myši před neinfikovanou
- Y – počet „zdarů“, $n = 50$, π – pst, že zvolí infikovanou
- Y má **binomické rozdělení**
  za $\mathrm{H}_0 : \pi = 1/2 (= \pi_0)$ (myši se neliší) $Y \sim \mathrm{bi}(50, 1/2)$
- **alternativní hypotéza**: $\mathrm{H}_1 : \pi > 1/2$
- data: z 50 případů dal kalous ve 33 případech přednost infikované myši před neinfikovanou
- **kritický obor**: velká hodnota Y (tj. velké $\hat{\pi}$ resp. velké Z)
  $$z = \frac{33 - 50 \cdot 0{,}5}{\sqrt{50 \cdot 0{,}5 \cdot 0{,}5}} = 2{,}263 \qquad p = \mathrm{P}(Z \geq 2{,}263) = 0{,}0118$$
- s opravou na spojitost:
  $$z = \frac{33 - 50 \cdot 0{,}5 - 0{,}5}{\sqrt{50 \cdot 0{,}5 \cdot 0{,}5}} = 2{,}121 \qquad p = \mathrm{P}(Z \geq 2{,}121) = 0{,}0169$$

příklad kalous

- prop.test() počítá Z^2, která má za H_0 : rozdělení χ^2_1
  [prop.test(33,50,alternative="greater",correct=FALSE)]
  [prop.test(33,50,alternative="greater")]
  [binom.test(33,50,alternative="greater")]
- **dosažená hladina**: za H_0 počítaná pst, že dostaneme výsledek aspoň tolik odporující nulové hypotéze, jako ve skutečném pokusu:

$$\begin{aligned} p &= \mathrm{P}(Y \geq 33) = 1 - \mathrm{P}(Y \leq 32) \\ &= \sum_{k=33}^{50} \binom{50}{k} 0{,}5^k (1-0{,}5)^{50-k} \\ &= 0{,}0164 \end{aligned}$$

[1-pbinom(32,50,1/2)]

párové testy

(převedou se na jednovýběrové)

- $(U_1, V_1), \ldots, (U_n, V_n)$ – párová pozorování
  nezávislé dvojice (možná) závislých náhodných veličin
- těsná závislost uvnitř dvojic je výhodná
- $X_i = U_i - V_i$ (označení rozdílů)
- předpokládáme **stejné rozdělení** $X_1, \ldots, X_n$
- U_i, V_i – dvojice měření na stejných jedincích, např. hodnota zjištěná před ošetřením a po něm
- např. výška otce a jeho syna nebo věk otce a věk matky
- **nezajímá nás** zda je mezi nimi **závislost**, tu připouštíme,
  zajímá nás zda jsou **co do polohy stejné**,
  nebo např. synové v (populačním) průměru vyšší, než otcové
- H_0 tvrdí, že např. mezi výškami otců a synů **není rozdíl**, tedy že rozdíly X_i **kolísají kolem nuly**

párový t-test

předpoklad normálního rozdělení rozdílů

- **normální** rozdělení: $X_i = U_i - V_i \sim \mathrm{N}(\mu, \sigma^2)$, **nezávislé**
- $H_0 : \mu = 0$ (pak je $\mu_U = \mu_V$)
- odhad σ^2: $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- $T = \dfrac{\bar{X}}{\mathrm{S.E.}(\bar{X})} = \dfrac{\bar{X}}{S}\sqrt{n} = \dfrac{\bar{U} - \bar{V}}{\mathrm{S.E.}(\bar{U} - \bar{V})}$
- 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 $T \geq t_{n-1}(2\alpha)$
- vlastně jednovýběrový t-test pro $X_i = U_i - V_i$

příklad: výšky rodičů (párová pozorování!)

H_0: otcové jsou o 10 cm vyšší matek

- U – výška otce, V – výška matky, $X = U - V - 10$
- $\alpha = 0{,}05, H_0 : \mu_U - 10 = \mu_V$ resp. $\mu_U - \mu_V = 10$
- $n = 99, \bar{u} = 179{,}267, \bar{v} = 166{,}970$
- $\bar{x} = \bar{u} - \bar{v} - 10 = 2{,}293, s_X = s_{U-10-V} = s_{U-V} = 8{,}144$
- $t = \frac{2{,}293}{8{,}144}\sqrt{99} = 2{,}801$, tedy $|t| > t_{98}(0{,}05) = 1{,}9845 \quad \Rightarrow$ zamítnout H_0
- $p = \mathrm{P}(|T| \geq t) = 0{,}0061 \qquad (0{,}61\ \%)$
- 95% interval spolehlivosti pro $\mu_U - \mu_V$:

$$\left(12{,}293 - \frac{8{,}144}{\sqrt{99}} 1{,}9845\ ;\ 12{,}293 + \frac{8{,}144}{\sqrt{99}} 1{,}9845\right) = (10{,}67; 13{,}92)$$

[shapiro.test(vyska.o-vyska.m)] ověření normality
[t.test(vyska.o,vyska.m, mu=10, paired=TRUE)]
[t.test(vyska.o-vyska.m, mu=10)]

znaménkový test

bez předpokladu normálního rozdělení, stačí libovolné spojité

- stačí znát znaménka rozdílů $X_i = U_i - V_i$
- pozorování s $U_i = V_i$ (tj. $X_i = 0$) se vynechají, upraví se n
- Y – počet **kladných** znamének $X_i = U_i - V_i$
- H_0 : rozdělení U a V jsou stejná, pak je nutně $P(U_i > V_i) = P(X_i > 0) = 1/2$, tedy $Y \sim \mathrm{bi}(n, 1/2)$
- H_0 zamítáme pro velká nebo malá Y:

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

- pro malá n je bezpečnější použít Yatesovu korekci

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

příklad: věk rodičů (párová pozorování!)

normalitu věku rodičů sotva lze předpokládat

- celkem 99 dvojic (otec, matka), sledujeme jejich věk (U, V)
- $H_0 : \mathrm{E}\,U = \mathrm{E}\,V + 2$ (populační míra polohy věku otců je o 2 roky větší, než matek), H_1 oboustranná
- v jedenácti případech je vek.o – vek.m $= 2$, tyto dvojice nepoužijeme, proto $n = 99 - 11 = 88$
- u 50 dvojic je vek.o – vek.m > 2, proto

$$z = \frac{50 - 88/2}{\sqrt{88/4}} = 1{,}279, \qquad p = 0{,}201 \ (20{,}1\ \%)$$

- s Yatesovou korekcí: $z = 1{,}172$, $p = 0{,}241$ $(24{,}1\ \%)$

[n = sum(vek.o-vek.m != 2)] počet nenulových X_i
[y = sum(vek.o-vek.m > 2)] počet kladných X_i
[prop.test(y,n,correct=FALSE)] bez Yatesovy korekce
[prop.test(y,n,correct=TRUE)] s Yatesovou korekcí

párový Wilcoxonův test

(silnější předpoklad, než u znaménkového testu)

- nutné **spojité a symetrické** rozdělení $X_i = U_i - V_i$
- opět vyloučíme případy $U_i = V_i$ (tj. $X_i = 0$)
- určíme pořadí R_i^+ hodnot $|X_i| = |U_i - V_i|$
- W součet těch pořadí, kde bylo $U_i > V_i$ (tj. $X_i > 0$)

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

- pod odmocninou bývá ještě oprava na výskyt shodných hodnot, která jmenovatele poněkud zmenší
[wilcox.test(vyska.o,vyska.m,mu=10,paired=TRUE)]
- pro malá n se čitatel zpravidla přibližuje o $1/2$ k nule: (všimněte si zkrácených názvů parametrů – jednoznačnost!)
[wilcox.test(vyska.o,vyska.m,m=10,p=TRUE,cor=FALSE)]

příklad: porovnání dvou metod učení nazpaměť

- u devíti osob provedeno porovnávání dvou způsobů předávání informace (poslouchání, čtení)
- rozhodnout, zda je mezi oběma způsoby rozdíl
- H_0: rozdělení U a V stejná, tedy populační medián $X = U - V$ je roven 0
- znaménkový test s Yatesovou korekcí (málo pozorování):

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

$$z = \frac{|5 - 8/2| - 0{,}5}{\sqrt{8/4}} = 0{,}3536 \qquad p = 72{,}4\ \%$$

u_i	90	86	72	65	44	52	46	38	43
v_i	85	87	70	62	44	53	42	35	46
x_i	5	-1	2	3	0	-1	4	3	-3
r_i^+	8	1,5	3	5	–	1,5	7	5	5

příklad: porovnání dvou metod učení nazpaměť

- H_0 : populační medián rozdílů = 0
- nově předpokládáme symetrii
- Wilcoxonův test:

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

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

$$z = \frac{28 - 8 \cdot 9/4 - 1/2}{\sqrt{8 \cdot 9 \cdot 17/24}} = \frac{9{,}5}{\sqrt{51}} = 1{,}33$$

$$p = 18{,}3\ \%$$

- R dá $p = 18{,}1\ \%$, protože kromě opravy na spojitost bere ohled na shody (přesný výpočet dá $p = 19{,}5\ \%$)

dvouvýběrový t-test

(předpoklad **normálního rozdělení**)

- n_X nezávislých pozorování X, n_Y nezávislých pozorování Y
- tyto výběry musí být **nezávislé** (musí vyplynout ze způsobu pořízení dat)
- 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 nenormalita tolik nevadí)
- společný odhad rozptylu (vážený průměr odhadů z jednotlivých 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)

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

dvouvýběrový t-test

- $H_0 : \mu_X = \mu_Y$
  zamítnout ve prospěch alternativy
  - $H_1 : \mu_X \neq \mu_Y$ když $|T| \geq t_{n_X+n_Y-2}(\alpha)$
  - $H_1 : \mu_X > \mu_Y$ když $T \geq t_{n_X+n_Y-2}(2\alpha)$
  - $H_1 : \mu_X < \mu_Y$ když $T \leq -t_{n_X+n_Y-2}(2\alpha)$

  [t.test(hosi,divky,var.equal=TRUE)] nebo
  [t.test(vyska~Hoch,data=Vysky,var.equal=TRUE)]
- zamítáme-li H_0, říkáme, že rozdíl výběrových průměrů **je významný**
- pochyby o shodě rozptylů: Welchův test (modifikace t-testu)
  [t.test(hosi,divky,var.equal=FALSE)] (pro $\sigma_X \neq \sigma_Y$)
  [t.test(hosi,divky)] resp. [t.test(vyska~Hoch)] (pro $\sigma_X \neq \sigma_Y$)
- shodu rozptylů lze ověřit např. F-testem ($H_0 : \sigma_X = \sigma_Y$)
  [var.test(hosi,divky)]
- ověření normality nutně pro každý výběr zvlášť!

příklad: výšky dětí

	rozsah	průměr	výb. rozptyl
hoši	15	139,13	42,98
dívky	12	140,83	33,79

$$s^2 = \frac{15 - 1}{15 + 12 - 2} 42{,}98 + \frac{12 - 1}{15 + 12 - 2} 33{,}79 = 38{,}936$$

$$|t| = \frac{|139{,}13 - 140{,}83|}{\sqrt{38{,}936}} \sqrt{\frac{15 \cdot 12}{15 + 12}} = |-0{,}703| < 2{,}06 = t_{25}(0{,}05)$$

[shapiro.test(hosi)] $p = 80\ \%$
[shapiro.test(divky)] $p = 38\ \%$
[tapply(vyska,Hoch,shapiro.test)] (spočítá test pro oba výběry)
[var.test(hosi,divky)] $p = 70\ \%$
[t.test(hosi,divky,var.equal=TRUE)]

dvouvýběrový t-test a intervaly spolehlivosti
(poznámka na okraj)

- zpravidla platí
  - disjunktní intervaly spolehlivosti $\Rightarrow$ významný rozdíl
  - nevýznamný rozdíl průměrů $\Rightarrow$ překryv intervalů
  - rozdíl průměrů může být významný a současně se intervaly mohou překrývat
  - pokud každý z intervalů spolehlivosti obsahuje výběrový průměr druhého výběru, rozdíl průměrů není významný (nemusí platit v případě, kdy oba rozsahy výběru jsou do čtyř)
- příklad: váha v 24. týdnu dětí matek maturantek
  - 95% interval spolehlivosti pro hochy [kg]: (7,51; 8,25)
  - 95% interval spolehlivosti pro dívky [kg]: (6,98; 7,59)
  - intervaly se poněkud překrývají, přestože t-test dal: $t = 2{,}52$, $p = 1{,}5\ \%$, tedy na odpovídající 5% hladině je rozdíl významný

dvouvýběrový Wilcoxonův test (Mannův-Whitneyův)
(stačí **spojité rozdělení**)

- dva nezávislé výběry rozsahu n_X, n_Y
- spojitá rozdělení
- H_0: rozdělení jsou stejná, tedy i **mediány** jsou stejné
- za H_0 jsou výběry „dobře promíchané“
- určíme pořadí v rámci spojených výběrů
- kritický obor: průměrná pořadí se příliš liší
- 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

[wilcox.test(hosi,divky)]

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

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

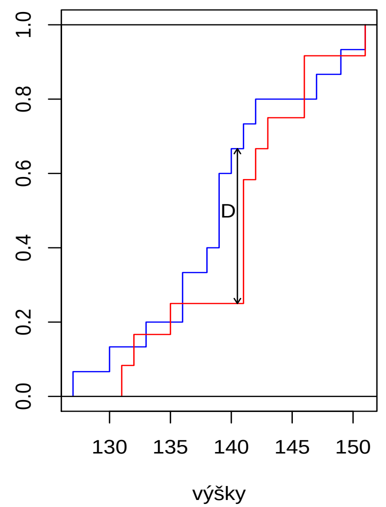
$$z = \frac{189 - 15 \cdot (15 + 12 + 1)/2}{\sqrt{15 \cdot 12(15 + 12 + 1)/12}} = -1{,}025$$

$p = 0{,}3055$

přesně: $p = 0{,}3149$

Kolmogorovův-Smirnovův test

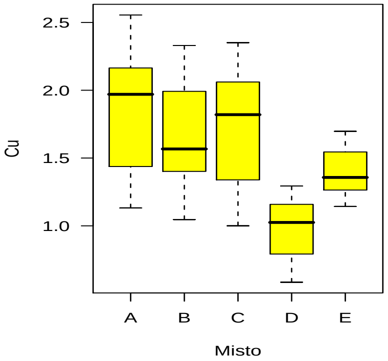
- porovná empirické distribuční funkce
- citlivý vůči všem neshodám (nejen co do populačního průměru či populačního mediánu)
- porovnání výšek hochů a dívek
- $D = \frac{10}{15} - \frac{3}{12} = 0{,}4167$
  $p = 19{,}7\ \%$

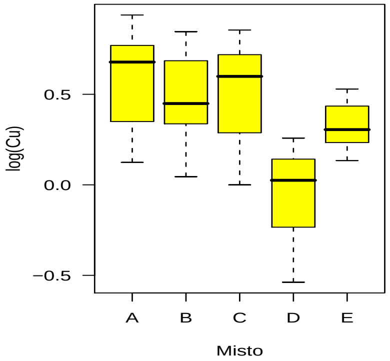


[ks.test(hosi,divky)]

motivační příklad pro analýzu rozptylu (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)

- $Y_{11}, \ldots, Y_{1n_1} \sim \mathrm{N}(\mu_1, \sigma^2)$ (první výběr, průměr $\bar{Y}_{1\bullet}$)
  $Y_{21}, \ldots, Y_{2n_2} \sim \mathrm{N}(\mu_2, \sigma^2)$ (druhý výběr, průměr $\bar{Y}_{2\bullet}$)
  ...
  $Y_{k1}, \ldots, Y_{kn_k} \sim \mathrm{N}(\mu_k, \sigma^2)$ (k-tý výběr, průměr $\bar{Y}_{k\bullet}$)
- **nezávislé** výběry (shodné rozptyly, normální rozdělení)
- $\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad (= \mu)$ $\quad \mathrm{H}_1$: neplatí H_0
- rozklad součtu čtverců (celkový průměr $\bar{Y}_{\bullet\bullet}$)

$$\sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{i\bullet})^2$$

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

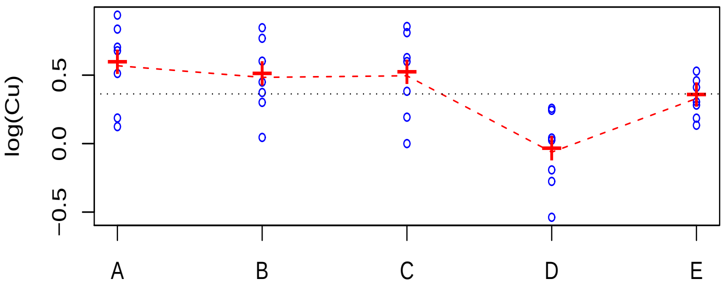
$$S_T = S_A + S_e$$
$$f_T = f_A + f_e$$
$$(n-1) = (k-1) + (n-k)$$

rozklad součtu čtverců

příklad játra (celkový průměr $\bar{y}_{\bullet\bullet} = 0{,}36$)

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

$$\sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{i\bullet})^2$$



tabulka analýzy rozptylu

$$\mathrm{H}_0 \text{ zamítnout, je-li } F_A = \frac{S_A/f_A}{S_e/f_e} \geq F_{f_A, f_e}(\alpha)$$

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

- S – součty čtverců, jejich rozklad
- f – počty stupňů volnosti
- S/f – průměrné čtverce
- F – F-statistika
- p – p-hodnota

příklad játra

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

$F = 5{,}862 > F_{4,30}(0{,}05) = 2{,}690$
na 5% hladině jsme **prokázali rozdíl**
[summary(aov(lnCu∼Misto,data=Med))]
nebo také
[anova(lm(lnCu∼Misto,data=Med))]

varianty zápisu modelu AR jednoduchého třídění

- **model** (měření = úroveň + „chyba“)

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

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

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

- $\mathrm{H}_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k$ (totéž, jako $\mu_1 = \mu_2 = \ldots = \mu_k$)
- pro $k = 2$ je $F_A = T^2$ (vztah s dvouvýběrovým t-testem)

ověření předpokladů

- **nezávislost**: dáno organizací (plánem) pokusu
  předpoklad nelze vynechat či nahradit
- **shoda rozptylů**: (vyvážený model málo citlivý na neshodu)
  - Leveneův test
    (vlastně ANOVA s $|Y_{it} - \mathrm{med}_t Y_{it}|$)
    $p = 64{,}8$ % [levene.test(lnCu,Misto)]
  - Bartlettův test
    (citlivý na splnění předpokladu o normálním rozdělení)
    $p = 45{,}3$ % [bartlett.test(lnCu,Misto)]
- **normální rozdělení**: (vyvážený model málo citlivý)
  test normality nutno uplatnit na rezidua $Y_{it} - \bar{Y}_{i\bullet}$
  $p = 6{,}8$ %
  [shapiro.test(resid(aov(lnCu Misto)))]
  nebo [shapiro.test(resid(lm(lnCu∼Misto)))]

mnohonásobná srovnání
(Tukeyův test, Kramerova verze)

- nutnost zachovat zvolenou hladinu testu i při současném rozhodování o řadě hypotéz
  (např. že $\mu_1 = \mu_2, \quad \mu_1 = \mu_3, \quad \mu_2 = \mu_3, \quad \ldots$)
- které dvojice úrovní faktoru (stř. hodnoty μ_i resp. efekty α_i) se liší?

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

kde $q_{k,n-k}(\alpha)$ je tabelovaná kritická hodnota

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

příklad játra

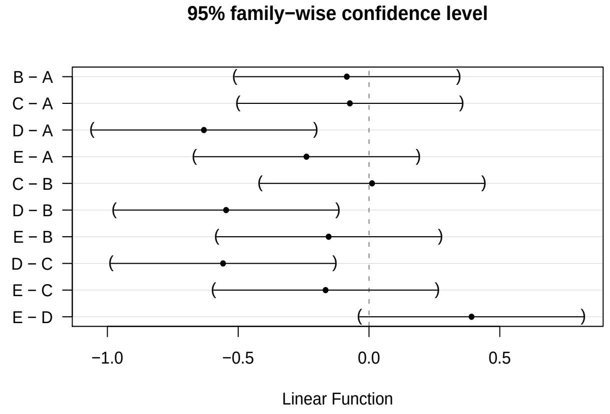
místo	počet	průměr	efekt	směr. odchylka
A	7	0,568	0,206	0,312
B	7	0,484	0,121	0,279
C	7	0,495	0,133	0,318
D	7	-0,063	-0,426	0,290
E	7	0,329	-0,034	0,144
celkem	35	0,363	0,000	0,104

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

$-0{,}063 + 0{,}428 = 0{,}365 \Rightarrow$ na 5% hladině se místa D s nejmenším průměrem liší všechna místa s průměry aspoň 0,365, tedy místa A, B, C, nikoliv E
[TukeyHSD(aov(lnCu~Misto,data=Med))]

příklad játra

funkce [TukeyHSD(aov(lnCu~Misto,data=Med))]
dá tabulku porovnání všech dvojic
pomocí knihovny Rcmdr dostaneme také graf



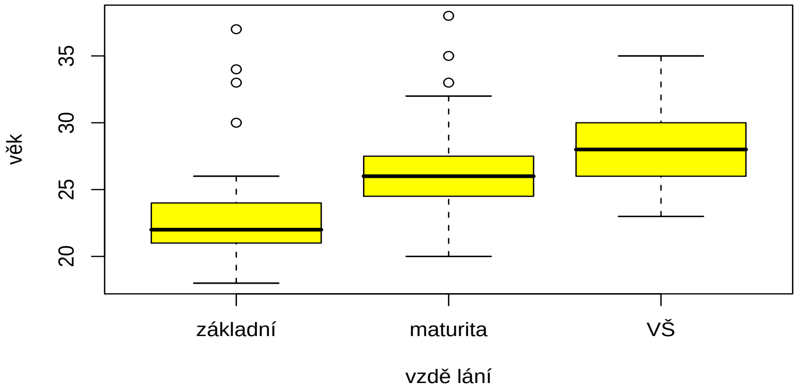
Kruskalův-Wallisův text
(neparametrický test)

- zobecnění dvouvýběrového Wilcoxonova testu (použije opět pořadí místo původních hodnot)
- předpoklady:
  - k nezávislých výběrů
  - spojitá rozdělení
- H_0: rozdělení jsou stejná (tedy i mediány jsou stejné)
- 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



je patrná nesymetrie, zejména u základního vzdělání

příklad kojení – věk matek podle vzdělání

vzdělání	n_i	průměrný věk	střední chyba	součet pořadí	průměrné 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		4950	50,00

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

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

[kruskal.test(vek.m~Vzdelani,data=Kojeni)]

motivační příklad diety: váhové přírůstky za danou dobu

vrh	dieta A	B	C	D	průměr
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
průměr	8,56	8,06	10,28	10,58	9,370

- $r = 4$ ošetření (pevné efekty, zvolili jsme je sami)
- $k = 5$ vrhů (náhodné efekty, zvolila je náhodně příroda)
- jsou patrné rozdíly mezi průměry pro jednotlivá ošetření i pro jednotlivé vrhy
- kdyby byly jen dvě diety ($r = 2$), použili bychom párový test

náhodné bloky

- zobecnění párových testů na r-tice
- **náhodný blok**
  - homogenní skupina r objektů
  - počet objektů ve skupině = počet ošetření nebo jeho násobek
  - ošetření se přiřadí uvnitř bloku **náhodně** (každému ošetření stejný počet objektů)
- bloky – náhodné efekty $A_i \sim \mathrm{N}(0, \sigma_A^2)$ (vliv bloku)
  ošetření – pevné efekty β_j $(\sum_{j=1}^r \beta_j = 0)$ (vliv ošetření)

  $$Y_{ij} = \mu + A_i + \beta_j + E_{ij}, \quad E_{ij} \sim \mathrm{N}(0, \sigma^2), \quad j = 1, \ldots, r, \quad i = 1, \ldots, k$$

  předpokládá se **aditivní** vliv, symbolicky zapisovaný $A + B$

náhodné bloky

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

  $$S_T = S_A + S_B + S_e$$

- vliv dvou **faktorů**
  - A – náhodný: nastavuje příroda, při opakování pokusu budou úrovně jiné
  - B – pevný: nastavuje experimentátor, při opakování pokusu budou úrovně stejné
  - rozhodování zda A je pevný nebo náhodný efekt závisí na cíli výzkumu, na interpretaci

příklad diety

- tabulka ANOVA

variabilita	S	f	S/f	F	p
vrhy	91,932	4	22,983	(22,26)	(<0,0001)
dieta	23,322	3	7,774	7,53	0,0043
reziduální	12,388	12	1,032	-	-
celk.	127,642	19	-	-	-

- na 5% hladině jsme prokázali rozdíl mezi dietami ($p = 0{,}4$ %)
- variabilita mezi vrhy je také průkazná ($p < 0{,}1$ %)
- [summary(aov(prirustek~Error(Vrh)+Dieta,data=Mysi))]
- pro takto jednoduchý model vyjde tabulka stejně i když považujeme faktor A za pevný (nenáhodný); porovnáváme pak konkrétních pět vrhů, vrhy nechápeme jako vzorek všech možných vrhů

příklad diety

- kdybychom **nesprávně** nevzali v úvahu závislost některých pozorování způsobenou náhodnými bloky (vrhy), dostali bychom model ANOVA jednoduchého třídění

variabilita	S	f	S/f	F	p
dieta	23,332	3	7,774	1,193	0,344
reziduální	104,320	16	6,520	-	-
celk.	127,642	19	-	-	-

- [summary(aov(prirustek~Dieta,data=Mysi))]
- porovnání se správnou tabulkou analýzy rozptylu

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

Friedmanův test, zobecnění znaménkového testu

(neparametrický test, bez předpokladu normality)

- 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 (nemusí být normální)
- $\mathrm{H}_0 : \beta_1 = \ldots = \beta_r$ (nezávisí na ošetření)
- urči pořadí v rámci každého bloku (řádku) R_{ij}
- za hypotézy je v každém řádku náhodná permutace čísel $1,\ldots,r$, součty ve sloupcích (pro ošetření) jsou 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

[friedman.test(prirustek~Dieta|Vrh,data=Mysi)]

	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

	dieta			
vrh	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

$k = 5$

$r = 4$

$$Q = \frac{12}{5 \cdot 4 \cdot 5}\left(9^2 + 7^2 + 17^2 + 17^2\right) - 3 \cdot 5 \cdot 5 = 9{,}96$$

$Q > \chi^2_3(0{,}05) = 7{,}8147$

$p = 0{,}0189$

dvojné třídění s interakcemi

opět **normální rozdělení**

- vliv dvou faktorů nemusí být aditivní ($1 \le t \le T$)

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

- 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), pak dvojné třídění bez interakcí (s opakováním pro $T > 1$)

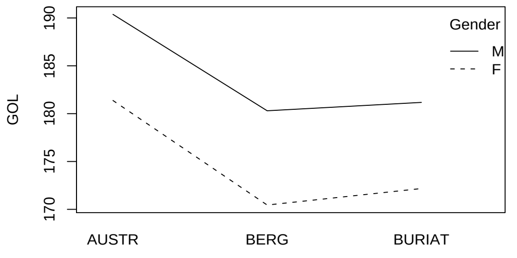
testy ve dvojném třídění

- $\mathrm{H}_{AB} : \gamma_{ij} = 0$ (aditivita obou faktorů)
  vliv úrovně faktoru A je stejný při všech úrovních faktoru B
  vliv úrovně faktoru B je stejný při všech úrovních faktoru A
- $\mathrm{H}_A : \alpha_i = 0$ (faktor A nemá vliv)
- $\mathrm{H}_B : \beta_j = 0$ (faktor B nemá vliv)
- pokud zamítneme H_{AB}, nemá smysl testovat $\mathrm{H}_A, \mathrm{H}_B$, neboť prostřednictvím interakcí oba faktory vliv mají
- v takovém případě je lépe přejít k modelu jednoduchého třídění s kombinovanými úrovněmi

příklad Howells

- lebky exhumované na třech místech (A)
- lebky jsou rozlišovány podle pohlaví (B)
- měříme největší délku mozkovny GOL

[anova(lm(gol~Gender*Popul))] nebo
[anova(lm(gol~Gender+Popul+Gender:Popul))]



$p_{AB} = 0{,}8872$

příklad Howells (GOL)

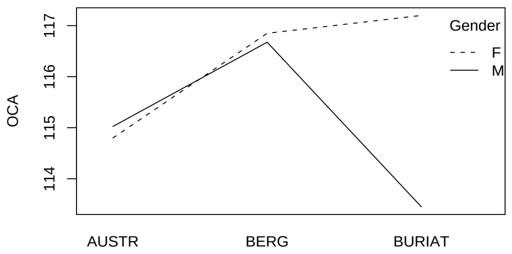
pohlaví	místo	n_{ij}	$\bar{y}_{ij}$	s_{ij}
M	Berg	40	180,300	7,293
F	Berg	40	170,450	6,641
M	Austrálie	40	190,375	5,555
F	Austrálie	40	181,375	6,632
M	Sibiř	40	181,175	6,468
F	Sibiř	40	172,175	5,228

variabilita	S	f	S/f	F	p
místa	5242,1	2	2621,1	65,2	$<0{,}0001$
pohlaví	5170,8	1	5170,8	128,6	$<0{,}0001$
interakce	9,6	2	4,8	0,1	0,8872
reziduální	9410,6	234	40,2		
celková	19833,2	239			

příklad Howells

- lebky exhumované na třech místech (A)
- lebky jsou rozlišovány podle pohlaví (B)
- měříme týlní úhel OCA

[anova(lm(oca~Gender*Popul))] nebo [anova(lm(oca~Gender+Popul+Gender:Popul))]



$p_{AB} = 0{,}0222$

příklad Howells (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

variabilita	S	f	S/f	F	p
místa	150,908	2	75,454	3,05	0,0493
pohlaví	91,267	1	91,267	3,69	0,0560
interakce	191,608	2	95,804	3,87	0,0222
reziduální	5789,550	234	24,742		
celková	6223,333	239			

porovnání populačních měr polohy

rozdělení	normální	spojité
populační parametr (o čem je hypotéza)	populační průměr	populační medián (distribuční funkce)
jeden výběr	jednovýběrový t-test	jednovýběrový 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
výběr r-tic	analýza rozptylu náhodné bloky	Friedman

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
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - 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 veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelační koeficient
(zavedení **výběrového** korelačního koeficientu)

- (populační) korelační koeficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ (zaveden na obr. 78)
  - $|\rho_{XY}| \leq 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy} (zaveden na obr. 34)

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

  - odhaduje ρ_{XY}
  - přesnost odhadu závisí na n
- alternativní označení: **Pearsonův** korelační koeficient, momentový korelační koeficient, [correlation coefficient]

dokazování závislosti X, Y

- k prokázání závislosti nutno **normální** rozdělení (X, Y)
- $H_0 : \rho_{XY} = 0$ se na hladině α zamítá:

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

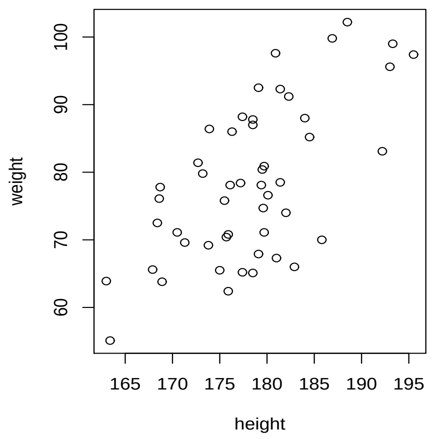
  (r je dost daleko od nuly)
- **Spearmanův** korelační koeficient
  - měří sílu **monotonní** závislosti
  - založen na **pořadích** R_i, Q_i hodnot X_i, Y_i

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

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

závislost váhy na výšce u mužů
data: Policie

[plot(weight∼height)]



[cor.test(weight,height)]

$r = 0{,}648$

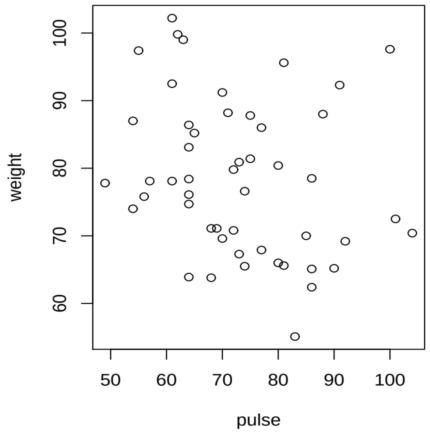
$t = 5{,}814$

$p < 0{,}001$

závislost váhy na pulsu u mužů

data: Policie

[plot(weight∼pulse]



[cor.test(pulse,weight)]

$r = -0{,}245$

$t = -1{,}752$

$p = 8{,}6\ \%$

Fisherova z-transformace

(přiblíží chování výběrového korelačního koeficientu r normálnímu rozdělení)

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

test shody dvou nezávisle odhadovaných korel. koeficientů

příklad **Kojeni**: výška rodičů chlapců a dívek

- dívky: $r_1 = 0{,}279$, $n_1 = 50$, $z_1 = \frac{1}{2}\ln\frac{1+0{,}5687}{1-0{,}5687} = 0{,}286$
- hoši: $r_2 = 0{,}150$, $n_2 = 49$, $z_2 = \frac{1}{2}\ln\frac{1+0{,}150}{1-0{,}150} = 0{,}151$
- test $H_0 : \rho_1 = \rho_2$ (odhady r_1, r_2 jsou **nezávislé!**)

$$z = \frac{0{,}286 - 0{,}151}{\sqrt{\frac{1}{50-3} + \frac{1}{49-3}}} = 0{,}650.$$

srovnej s kritickou hodnotou $z(0{,}05/2) = 1{,}960$, $p = 51{,}6\ \%$

interval spolehlivosti pro ρ

opět potřebujeme normální rozdělení (X, Y)

- ve dvou krocích:
  - interval spolehlivosti pro $\zeta = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$
  - pomocí inverzní transformace pak int. spol. pro ρ
- interval spolehlivosti součástí funkce cor.test()
- náš příklad:

skupina	r (bodový odhad)	95% int. spol. pro ρ	p
dívky	0,279	(0,000; 0,517)	5,01 %
hoši	0,150	(−0,137; 0,414)	30,3 %

- u chlapců nelze prokázat na 5% hladině závislost
- u děvčat je závislost na 10% hladině průkazná, na 5% hladině těsně nikoliv (interval spolehlivosti je jen přibližný!)

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 otců této výšky 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
- průměrné výšky synů nereprodukují celou odchylku výšky otce od průměru, je tu návrat k průměru (regrese)

regresní přímka

- **předpokládaná** závislost střední hodnoty Y na nenáhodné x:

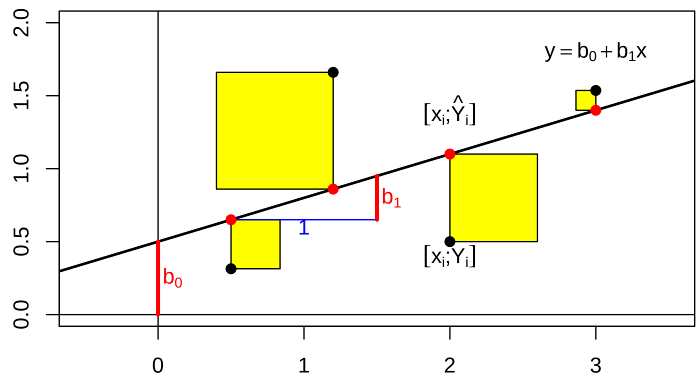
$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

- k daným $x_1, \dots, x_n$ zjistíme $Y_1, \dots, Y_n$
- předpoklady:
  - **nezávislá** pozorování $Y_1, \dots, Y_n$
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (potřebné až pro testy, normalitu nelze ověřovat testováním přímo $Y_1, \dots, Y_n$!)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

  minimalizovat přes β_0, β_1 výraz $\displaystyle\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$
- odhady označíme b_0, b_1

metoda nejmenších čtverců

odhadovaná závislost:	$y = \beta_0 + \beta_1 \cdot x$	(populace)
odhad závislosti:	$y = b_0 + b_1 \cdot x$	(výběr)
i-tá vyrovnaná hodnota	$\hat{Y}_i = b_0 + b_1 x_i$	(výběr)
i-té reziduum	$U_i = Y_i - \hat{Y}_i$	(výběr)
celková plocha čtverců:	$S_e = \sum_{i=1}^{n} U_i^2$	(výběr)



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

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

- **reziduální rozptyl**

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

alternativní formulace

- uvažovanou závislost lze psát ve tvaru

$$\begin{aligned} Y_i &= (\beta_0 + \beta_1 \bar{x}) + \beta_1 (x_i - \bar{x}) + E_i \\ &= \beta_0^* + \beta_1 (x_i - \bar{x}) + E_i \end{aligned}$$

- β_0^* vyjadřuje střední úroveň vysvětlované proměnné Y při průměrné hodnotě nezávisle proměnné x
- β_1 vyjadřuje citlivost, s jakou reaguje střední hodnota vysvětlované proměnné Y na jednotkovou odchylku nezávisle proměnné x od jejího průměru $\bar{x}$
- E_i vyjadřuje náhodnou složku i-tého pozorování, $E_i \sim \mathrm{N}(0, \sigma^2)$
- odhadem závislosti je (b_1 je stejné jako při klasickém vyjádření)

$$\hat{Y}_i = \bar{Y} + b_1 (x_i - \bar{x})$$

prokazování závislosti

- modelujeme závislost $\mathrm{E}\,Y$ na x pomocí $\mathrm{E}\,Y = \beta_0 + \beta_1 x$
- nezávislost $y = \beta_0 + \beta_1 x$ na x znamená $\beta_1 = 0$
- hypotézu $\mathrm{H}_0 : \beta_1 = 0$ testujeme pomocí statistiky

$$T = \frac{b_1}{\mathrm{S.E.}(b_1)}$$

- hypotézu zamítáme, je-li $|T| \geq t_{n-2}(\alpha)$
  tj. je-li příslušná p-hodnota $\leq \alpha$
- pokud H_0 zamítneme, říkáme, na hladině α je **závislost průkazná**

koeficient determinace
[coefficient of determination]

- podíl variability Y vysvětlené uvažovanou závislostí
  (jakou část variability Y se podařilo závislostí na x vysvětlit)
-

$$\begin{aligned} R^2 &= \frac{\text{variabilita vysvětlená}}{\text{variabilita vysvětlovaná}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita vysvětlovaná}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2} \end{aligned}$$

- R^2 je bezrozměrné číslo, často vyjádřeno v procentech
- R^2 ukazuje, zda má smysl předpovídat pomocí regrese

příklad závislost procenta tuku na výšce
data: Policie

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: $\hat{Y}_i = -53{,}870 + 0{,}379 x_i$
- $\widehat{\text{fat}} = -\,53{,}870 + 0{,}379 \cdot \text{height}$
- závislost procenta tuku na výšce je na 5% hladině průkazná
- na každý centimetr výšky *v průměru* přibude 0,379 procentního bodu tuku
- [summary(lm(fat~height))]

tabulka analýzy rozptylu

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)		

- $s^2 = 48{,}22$
-

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

- závislostí na výšce jsme vysvětlili jen 13,5 % variability procenta tuku
- [anova(lm(fat~height))]

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{\mathrm{E}\, Y_i} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}\big(0, \sigma^2\big)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

interpretace

- 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**

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

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

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

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

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

- $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$ (chování Y nezávisí ani na x ani na v)

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

- p-hodnota tohoto testu bývá uváděna spolu s R^2

testy o přínosu jednotlivých regresorů

- model $y = \beta_0 + \beta_1 x + \beta_2 v$
- $\mathrm{H}_0 : \beta_2 = 0$
  k vysvětlení chování Y stačí x, tj. $y = \beta_0 + \beta_1 x$

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

- $\mathrm{H}_0 : \beta_1 = 0$
  k vysvětlení chování Y stačí v, tj. $y = \beta_0 + \beta_2 v$

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

- $\mathrm{H}_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

příklad: závislost procenta tuku na výšce a váze

data: Policie

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

- [summary(lm(fat~height+weight))]
- 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
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

tabulka analýzy rozptylu

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)		

- $R^2 = 1833{,}11/2676{,}95 = 1 - 843{,}85/2676{,}95 = 0{,}685$
- závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- $s^2 = 17{,}95$
- na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

regresní diagnostika

zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?
b) je **rozptyl** všude **stejný**?
c) je přiměřeně splněn předpoklad o **normálním rozdělení**?
d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c), např. logaritmování závisle proměnné
- [plot(lm(fat~height+weight))]

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- 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
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

multinomické rozdělení

- v dílčím pokusu k možných výsledků (jevů) $A_1, \dots, A_k$ neslučitelné jevy, sjednocení všech je jev jistý
- π_j je pst, že vyjde A_j $(\pi_1 + \pi_2 + \dots + \pi_k = 1)$
- n **nezávislých** dílčích pokusů (opakování)
- N_j – počet dílčích pokusů, kdy nastalo A_j
- $(N_1, \dots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \dots, \pi_k$
- **pravděpodobnost** toho, že $N_1 = n_1, \dots, N_k = n_k$

$$\mathrm{P}(N_1 = n_1, \dots, N_k = n_k) = \frac{n!}{n_1! \dots n_k!} \pi_1^{n_1} \dots \pi_k^{n_k}$$

souvislost s binomickým rozdělením

- pro $k = 2$ jsou v dílčím pokusu jen dva možné výsledky, binomické rozdělení je speciálním případem multinomického

$$\mathrm{P}(N_1 = n_1, N_2 = n_2) = \frac{n!}{n_1! n_2!} \pi_1^{n_1} \pi_2^{n_2}$$

je totéž jako (platí přece $n_1 + n_2 = n$)

$$\mathrm{P}(N_1 = n_1) = \binom{n}{n_1} \pi_1^{n_1} \pi_2^{n-n_1}$$

- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tedy

$$\boxed{N_j \sim \mathrm{bi}(n, \pi_j), \quad \mathrm{E}\, N_j = n\pi_j}$$

vlastnost χ^2 (chí-kvadrát)

(X^2 – velké χ^2)

- platí pro velká n, např. pokud $n\pi_j \geq 5$ pro všechna j

$$\boxed{X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j}} \text{ má přibližně rozdělení } \chi^2_{k-1}$$

- **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 $\mathrm{E}\, N_j = n\pi_j^0$:
- H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(\alpha)$, $\boxed{X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}}$
- N_j – **experimentální** četnosti, $n\pi_j^0$ – **očekávané** (teoretické) četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti (měří jejich neshodu)

počty studentů biologie narozených v jednotlivých měsících

nulová hypotéza: děti se rodí během roku **rovnoměrně**

[chisq.test(nn,p=c(31,28,31,30,31,30,31,31,30,31,30,31)/365)]

měsíc	n_j	$n\pi_j^0$	přínos k chí-kvadrát
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

$X^2 = 7{,}4115 < \chi^2_{12-1}(0{,}05) = 19{,}675$ $\quad p = 76{,}5\ \%$

příklad: reprezentativnost výběru
(porovnat procenta v populaci a výběru **nestačí**)

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 % (to určí H_0)
- ve vzorku pacientů byly počty osob s krevními skupinami 0, A, B a AB po řadě 56, 72, 54, 18 (tedy $n = 200$)
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?

$$\chi^2 = \frac{(56-70)^2}{70} + \frac{(72-70)^2}{70} + \frac{(54-40)^2}{40} + \frac{(18-20)^2}{20}$$
$$= 7{,}96 \qquad p = 4{,}7\ \%$$

- výběr **nelze** považovat za reprezentativní
- při polovičních četnostech ve výběru (28, 36, 27, 9) by vyšlo $\chi^2 = 3{,}98$, $p = 26{,}4\ \%$ (**lze** považovat za reprezentativní)

příklad: barva květů a tvar pylových zrnek
segregace dvou typů genů (C. R. Rao: Lineární metody statistické indukce ..., str. 439)

- barva květů – purpurová : červená v poměru 3 : 1 (dáno)
- tvar pylu – oválný : kulatý v poměru 3 : 1 (dáno)
- platí-li nulová hypotéza (H_0 : jde o **nezávislou** segregaci), pak čtyři možné kombinace musí být v poměru 9 : 3 : 3 : 1

barva tvar	pupurová oválný	červená oválný	purpurová kulatý	červená kulatý	celkem
n_j	296	27	19	85	427
o_j	3843/16	1281/16	1281/16	427/16	427
$\frac{(n_j-o_j)^2}{o_j}$	12,97	35,17	46,57	127,41	222,12

$$\chi^2 = 222{,}12 > \chi^2_3(0{,}05) = 7{,}81$$

- nezávislost jsme **zamítli**

příklad: barva květů a tvar pylových zrnek

- co způsobilo zamítnutí hypotézy?

barva	purpurová	červená	celkem
oválný tvar	296	27	323
kulatý tvar	19	85	104
celkem	315	112	427

- jsou barvy v očekávaném poměru 3 : 1?
[chisq.test(c(315,112),p=c(3/4,1/4))]

$$\chi^2 = 0{,}3443 \qquad p = 55{,}7\ \%$$

- jsou tvary v očekávaném poměru 3 : 1?

$$\chi^2 = 0{,}0945 \qquad p = 75{,}9\ \%$$

- důvodem zamítnutí určitě závislost

složená nulová hypotéza (hypotéza o struktuře)

- hypotéza určuje vztahy mezi pravděpodobnostmi $\pi_1, \ldots, \pi_k$ některé parametry zůstávají volné, je třeba je odhadnout
- příklad antigen: (Hardy-Weinberg equilibrium) model pro fenotypy AA, Aa, aa

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

- neurčený parametr θ – pravděpodobnost alely A
- jsou zjištěné četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$ v souladu s modelem, tj. s H-W rovnováhou?

- odhad θ maximalizací *logaritmické věrohodnostní funkce*

$$\begin{aligned}\ell(\theta) &= \ln(\mathrm{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3))\\ &= \ln\left(c_1\left(\theta^2\right)^{n_1}(2\theta(1-\theta))^{n_2}\left((1-\theta)^2\right)^{n_3}\right)\\ &= c_2 + (2n_1 + n_2)\ln\theta + (n_2 + 2n_3)\ln(1-\theta)\end{aligned}$$

$$\hat{\theta} = \frac{2\cdot N_1 + N_2}{2n} \quad \left(= \frac{2\cdot 18 + 17}{82} = 0{,}646\right)$$

- obecně se H_0 zamítá, pokud (θ má q nezávislých složek)

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

- příklad antigen: $\chi^2 = 0{,}355 < \chi^2_{3-1-1}(0{,}05) = 3{,}84$
  $p = 55{,}1\ \%$ hypotézu na 5% hladině nezamítáme

nezávislost **nominálních** znaků

- 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

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

- charakteristika nezávislosti: z **marginálních** pstí jevů A_i, B_j dokážeme rekonstruovat **sdružené psti** jevů $A_i \cup B_j$

test nezávislosti dvou kvalitativních znaků

- teoretické četnosti (protějšek N_{ij}) –
  četnosti, které **v průměru očekáváme**, platí-li hypotéza

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

- H_0 : znaky jsou **nezávislé**

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

- nezávislost se zamítá pokud $X^2 \geq \chi^2_{(r-1)(c-1)}(\alpha)$
- stupně volnosti $n - 1 - q = r\cdot c - 1 - (r-1) - (c-1) =$
  $r\cdot c - r - c + 1 = (r-1)(c-1)$
- musí být $o_{ij} \geq 5\ \forall\ (i,j)$ (tj. pro všechny dvojice)

příklad: kouření u mužů

data: Ichs

empirické sdružené a marg. četnosti

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

očekávané sdružené a marg. četnosti

vzdělání	zákl.	odb.	mat.	VŠ	celk.
nekuřák	24,3	61,4	61,9	49,4	197
bývalý k.	15,4	39,0	39,3	31,3	125
kuřák	9,7	24,6	24,8	19,8	79
silný k.	67,6	170,9	172,1	137,4	548
celkem	117	296	298	238	949

$$\chi^2 = \frac{(14-24{,}3)^2}{24{,}3} + \ldots + \frac{(106-137{,}4)^2}{137{,}4} = 38{,}68$$

$$f = (4-1)(4-1) = 9$$

$$p < 0{,}0001$$

[chisq.test(matrix(c(14,11,14,8,55,28,24,189,55,44,24,175,73,42,17,106),nr=4,nc=4))]

závislost jsme na 5% hladině prokázali

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

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

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

$$\chi^2 = \frac{(121 - 353 \cdot 239/717)^2}{353 \cdot 239/717} + \ldots = 11{,}742 > \chi^2_3(0{,}05) = 7{,}815$$

nejm. teoretická četnost: $353 \cdot 63/717 = 31{,}02 > 5$, $p = 0{,}8\ \%$

McNemarův test (test symetrie)

nezaměňovat s testem nezávislosti!

- **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
- **nulová hypotéza**: pravděpodobnosti možných hodnot znaku jsou **stejné** za obojích okolností (před ošetřením i po něm)

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

- 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

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

- $\chi^2_3(0{,}05) = 7{,}8147$, $\quad p = 36{,}0\ \%$
- rozdíl mezi sezónami jsme neprokázali
- [mcnemar.test(matrix(c(4,7,1,3,21,15,3,11,35),3,3))]

čtyřpolní tabulka
znovu test nezávislosti či homogenity

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

- speciální případ kontingenční tabulky pro $r = c = 2$
- test nezávislosti i test homogenity
  statistiku lze upravit na pohodlnější vyjádření

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

zamítá se pro $X^2 \geq \chi_1^2(\alpha) = z(\alpha/2)^2$

případ malých četností

- je-li některá očekávaná četnost malá, pak lze u čtyřpolní tabulky použít upravený postup: **Yatesova korekce**

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

- **Fisherův exaktní** test počítá přímo dosaženou hladinu p
- pro tabulku s velkými četnostmi je výpočet Fisherova testu výpočetně náročný (paměťové nároky, trvání výpočtu)
- existuje zobecnění Fisherova testu i pro větší tabulky, než je čtyřpolní

příklad **hraboš**

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

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

$$\chi^2 = \frac{515(4 \cdot 473 - 11 \cdot 27)^2}{15 \cdot 500 \cdot 31 \cdot 484} = 11{,}643, \quad p = 0{,}06\ \%$$

- [chisq.test(matrix(c(4,11,27,473),2,2),correct=FALSE)]

příklad **hraboš**

- nejmenší očekávaná četnost: $15 \cdot 31/515 = 0{,}9 < 5$
- **Yates**: $\chi^2 = 8{,}187$ $p = 0{,}42\ \%$
  [chisq.test(matrix(c(4,11,27,473),2,2))]
- **Fisherův test**: $p = 0{,}92\ \%$
  [fisher.test(matrix(c(4,11,27,473),2,2))]
- na 5% hladině závislost **prokázána**
- **vyskytují se dvojí cizopasníci se stejnou pstí?**
  (zcela jiná otázka, než na nezávislost)
- odpověď dá McNemarův test:

$$\chi^2 = \frac{(11-27)^2}{11+27} = 6{,}7368, \quad p = 0{,}94\ \%$$

[mcnemar.test(matrix(c(4,11,27,473),2,2),correct=FALSE)]

příklad: barva květů a tvar pylových zrnek
(jiný postup)

- připomeňme data

barva	purpurová	červená	celkem
oválný tvar	296	27	323
kulatý tvar	19	85	104
celkem	315	112	427

- kdybychom neznali předem teoretické poměry u barvy a tvaru, použijeme běžný postup pro čtyřpolní tabulku

$$\chi^2 = \frac{427 \cdot (296 \cdot 85 - 19 \cdot 27)^2}{315 \cdot 112 \cdot 323 \cdot 104} = 218{,}9$$

- porovnat s $\chi_1^2(0{,}05) = 3{,}84$ a nikoliv s $\chi_3^2(0{,}05) = 7{,}81$
- nyní marginální psti odhadujeme, kdežto v 11. přednášce (obrázek 203) jsme je znali

jak statistiku použijeme

- 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 (např. 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 rozdělení (Kolmogorov-Smirnov)
- **prokázat závislost**
  - obě spojité (korelační koeficient, regrese)
  - spojitá na kvalitativními (ANOVA)
  - obě kvalitativní (kontingenční tabulka)
  - **predikce** spojité veličiny na spojitých či kvalitativních (regrese)

výběr metody

- jakou úlohu řešíme?
- jsou výběry nezávislé?
  - z organizace pokusu
- lze předpokládat normální rozdělení?
  - 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)
  - u regrese lze ověřit pomocí Breuschova-Paganova testu

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