Statistika

(MD360P03Z, MD360P03U)
ak. rok 2011/2012

Karel Zvára

karel.zvara@mff.cuni.cz
http://www.karlin.mff.cuni.cz/∼zvara

(naposledy upraveno 9. ledna 2012)

literatura

- K. Zvára: Biostatistika, Karolinum Praha, 1998, 2000, 2001, 2003, 2006, 2008
- Z. Pavlík, K. Kühnl: Úvod do kvantitativních metod pro geografy, SPN Praha, 1981
- T. H. Wonnacot, R. J. Wonnacot: Statistika pro obchod a hospodářství, Victoria Publishing Praha, 1992
- slajdy přednášky na adrese http://www.karlin.mff.cuni.cz/∼zvara
- může dojít k drobným úpravám slajdů po přednášce i před ní

cvičení, zápočet, zkouška

- cvičení v počítačové učebně PUA (suterén Albertov 6) (začátek v týdni od 10. října)
- MS Excel [funkce Excelu]
- volně šiřitelný program R (http://cran.r-project.org/) [funkce R]
- (aktivní účast na cvičení, maximálně dvě absence) & (napsání zápočtového testu) ⇒ zápočet
- obsah cvičení více přizpůsoben studovanému oboru
- přednášky jsou formulovány obecněji
- zkouška kombinovaná (písemná s počítačem i ústní), zápočet **musí** zkoušce **předcházet**; přihlašování ke zkoušce přes SIS

přehled témat

- popisná statistika (měřítka, charakteristiky polohy, variability, souvislost znaků)
- statistika v geografických/demografických/sociálních vědách
- pravděpodobnost (základní kombinatorické pojmy, klasická definice, podmíněná pravděpodobnost, nezávislost)
- náhodná veličina (rozdělení, střední hodnota, rozptyl, hustota, distribuční funkce)
- důležitá rozdělení (normální, binomické, Poissonovo)
- statistické usuzování (populace a výběr, parametry a jejich odhady, interval spolehlivosti, volba rozsahu výběru)
- testování hypotéz (chyba 1. druhu, chyba 2. druhu, hladina významnosti testu, síla testu, p-hodnota)
- některé testy (o populačním průměru či průměrech, populačním podílu či podílech, nezávislosti, regresních koeficientech)
- regrese, kontingenční (čtyřpolní) tabulky

příklad statistického zjišťování I

- zjišťování se týká mužů středního věku
- v souboru je 80 kuřáků a 120 nekuřáků
- 85 mužů má oči modré, 25 hnědé, 90 jiné barvy
- 27 mužů má jen základní vzdělání, 44 neúplné střední, 65 maturitu, 64 vysokoškolské
- 22 se jich narodilo v roce 1942, 19 v roce 1943, 25 v roce 1944, ..., 18 v roce 1951
- hmotnosti jednotlivých mužů jsou 83, 92, ..., 63 kg
- výška jednotlivých mužů jsou 172, 176, ..., 178 cm
- Co mají tyto údaje společného? Čím se údaje v jednotlivých podskupinách liší? Souvisí kouření a vzdělání? Souvisí příjem se vzděláním? Souvisí váha s výškou? Je tato souvislost stejná, jako v zemi XY?

příklad statistického zjišťování II

- zjišťování se týká příjmů obyvatel
- hodnotíme hrubý příjem za rok
- přihlížíme k místu trvalého bydliště (velikost obce, který kraj)
- přihlížíme k vzdělání (druh, délka školní docházky)
- přihížíme k délce praxe v oboru
- přihlížíme k věku a pohlaví
- Co mají tyto údaje společného? Čím se údaje liší?

co a jak měříme (zjišťujeme)

- měříme na mnoha **statistických jednotkách** (osoba, domácnost, obec, okres, stát, pokusné pole ...)
- měříme (zjišťujeme) hodnoty statistických **znaků**
- zjištěnou hodnotu znaku vyjadřujeme ve zvoleném **měřítku** (stupnici)
- na jedné jednotce můžeme měřit několik znaků (to umožní vyšetřovat závislost)
- měříme na skupinách jednotek – **souborech**
- zajímají nás **hromadné** vlastnosti ve velkých souborech
- můžeme **porovnávat** vlastnosti znaku **mezi soubory**

měřítka

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (země původu, barva očí) jednoznačně dané hodnoty (úrovně znaku)
- **ordinální** (dosažené vzdělání, stupeň bolesti) jednoznačně dané hodnoty, možné hodnoty jsou *uspořádané*
- **intervalové** (teplota v Celsiově stupnici, rok narození) konstantní vzdálenosti mezi sousedními hodnotami, umístění nul je jen konvence;
  o *kolik* stupňů je je dnes tepleji, než bylo včera?
- **poměrové** (hmotnost, výška, HDP, počet obyvatel, věk) násobek zvolené jednotky
  nula = neexistence měřené vlastnosti
  kolikrát je A starší (vyšší ...) než B
  kolikrát je dnes tepleji? nedává smysl

měřítka (stručnější dělení)

- **kvalitativní**: nula-jedničkové, nominální, často i ordinální
- u kvalitativního měřítka se zpravidla udávají **četnosti** jednotlivých hodnot (kolikrát která hodnota nastala)
- **kvantitativní** (spojité): intervalové, poměrové, někdy ordinální (není spojité)
- hodnoty v kvantitativním měřítku – čísla
- zařazení znaku k určitému měřítku může záviset na účelu šetření (např. barva nominální pro biologa, přinejmenším ordinální pro fyzika)

veličina

- číselně vyjádřený výsledek měření (zjišťování)
- *hodnoty* znaků v intervalovém a poměrovém měřítku jsou husté – **spojitá veličina**
- *četnosti hodnot* znaků v nula-jedničkovém, nominálním (či ordinálním) měřítku – **diskrétní veličina**
- pro veličiny máme charakteristiky některých jejich hromadných vlastností (**charakteristiky polohy, variability, tvaru rozdělení**)
- charakteristiky (statistiky) mají jedním číslem vyjádřit danou vlastnost
- Kdo je vyšší – dvanáctiletí hoši nebo dvanáctileté dívky? (potřebujeme výšky **všech** dvanáctiletých hochů charakterizovat **jediným** číslem, má vyjádřit **úroveň** výšek, podobně pro dívky)

příklad: 100 hodů kostkou

počty puntíků na kostce coby různé obrázky – nominální znak

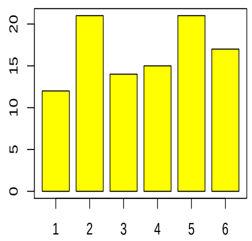
kostka A										kostka B									
4	2	5	6	3	1	1	2	2	2	1	4	6	2	3	2	6	1	5	2
2	4	5	3	1	1	3	5	5	5	5	6	5	5	6	4	2	4	5	6
4	3	2	5	5	5	2	2	5	2	3	6	3	6	5	6	1	3	5	1
2	6	5	5	2	3	6	6	4	6	6	6	2	1	1	2	6	3	2	3
5	4	1	4	2	2	4	5	2	5	4	4	1	6	6	2	6	3	2	6
5	5	3	3	5	3	6	6	6	5	2	6	1	2	6	1	5	5	6	5
3	5	4	5	1	1	4	3	2	4	6	6	5	1	6	6	6	1	2	6
1	2	4	6	6	3	4	6	1	2	6	2	5	6	2	6	6	5	6	4
6	6	1	2	6	2	4	3	2	3	6	1	2	6	2	1	6	6	6	6
1	1	6	5	2	6	4	4	6	3	6	5	1	5	6	6	1	6	6	6

hody kostkou jako hromadný jev

- chceme $n = 100$ zjištěných hodnot (počtů puntíků) vyjádřit názorně, aby vypovídaly o vlastnostech kostky
- n_j (absolutní) **četnost** [frequency] j-té hodnoty (kolikrát nastala)
- $f_j = \frac{n_j}{n}$ **relativní četnost** j-té hodnoty (lze vyjádřit v %) v jakém dílu měření nastala
- nutně platí $n = n_1 + n_2 + \ldots + n_k = \sum_{j=1}^{k} n_j$, $\sum_{j=1}^{k} f_j = 1$
- tabulka četností (absolutních, relativních)
- grafické vyjádření četností – **histogram** [histogram, barplot] (velikost plochy je úměrná četnosti)
- rozhodování o kvalitě kostky (zda je symetrická) je úlohou **statistické indukce** [inference] – bude později

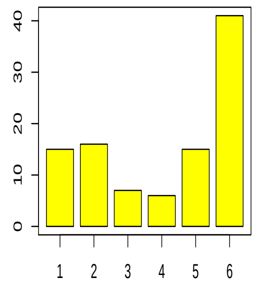
zpracování četností (kostka A)

j	n_j	$f_j = n_j/n$
1	12	0,12
2	21	0,21
3	14	0,14
4	15	0,15
5	21	0,21
6	17	0,17
	$n = 100$	1,00



zpracování četností (kostka B)

j	n_j	$f_j = n_j/n$
1	15	0,15
2	16	0,16
3	7	0,07
4	6	0,06
5	15	0,15
6	41	0,41
	$n = 100$	



příklad: věk 99 matek

99 zjištěných hodnot – soubor naměřených hodnot

26	35	21	25	27	24	24	30	23	18
35	21	25	26	26	19	29	22	21	27
26	30	28	28	27	29	27	26	21	23
24	21	28	25	34	24	21	28	25	28
22	26	32	22	32	25	21	25	24	32
24	22	31	33	23	30	26	27	25	24
24	23	25	23	26	28	24	25	25	26
28	28	22	23	20	20	21	31	24	21
29	28	26	38	20	23	25	37	33	23
27	23	21	25	21	33	22	29	21	

▸ Jdi k variační řadě

variační řada, pořadí

- $x_1, x_2, \ldots, x_n$ původní (neuspořádaná) data – hodnoty znaku uvedené v původním pořadí, bez ohledu na případná opakování
- **variační řada** (uspořádaný výběr) [sort(x)]

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

  data v měřítku aspoň ordinálním uspořádána tak, aby hodnoty neklesaly; proto **závorky u indexů**
- **pořadí** [rank] – umístění pozorování ve variační řadě; shodným hodnotám dáváme průměrné pořadí [rank(x)]
- příklad

x_j	22	15	17	15	21	13	18
pořadí R_j	7	2,5	4	2,5	6	1	5

- v Excelu má funkce RANK() poněkud jiný význam, lze použít opravu na shody (viz nápovědu pro RANK)

příklad: věk 99 matek – variační řada

uspořádaný soubor hodnot – variační řada

18	19	20	20	20	21	21	21	21	21
21	21	21	21	21	21	21	22	22	22
22	22	22	23	23	23	23	23	23	23
23	23	24	24	24	24	24	24	24	24
24	24	25	25	25	25	25	25	25	25
25	25	25	25	26	26	26	26	26	26
26	26	26	26	27	27	27	27	27	27
28	28	28	28	28	28	28	28	28	29
29	29	29	30	30	30	31	31	32	32
32	33	33	33	34	35	35	37	38	

Jdi k původním pozorováním Jdi k mírám variability

třídění, třídní četnosti

- spojitá veličina s velkým počtem naměřených hodnot
- obor hodnot rozdělíme na k nepřekrývajících se tříd (intervalů), nejlépe stejné délky (ale ne vždy je to praktické či možné)
- všechna pozorování z daného intervalu nahradíme zástupnou hodnotou (zpravidla středem intervalu) x_j^* $(x_1^* < \ldots < x_k^*)$
- zjistíme (**absolutní**) **četnosti** $n_1, \ldots, n_k$ jednotlivých tříd
- **kumulativní četnost** N_j udává počet hodnot v dané třídě a třídách předcházejících $(1 \leq j \leq k)$ [cumsum( )]

$$N_j = n_1 + n_2 + \ldots + n_j = \sum_{i=1}^{j} n_i$$

věk matek – třídní četnosti

$k = 7$

interval	x_j^*	n_j	$f_j = n_j/n$	N_j	N_j/n
do 20	19	5	0,051	5	0,051
21 až 23	22	27	0,273	32	0,324
24 až 26	25	32	0,322	64	0,646
27 až 29	28	19	0,192	83	0,838
30 až 32	31	8	0,081	91	0,919
33 až 35	34	6	0,061	97	0,980
36 a více	37	2	0,020	99	1,000
celkem	–	99	1,000	–	–

Jdi k histogramu věku matek Jdi k mírám polohy věku matek

grafické znázornění třídních četností

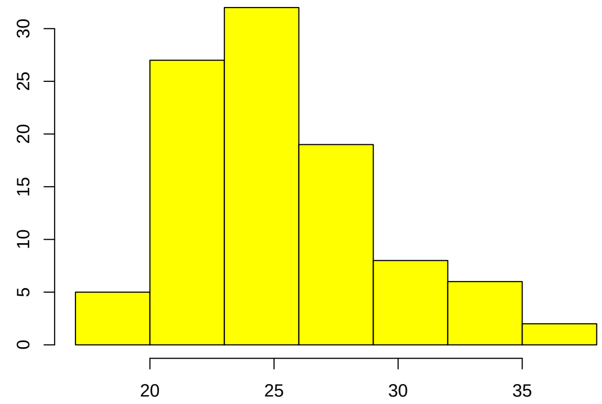
- **histogram** je založen na třídění do intervalů, výjimečně zobrazuje přímo četnosti jednotlivých hodnot (barplot)[hist( )]
- každé třídě odpovídá obdélník o **ploše úměrné četnosti** (absolutní nebo relativní)
- při stejných šířkách intervalů h odpovídají četnostem výšky obdélníků (protože základny jsou stejně dlouhé)
- počet intervalů k: volí se 5–15 tak, aby středy byly okrouhlé
- pomůckou je např. Sturgesovo pravidlo

$$k \approx 1 + 3{,}3 \cdot \log_{10} n = 1 + \log_2 n$$

- příklad věk matek: $k \approx 1 + 3{,}3 \cdot \log_{10} 99 \approx 7{,}6$

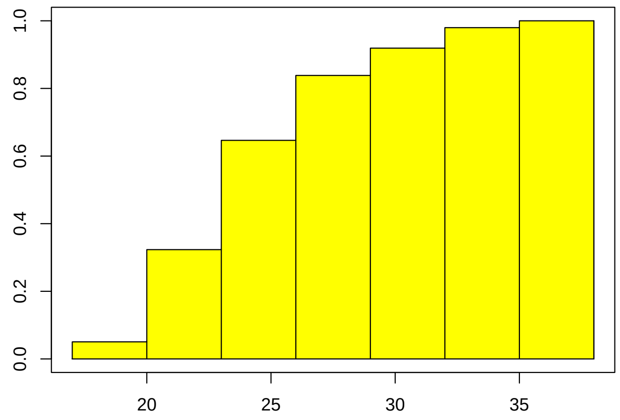
příklad (věk matek): histogram, $h = 3$ ($k = 7$)

[hist(vek.m,seq(17,38,by=3),col="yellow")]



▸ Jdi k četnostem věku mate

příklad (věk matek): kumulativní relativní četnosti



třídění při nestejně dlouhých intervalech

- někdy jsou data nepravidelně rozmístěna
- zpravidla jsou soustředěna u levého okraje rozmezí hodnot (věkové či příjmové složení obyvatelstva)
- pak vhodné zvolit nestejně dlouhé intervaly
- je vhodné zvolit délky intervalů tak, aby delší byly násobkem kratších
- při nestejně dlouhých intervalech musí zjištěné četnosti odpovídat **plocha**, nikoliv výška; pak se na svislou osu nanáší **relativní** četnosti

příklad: tolary

měsíční příjmy 99 osob v tolarech

četnosti

x_j^*	10	11	12	13	14	15	16	17	18	19	20	
n_j	7	14	16	10	6	3	9	3	1	5	3	
x_j^*	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1

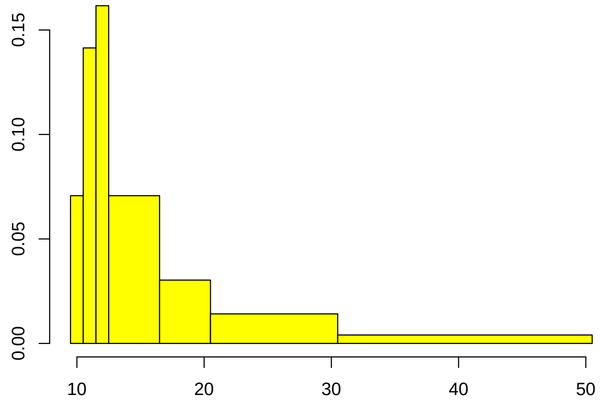
třídní četnosti (hustota = četnost na jednotku délky intervalu$/n$)

třída	10	11	12	13–16	17–20	21–30	31–50	celkem
x_j^*	10	11	12	14,5	18,5	25,5	40,5	
n_j^*	7	14	16	28	12	14	8	99
hustota*99	7	14	16	7	3	1,4	0,4	

▸ Jdi k hodnocení tolarů

příklad (tolary): histogram

na svislé ose je hustota (celková plocha obdélníků = 1)



výběrové charakteristiky polohy: medián

- snaha charakterizovat úroveň (malé či velké hodnoty) číselné veličiny jediným číslem
- medián je číslo, které dělí data na dvě stejně velké části (větších hodnot a menších hodnot, medián je ve variační řadě uprostřed)
- **medián** [median] (prostřední hodnota) $\tilde{x}$ [median(x)]

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} \qquad \text{pro } n \text{ liché}$$

$$\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) \qquad \text{pro } n \text{ sudé}$$

- závorky u indexů jsou nutné: znamenají, že hodnoty byly předem uspořádány do variační řady
- 5, 3, 4, 9, 6 $\tilde{x} = 5$ $(3 < 4 < 5 < 6 < 9)$

kvartily, percentily

- **dolní (horní) kvartil** Q_1 (Q_3) [lower (upper) quartile] vyděluje čtvrtinu nejmenších (největších) hodnot
- kvartil – speciální případ percentilu
- **percentil** [percentile] x_p vyděluje $100p$ % nejmenších hodnot od ostatních
- výpočet percentilů – mnoho vzorečků
- medián je také percentil, totiž $x_{0,5}$
- podobně $Q_1 = x_{1/4} = x_{0,25}$, $Q_3 = x_{3/4} = x_{0,75}$ [quantile(x,probs=c(1/4,3/4))]

výpočet percentilů (jako v R), jen pro ilustraci

jedna z možných definic – Gumbel(1939)

- k danému p se najde celé číslo k splňující

$$\frac{k-1}{n-1} \leq p < \frac{k}{n-1}$$

- tedy $k = \lfloor 1 + (n-1) \cdot p \rfloor$ ($\lfloor x \rfloor$ znamená celou část z x)
- provede se lineární interpolace mezi $x_{(k)}$ a $x_{(k+1)}$:

$$q = (1 + (n-1) \cdot p) - k = \{1 + (n-1) \cdot p\}$$

$$x_p = (1-q) \cdot x_{(k)} + q \cdot x_{(k+1)}$$

  ($\{x\}$ znamená zlomkovou část x, o kolik přesahuje celé číslo)
- např. pro $n = 99, p = 0{,}25$ bude

$$k = \lfloor 1 + (99-1) \cdot 0{,}25 \rfloor = \lfloor 1 + 24{,}5 \rfloor = \lfloor 25{,}5 \rfloor = 25$$

$$q = 25{,}5 - 25 = 0{,}5$$

$$Q_1 = x_{0{,}25} = (1-0{,}5) \cdot x_{(25)} + 0{,}5 \cdot x_{(26)} = 23$$

příklad: věk 99 matek – variační řada

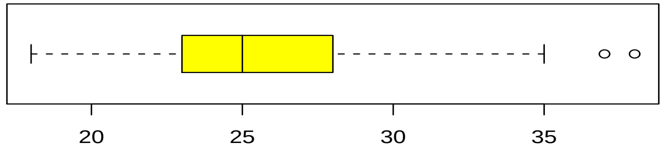
variační řada, medián $\tilde{x}$ =25
kvartily Q_1 =(23+23)/2=23, Q_3 =(28+28)/2=28

18	19	20	20	20	21	21	21	21	21
21	21	21	21	21	21	21	22	22	22
22	22	22	23	23	23	23	23	23	23
23	23	24	24	24	24	24	24	24	24
24	24	25	25	25	25	25	25	25	25
25	25	25	25	26	26	26	26	26	26
26	26	26	26	27	27	27	27	27	27
28	28	28	28	28	28	28	28	28	29
29	29	29	30	30	30	31	31	32	32
32	33	33	33	34	35	35	37	38	

Návrat míry var. věku matek

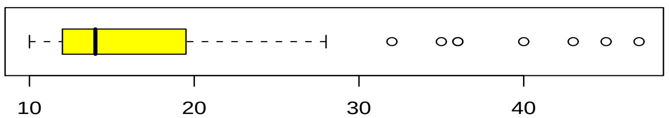
krabicový diagram

- **krabicový diagram** [box-plot] zobrazuje kvartily, medián, minimum, maximum, případně odlehlá pozorování (jsou od bližšího kvartilu dál než $3/2\cdot(Q_3 - Q_1)$) [boxplot(x)]
- příklad: věk matek ($Q_1 = 23$, $\tilde{x} = 25$, $Q_3 = 28$, dvě odlehlá pozorování)



příklad: tolary ($\tilde{x} = 14$, $Q_1 = 12$, $Q_3 = 19{,}5$)

10	10	10	10	10	10	10	11	11	11
11	11	11	11	11	11	11	11	11	11
11	12	12	12	12	12	12	12	12	12
12	12	12	12	12	12	12	13	13	13
13	13	13	13	13	13	13	14	14	14
14	14	14	15	15	15	16	16	16	16
16	16	16	16	16	17	17	17	18	19
19	19	19	19	20	20	20	21	21	21
21	22	22	22	24	24	24	26	27	27
28	32	35	36	36	40	43	45	47	



průměr

- **průměr** [mean] (kdyby bylo všech n hodnot stejných) [mean(x)]

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

- **vážený průměr**: [weighted mean] založen na četnostech

$$\bar{x} = \frac{1}{n}(n_1 x_1^* + \ldots + n_k x_k^*) = \frac{1}{n}\sum_{j=1}^{k} n_j x_j^* = \sum_{j=1}^{k} \frac{n_j}{n} x_j^* = \frac{\sum_{j=1}^{k} n_j x_j^*}{\sum_{j=1}^{k} n_j}$$

- obecněji s vahami $w_1, \ldots, w_k$ hodnot $x_1^*, \ldots, x_k^*$

$$\boxed{\frac{\sum_{j=1}^{k} w_j x_j^*}{\sum_{j=1}^{k} w_j}}$$

váhy nutně nezáporné: $w_j \geq 0$, $\sum_{j=1}^{k} w_j > 0$

příklad: vážený průměr známek

předmět	známka	kredity	součin
A	1	6	6
B	1	6	6
C	2	4	8
D	3	3	9
celkem	7	19	29

- průměr (nevážený): $\bar{x} = 7/4 = 1{,}75$
- vážený průměr (vahami kredity): $\bar{x} = 29/19 \doteq 1{,}53$

průměr pro nula-jedničkovou veličinu

- průměr = relativní četnost jedniček
- počet jedniček/počet všech hodnot (nul i jedniček)
- procento jedniček mezi všemi hodnotami (nulami a jedničkami)
- procento jedinců s danou vlastností
- zpravidla to není pravděpodobnost, nanejvýš její odhad!
- o pravděpodobnost jedničky by šlo, kdybychom měli **náhodně** ze všech hodnot jedinou vybírat a každé z n pozorování mělo **stejnou pravděpodobnost**, že bude vybráno

modus

- **modus** $\hat{x}$ [mode] nejčastější hodnota
- modus lze počítat také pro nominální či ordinální měřítko
- modus nemusí být určen jednoznačně, např. věk matek:

x_j^*	18	19	20	21	22	23	24	25	26	27
n_j	1	1	3	12	6	9	10	12	10	6
x_j^*	28	29	30	31	32	33	34	35	37	38
n_j	9	4	3	2	3	3	1	2	1	1

- maximální četnost nastává pro 21 i pro 25

příklad – věk matek

- již známe $\tilde{x} = 25$, $Q_1 = 23$, $Q_3 = 28$
- modus není určen jednoznačně: $\hat{x} = 21$, $\hat{x} = 25$
- průměr

$$\bar{x} = \frac{1}{99}(26 + 35 + \ldots + 21 + 23) = \frac{2544}{99} \doteq 25{,}7$$

- vážený průměr založený na třídění ($h = 3$, viz str. 19)

$$\bar{x} = \frac{5 \cdot 19 + 27 \cdot 22 + 32 \cdot 25 + 19 \cdot 28 + 8 \cdot 31 + 6 \cdot 34 + 2 \cdot 37}{5 + 27 + 32 + 19 + 8 + 6 + 2}$$
$$= \frac{2547}{99} \doteq 25{,}7 \qquad (\text{ale} \neq \frac{2544}{99})$$

Třídění: věk matek

příklad – tolary

- průměr

$$\bar{x} = \frac{1}{99}(26 + 20 + \ldots + 12 + 10) = \frac{1687}{99} \doteq 17{,}04$$

- vážený průměr založený na četnostech jednotlivých hodnot

$$\bar{x} = \frac{7 \cdot 10 + 14 \cdot 11 + 16 \cdot 12 + \cdots + 1 \cdot 47}{7 + 14 + 16 + \cdots + 1} = \frac{1687}{99} \doteq 17{,}04$$

- vážený průměr založený na třídních četnostech (obr. 24)

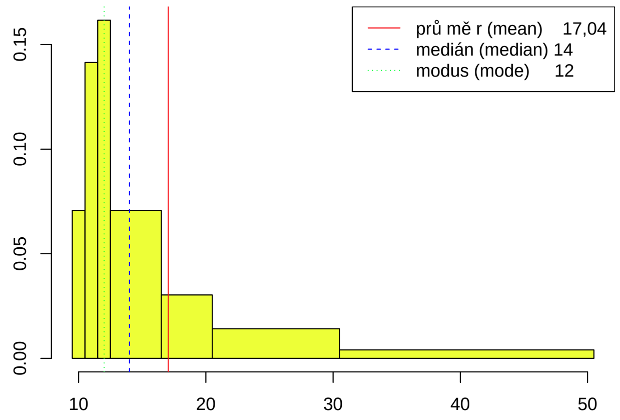
$$\bar{x} = \frac{7 \cdot 10 + 14 \cdot 1 + 16 \cdot 12 + 28 \cdot 14{,}5 + \cdots + 8 \cdot 40{,}5}{7 + 14 + 16 + 28 + 12 + 14 + 8} = \frac{1725}{99} \doteq 17{,}42$$

- modus: $\hat{x} = 12$ medián: $\tilde{x} = 14$

Jdi k četnostem tolarů

příklad (tolary): porovnání tří měr polohy

zpravidla je modus ≤ median ≤ mean



useknutý průměr

- **alfa-useknutý průměr** [trimmed mean]:
  nejprve se oddělí (usekne) 100α % nejmenších a 100α % největších hodnot, ze zbytku se spočítá průměr
- je robustní (necitlivý) vůči odlehlým hodnotám
- volí se zpravidla $\alpha = 0{,}1$
- příklad: věk matek [mean(vek.m,trim=0.1)]

$$\frac{1}{99-18}\left(x_{(10)} + x_{(11)} + \ldots + x_{(89)} + x_{(90)}\right) = 25{,}3$$

příklad (věk matek):useknutý průměr

(průměr počítán pouze z černých čísel)

vyloučí se $\lfloor 0{,}1 \cdot 99 \rfloor = \lfloor 9{,}9 \rfloor = 9$ ($\lfloor x \rfloor$ znamená celou část z x) nejmenších a 9 největších hodnot

18	19	20	20	20	21	21	21	21	21
21	21	21	21	21	21	21	22	22	22
22	22	22	23	23	23	23	23	23	23
23	23	24	24	24	24	24	24	24	24
24	24	25	25	25	25	25	25	25	25
25	25	25	25	26	26	26	26	26	26
26	26	26	26	27	27	27	27	27	27
28	28	28	28	28	28	28	28	28	29
29	29	29	30	30	30	31	31	32	32
32	33	33	33	34	35	35	37	38	

vlastnosti charakteristik polohy

- charakteristiky (míry) polohy mají měřit úroveň spojitého znaku (velký – malý, hodně – málo, …)
- změníme-li všechny hodnoty x_i tak, že přidáme ke každé stejnou konstantu a, změní se o tutéž konstantu také charakteristika polohy (posunutí)
- změníme-li všechny hodnoty x_i tak, že je vynásobíme kladnou konstantou b, toutéž konstantou musíme vynásobit původní charakteristiku polohy, abychom dostali charakteristiku polohy pro upravená data (změna měřítka)
- obecně pro míru polohy $m(x)$ platí

$$\begin{aligned} m(a+x) &= a + m(x), \\ m(b \cdot x) &= b \cdot m(x), \qquad b > 0 \end{aligned}$$

- v **obou** případech míra polohy **reaguje**

charakteristiky variability

- měří rozptýlení (nestejnost, **variabilitu**) hodnot číselné veličiny
- obecně pro míru variability $s(x)$ by mělo platit:

$$\begin{aligned} s(a+x) &= s(x), \qquad \text{(srovnej s } m(a+x) = a + m(x)) \\ s(b \cdot x) &= b \cdot s(x), \qquad b > 0 \end{aligned}$$

- přičtením stejné konstanty a (posunutím) se charakteristika variability nezmění (nezávisí na poloze)
- vynásobení kladnou konstantou znamená, že stejnou konstantou nutno vynásobit charakteristiku variability
- **rozpětí** [range] $R = x_{(n)} - x_{(1)}$
- **kvartilové rozpětí** [quartile range] $R_Q = Q_3 - Q_1$

rozptyl (variance)

- (výběrový) **rozptyl** (variance) [variance] [VAR.VÝBĚR][var(x)] (nevyhovuje druhému požadavku, platí $s^2_{a+b\cdot x} = b^2 \cdot s^2_x$)

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\left((x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right) \\ &= \frac{1}{n-1}\sum_{j=1}^{k} n_j(x_j^*-\bar{x})^2 = \frac{1}{n-1}\left(\sum_{j=1}^{k} n_j x_j^{*2} - n\cdot\bar{x}^2\right) \end{aligned}$$

- nechť $x_1 = 1, x_2 = 3, x_3 = 8$, pak je $\bar{x} = (1+3+8)/3 = 12/3 = 4$

$$s_x^2 = \frac{1}{3-1}\left((1-4)^2 + (3-4)^2 + (8-4)^2\right) = \frac{26}{2} = 13 \doteq 3{,}6^2$$

směrodatná odchylka

- rozptyl měří průměrný čtverec vzdálenosti od průměru
- **směrodatná odchylka** [std. deviation]: odmocnina z rozptylu [SMODCH.VÝBĚR][sd(x)]

$$s_x = \sqrt{s_x^2}$$

- zcela vyhovuje požadavkům na míry variability
- výhoda směrodatné odchylky: stejný fyzikální rozměr jako původní data
- výběrový rozptyl počítaný z *třídních* četností: Sheppardova korekce (jsou-li všechny intervaly délky h):

$$\text{odečti } \frac{h^2}{12}$$

příklad – věk matek

- rozpětí: $R = 38 - 18 = 20$
- kvartilové rozpětí: $R_Q = 28 - 23 = 5$
- rozptyl

$$s^2 = \frac{1}{98}\left((26^2 + 35^2 + \ldots + 21^2 + 23^2) - 99 \cdot \left(\frac{2544}{99}\right)^2\right)$$
$$= 16{,}97 \doteq 4{,}12^2$$

- směrodatná odchylka je 4,12

▸ Var. řada věku matek

příklad – věk matek 2

- pomocí třídních četností

$$s^2 = \frac{1}{98}\left((5 \cdot 19^2 + 27 \cdot 22^2 + \ldots + 2 \cdot 37^2) - 99 \cdot \left(\frac{2547}{99}\right)^2\right)$$
$$= 16{,}36 = (4{,}05)^2$$

- navíc Sheppardova korekce (všechny třídy mají délku $h = 3$)

$$s^2 = 16{,}36 - \frac{3^2}{12} = 15{,}61 = (3{,}95)^2$$

střední odchylka

- **střední odchylka** [mean deviation]: průměr odchylek od mediánu (někdy od průměru) [mean(abs(x-median(x)))]

$$d = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}|$$

- **střední diference** [mean difference]: průměr vzájemných vzdáleností všech n^2 dvojic [mean(abs(outer(x,x,"-")))]

$$\Delta = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |x_i - x_j|$$
$$= \frac{2}{n^2}\sum_{j>i}\sum \left(x_{(j)} - x_{(i)}\right)$$

normované charakteristiky rozptýlenosti

- dosud zavedené charakteristiky variability závisejí na volbě měřítka (např. délka v m nebo v km)
- hledáme charakteristiky nezávislé na měřítku, nutně *poměrové* měřítko, *kladné* hodnoty
- umožní **porovnání** z různých souborů
- **variační koeficient** [sd(x)/mean(x)]

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

- **(Giniho) koeficient koncentrace**

$$G = \frac{\Delta}{2\bar{x}}\left(= \frac{2\sum_{i=1}^{n} i \cdot x_{(i)}}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n}\right)$$

souvisí s plochou u Lorenzovy křivky
měří například nerovnoměrnost příjmů, velikostí územních jednotek ...

z-skór, standardizace

- variační koeficient v, Giniho koeficient G jsou příklady bezrozměrných veličin (zásluhou průměru ve jmenovateli závisí G i v na posunutí!)
- z-skóry [STANDARDIZE(x;průměr(x);smodch.výběr(x))] [(x-mean(x))/sd(x)] nebo [c(scale(x))]

$$z_i = \frac{x_i - \bar{x}}{s_x}, \qquad i = 1, 2, \ldots, n$$

- dostaneme nulový průměr ($\bar{z} = 0$), jednotkový rozptyl ($s_z = 1$)
- z-skóry jsou bezrozměrné $\Rightarrow$ umožní hodnotit vlastnosti nezávislé na poloze a variabilitě, např. tvar rozdělení
- $x_1 = 1, x_2 = 2, x_3 = 3 \quad \Rightarrow \quad \bar{x} = 2, s_x = 1$
  $z_1 = \frac{1-2}{1} = -1, z_2 = \frac{2-2}{1} = 0, z_3 = \frac{3-2}{1} = 1$

charakteristiky tvaru: šikmost [skewness]

- invariantní vůči posunutí i změně měřítka:

$$\gamma(a + x) = \gamma(x)$$
$$\gamma(b \cdot x) = \gamma(x) \qquad b > 0$$

proto použijeme z-skóry

- **šikmost** $\sqrt{b_1}$ – **průměr** z 3. mocnin z-skórů [SKEW()] [mean(scale(x)^3)]

$$\sqrt{b_1} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right)^3$$

- pro symetrický histogram $\sqrt{b_1}$ blízké nule
- doprava protažený histogram pro $\sqrt{b_1} >> 0$
- doleva protažený histogram pro $\sqrt{b_1} << 0$

charakteristiky tvaru: špičatost [kurtosis]

- **špičatost** b_2 – **průměr** ze 4. mocnin z-skórů (někdy se odečítá 3) [KURT()] [mean(scale(x)^4)]

$$b_2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right)^4$$

- někdy se počítají odhady populační šikmosti a špičatosti jinak (Excel: s_x jinak, Fisherovo g_1, g_2 – pro zajímavost)

$$g_1 = \frac{\sqrt{n(n-1)}}{n-2} \sqrt{b_1}, \qquad g_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)} \left(b_2 - \frac{3(n-1)}{n+1} \right)$$

- šikmost a špičatost slouží k hodnocení, zda lze předpokládat *normální rozdělení* (bude zavedeno později)

charakteristiky polohy v geografii/demografii

- místo x budeme tuto přednášku označovat měřené hodnoty jako y, princip pojmů je stejný, označení je jen konvence
- často známe jen průměry v dílčích souborech a četnosti: průměry se použijí jako y_j^*, četnosti standardně
- příklad: věk nových profesorů a docentů UK 2002:
  41 profesorů, průměrný věk 51,1 ($n_1 = 41$, $y_1^* = 51{,}1$)
  77 docentů, průměrný věk 47,8 ($n_2 = 77$, $y_2^* = 47{,}8$)
  průměr nových habilitovaných akademických pracovníků (**vážený průměr**):
  [weighted.mean(c(51.1,47.8),c(41,77))]

$$\frac{41 \cdot 51{,}1 + 77 \cdot 47{,}8}{41 + 77} = 48{,}9$$

nikoliv [mean(c(51.1,47.8))]

$$\frac{51{,}1 + 47{,}8}{2} = 49{,}4$$

charakteristiky polohy v geografii/demografii (2)

- **geografický střed**
  - bod
  - průsečík průměrné zeměpisné šířky a průměrné zeměpisné délky; průměry vážíme velikostí sledovaného jevu
- **geografický medián** – obdoba mediánu,
  - čára, která rozděluje geografické objekty do dvou disjunktních souvislých skupin
  - hodnocená vlastnost určí váhy objektů
  - uspořádání hodnocení znaků dáno zvolenou geografickou vlastností (např. zeměpisnou délkou)

příklad: geografický střed obyvatel ČR

použijeme jen údaje o krajích

```
   Zkratka        kraj rozloha obyvatel šířka délka
1        A       Praha   49609  1188126 50,08 14,42
2        S Středočeský 1101473  1175254 50,08 14,42
3        C   Jihočeský 1005688   630006 48,98 14,47
. . . . . . . . . . . .
14       T M-Slezský    542698  1249290 49,84 18,32
```

- šířka

$$\frac{1188126 \cdot 50{,}08 + 1175254 \cdot 50{,}08 + \ldots 1249290 \cdot 49{,}84}{1188126 + 1175254 + \ldots + 1249290} = 49{,}84$$

- délka

$$\frac{1188126 \cdot 14{,}42 + 1175254 \cdot 14{,}42 + \ldots 1249290 \cdot 18{,}32}{1188126 + 1175254 + \ldots + 1249290} = 15{,}59$$

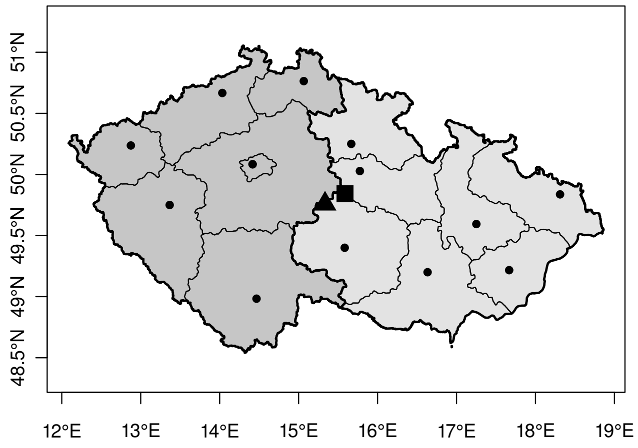
příklad: geografický medián

sčítáme obyvatele postupně od západu na východ, poslední je v západní polovině Liberecký kraj L, ve východní polovině je první kraj Vysočina označený symbolem J

	délka	obyvatel	součet	podíl
K	12,87504	304602	304602	0,02960984
P	13,36667	554537	859139	0,08351543
U	14,03333	823265	1682404	0,16354361
A	14,41667	1188126	2870530	0,27903930
S	14,41667	1175254	4045784	0,39328372
C	14,46667	630006	4675790	0,45452553
L	15,06528	430774	5106564	0,49640033
J	15,58333	511645	5618209	0,54613646
H	15,66667	549643	6167852	0,59956631
E	15,77583	507751	6675603	0,64892392
B	16,63333	1132563	7808166	0,75901843
M	17,25119	639894	8448060	0,82122142
Z	17,66667	589839	9037899	0,87855866
T	18,31117	1249290	10287189	1,00000000

příklad: geografický střed obyvatel ČR

použijeme jen údaje o krajích, □ – střed obyvatel (△ – velikost kraje)



Giniho index

(místo x budeme z důvodů, které vysvitnou později, psát y)

- **Giniho index** charakterizuje jediným číslem nerovnoměrnost rozdělení (bohatství příjmů, ...) $\boxed{G = \Delta/(2\bar{y})}$
- průměrný rozdíl v bohatství vztažený k dvojnásobku průměru
- mají-li všichni stejně ($y_{(1)} = \ldots = y_{(n)} > 0$), je nutně $\Delta = 0$ a tedy $G = 0$
- má-li jeden všechno, ostatní nic ($0 = y_{(1)} = \ldots = y_{(n-1)} < y_{(n)} = a$), pak je

$$\bar{y} = \frac{a}{n} \qquad \Delta = \frac{2(n-1)a}{n^2}$$

$$G = \frac{2(n-1)a}{n^2} \cdot \frac{n}{2a} = \frac{n-1}{n}$$

- jemnějším nástrojem bude Lorenzova křivka

příklad: tolary (rozdělení příjmů)

různé hodnoty jsme označili $y_1^* < \ldots < y_k^*$

jaké procento nejchudších získá **desetinu** celkového příjmu?
četnosti 99 osob (celkový měsíční příjem je 1687)

y_j^*	10	11	12	13	14	15	16	17	18	19	20	
n_j	7	14	16	10	6	3	9	3	1	5	3	
y_j^*	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1

sčítejme příjmy nejchudších, dokud nenasčítáme 10 % z 1687

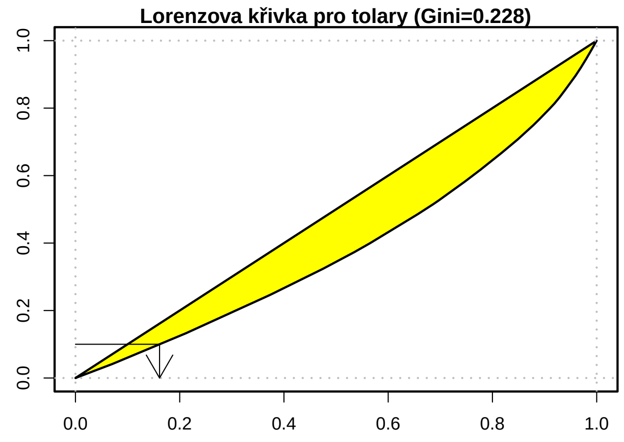
$$(7 \cdot 10 + 8 \cdot 11)/1687 = 158/1687 = 0{,}0937 = 9{,}37\ \%$$
$$(7 \cdot 10 + 9 \cdot 11)/1687 = 169/1687 = 0{,}1002 = 10{,}02\ \%$$

u jaké části z 99 osob jsme sčítali příjmy?

$$(7 + 8)/99 = 15/99 = 0{,}152 = 15{,}2\ \%$$
$$(7 + 9)/99 = 16/99 = 0{,}162 = 16{,}2\ \%$$

Lorenzova křivka (Tolary)

10 % příjmů u nejchudších 16 %



příklad: tolary (rozdělení příjmů)

různé hodnoty jsme označili $y_1^* < \ldots < y_k^*$

jaké procento nejchudších získá **dvě desetiny** celkového příjmu?
četnosti 99 osob (celkový měsíční příjem je 1687)

y_j^*	10	11	12	13	14	15	16	17	18	19	20	
n_j	7	14	16	10	6	3	9	3	1	5	3	
y_j^*	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1

sčítejme příjmy nejchudších, dokud nenasčítáme 337 (20 % z 1687)

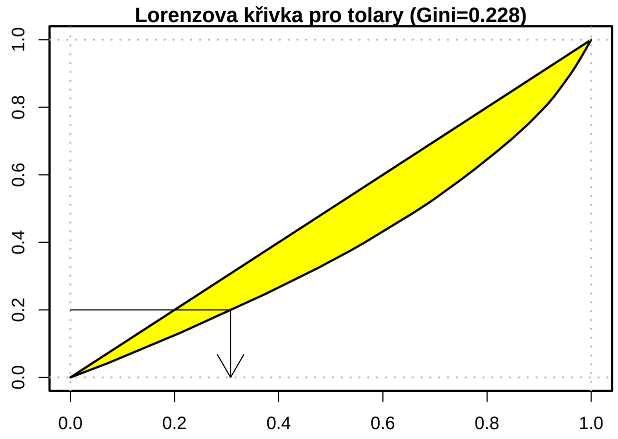
$$(7 \cdot 10 + 14 \cdot 11 + 9 \cdot 12)/1687 = 332/1687 = 0{,}1938 = 19{,}68\ \%$$
$$(7 \cdot 10 + 14 \cdot 11 + 10 \cdot 12)/1687 = 344/1687 = 0{,}2039 = 20{,}39\ \%$$

u jaké části z 99 osob jsme sčítali příjmy?

$$(7 + 14 + 9)/99 = 30/99 = 0{,}303 = 30{,}3\ \%$$
$$(7 + 14 + 10)/99 = 31/99 = 0{,}313 = 31{,}3\ \%$$

Lorenzova křivka (Tolary)

20 % příjmů u nejchudších 30 %



příklad: tolary (rozdělení příjmů)

jaké procento nejchudších získá **polovinu** celkového příjmu?
četnosti (celkový měsíční příjem je 1687)

y_j	10	11	12	13	14	15	16	17	18	19	20
n_j	7	14	16	10	6	3	9	3	1	5	3

y_j	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1

sčítejme příjmy nejchudších, dokud nenasčítáme 50 % z 1687

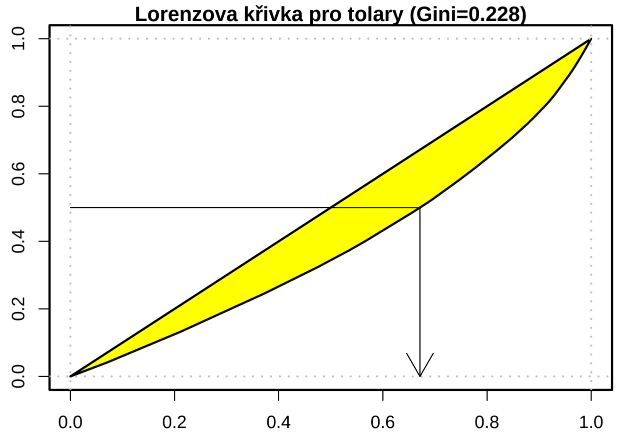
$$(7 \cdot 10 + \ldots + 9 \cdot 16 + 17)/1687 = 836/1687 = 0{,}4956 = 49{,}56\ \%$$
$$(7 \cdot 10 + \ldots + 9 \cdot 16 + 2 \cdot 17)/1687 = 853/1687 = 0{,}5056 = 50{,}56\ \%$$

u jaké části z 99 osob jsme sčítali příjmy?

$$(7 + \ldots + 9 + 1)/99 = 66/99 = 0{,}6667 = 66{,}67\ \%$$
$$(7 + \ldots + 9 + 2)/99 = 67/99 = 0{,}6768 = 67{,}68\ \%$$

Lorenzova křivka (Tolary)

50 % příjmů u nejchudších 67 %



příklad: tolary (rozdělení příjmů)

jaké procento získají čtyři (tj. asi 4 %) nejbohatší?
četnosti (celkový měsíční příjem je 1687)

y_j	10	11	12	13	14	15	16	17	18	19	20
n_j	7	14	16	10	6	3	9	3	1	5	3

y_j	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1

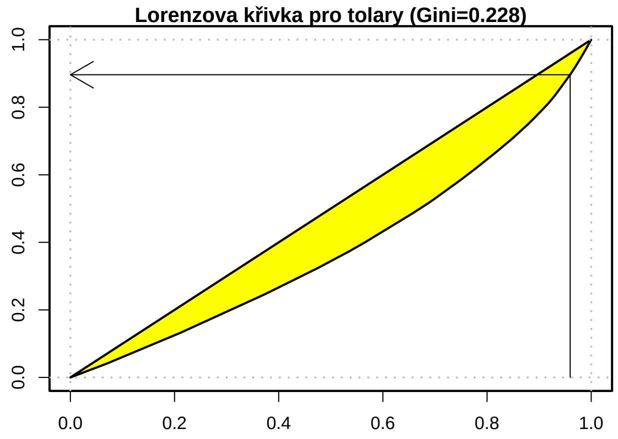
sečteme příjmy oněch čtyř nejbohatších

$$(47 + 45 + 43 + 40)/1687 = 175/1687 = 0{,}1037 = 10{,}37\ \%$$

čtyři nejbohatší tedy dostanou přes 10 % bohatství, na ostatních 95,6 % zbude jen $1 - 0{,}1037 = 0{,}8963 = 89{,}63$ % bohatství

Lorenzova křivka (Tolary)

nejchudších 96 % má 90 % příjmů



Lorenzova křivka, její konstrukce

(pozor na rozlišování velikosti písmen y a Y!!!!!!!!)

- variační řada: $0 < y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ [sort(y)]
- kumulativní součty pro $j = 0, 1, \ldots, n$ [cumsum(sort(y))] (kolik celkem patří j nejchudším)

$$Y_{(0)} = 0 \qquad Y_{(j)} = y_{(1)} + y_{(2)} + \ldots + y_{(j)} = \sum_{i=1}^{j} y_{(i)}$$

- úsečkami spojit body $[j/n; Y_{(j)}/Y_{(n)}]$, $0 \leq j \leq n$, j/n – díl populace $Y_{(j)}/Y_{(n)}$ – díl bohatství
- [n = length(y)]
- [Y = cumsum(sort(y))]
- [plot((0:n)/n,c(0,Y)/Y[n],type="l")] ("malé el")
- [abline(a=0,b=1); abline(h=0:1,v=0:1,lty=3)]

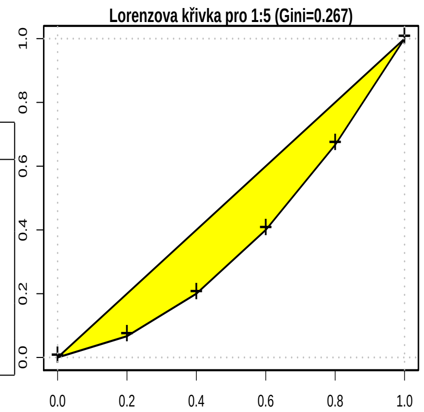
Lorenzova křivka

- vodorovná osa: postupné načítání lidí od nejchudších, jako díl celku
- svislá osa: postupné načítání bohatství od nejchudších, jako díl celku
- zajímá nás plocha nad touto lomenou čarou a pod úhlopříčkou jednotkového čtverce
- plocha měří nerovnoměrnost rozdělení nějakého zdroje
- kdyby dostal každý stejně, bude velikost plochy nulová
- kdyby všechno dostala jediná z n osob, lomená čára bude nulová až do $(n-1)/n$
- pro $n \to \infty$ je $(n-1)/n \to 1$
- Giniho koeficient koncentrace je **dvojnásobkem** této plochy, porovnává ji s plochou dolního trojúhelníku

umělý příklad

$y_1, \ldots, y_5$: 1, 2, 3, 4, 5

j	j/n	$y_{(j)}$	$Y_{(j)}$	$Y_{(j)}/Y_{(n)}$
0	0,0		0	0,000
1	0,2	1	1	0,067
2	0,4	2	3	0,200
3	0,6	3	6	0,400
4	0,8	4	10	0,667
5	1,0	5	15	1,000



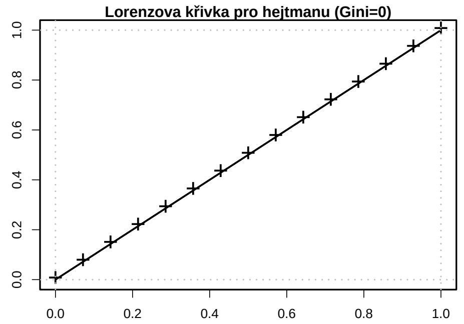
umělý příklad - pokračování

výpočet Giniho koeficientu ($n = 5$)

$$\begin{aligned}
5^2 \cdot \Delta &= |1-1| + |1-2| + |1-3| + |1-4| + |1-5| \\
&+ |2-1| + |2-2| + |2-3| + |2-4| + |2-5| \\
&+ |3-1| + |3-2| + |3-3| + |3-4| + |3-5| \\
&+ |4-1| + |4-2| + |4-3| + |4-4| + |4-5| \\
&+ |5-1| + |5-2| + |5-3| + |5-4| + |5-5| \\
&= 10 + 7 + 6 + 7 + 10 \\
\Delta &= 40/25 = 1{,}6 \\
\bar{y} &= 3 \\
G &= \frac{1{,}6}{2 \cdot 3} = \frac{1{,}6}{6} = 0{,}267
\end{aligned}$$

Lorenzova křivka počet hejtmanů v krajích ČR

- v každém kraji je stejně hejtmanů, proto postupné součty rovnoměrně rostou, totéž platí pro $Y_{(j)}/Y_{(n)}$ $(= j/n)$
- lomená čára Lorenzovy křivky přejde v úsečku a plocha zmizí
- průměrná diference je nulová (všechny rozdíly $|y_i - y_j|$ u počtu hejtmanů jsou nulové)

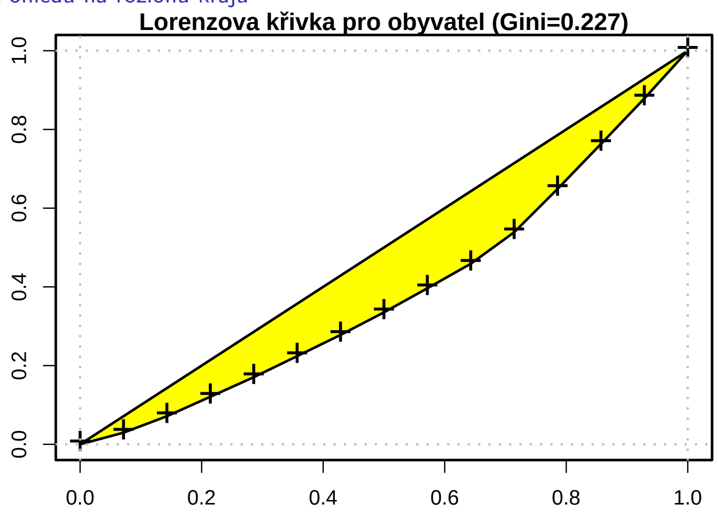


příklad: obyvatelé v krajích ČR ke konci roku 2006

kraj	obyvatel	rozloha[km^2]	hustota na km^2
i	y_i	x_i	$y_i^{\text{prům}} = y_i/x_i$
Hlavní město Praha	1 188 126	496,1	2 395,0
Středočeský kraj	1 175 254	11 014,7	106,7
Jihočeský kraj	630 006	10 056,9	62,6
Plzeňský kraj	554 537	7 561,1	73,3
Karlovarský kraj	304 602	3 314,6	91,9
Ústecký kraj	823 265	5 334,5	154,3
Liberecký kraj	430 774	3 163,0	136,2
Královéhradecký kraj	549 643	4 758,4	115,5
Pardubický kraj	507 751	4 518,6	112,4
Vysočina	511 645	6 795,6	75,3
Jihomoravský kraj	1 132 563	7 196,3	157,4
Olomoucký kraj	639 894	5 266,8	121,5
Zlínský kraj	589 839	3 963,5	148,8
Moravskoslezský kraj	1 249 290	5 427,0	230,2
celkem	10 287 189	78 867,0	130,4

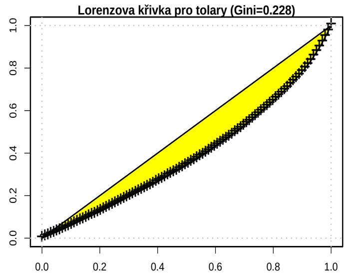
Lorenzova křivka (obyvatelé – kraje)

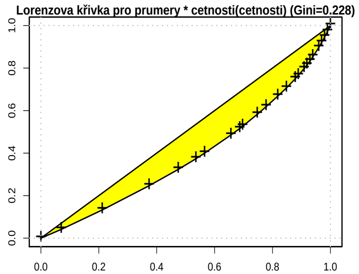
zatím bez ohledu na rozlohu krajů



Lorenzova křivka pro tolary ještě jinak

- spousta hodnot proměnné tolary se opakuje, mohli jsme použít četnosti
- hodnota $y^*_{(j)}$ se vyskytuje $n_j$krát
  - o 10*7=70 tolarů se rozdělilo 7 „nejchudších“ osob
  - o 11*14=154 tolarů se rozdělilo 14 „druhých nejchudších“
  - posledních 47 tolarů připadlo jedinému nejbohatšímu
  - obě křivky (až na křížky) jsou totožné





obyvatelstvo ČR (hustota vážená rozlohou)

- hodnotíme nerovnoměrnost rozmístění obyvatel v republice, ale k disposici kumulované údaje (za celé kraje)
- ideálně bychom pro každou jednotlivou jednotku plochy (např. km^2) potřebovali znát počet obyvatel zde žijících
- známe jen počty obyvatel y_i v krajích a rozlohu krajů x_i
- předpokládáme rovnoměrné rozmístění uvnitř kraje, tedy $y_i^{\text{prům}} = y_i/x_i$ obyvatel na každý km^2 v i-tém kraji
- každou takovou hustotu $y_i^{\text{prům}}$ musíme započítat x_i-krát
- celková plocha $x_1 + \ldots + x_{14} (= X_{14})$
- průměrný počet obyvatel na km^2 (vážený prům. hustot $y_i^{\text{prům}}$)

$$\bar{y}^{\text{prům}} = \frac{\sum_i x_i y_i^{\text{prům}}}{\sum_i x_i} = \frac{\sum_i x_i (y_i/x_i)}{\sum_i x_i} = \frac{\sum_i y_i}{\sum_i x_i}$$

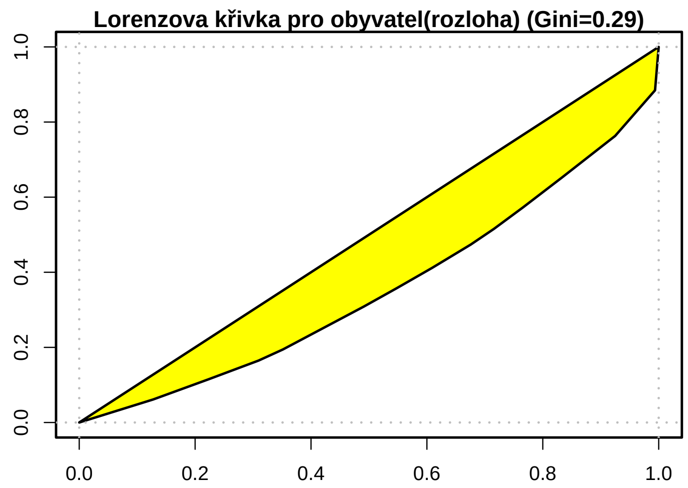
- **dál budeme předpokládat** $\boxed{y_1^{\text{prům}} \leq \ldots \leq y_k^{\text{prům}}}$

příklad: kraje ČR ke konci roku 2006

relativní kumulativní podíly = souřadnice bodů na vodorovné a svislé ose,
pořadí krajů podle hustoty obyvatel!!!

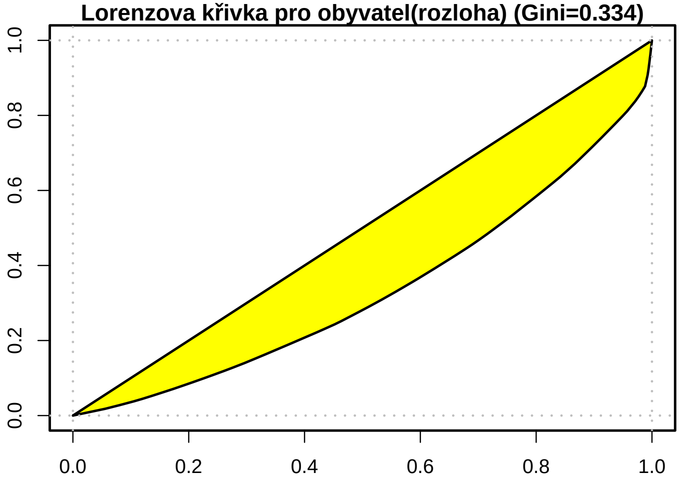
kraj	obyvatel			rozloha[km^2]			hustota
	abs.	kumul.	rel.	abs.	kum.	rel.	osob/km^2
j	y_j	Y_j	Y_j/Y_{14}	x_j	X_j	X_j/X_{14}	$y_j^{\text{prům}}$
C	630 006	630 006	0,061	10 056,9	10 056,9	0,128	62,6
P	554 537	1 184 543	0,115	7 561,1	17 618,0	0,223	73,3
J	511 645	1 696 188	0,165	6 795,6	24 413,6	0,310	75,3
K	304 602	2 000 790	0,194	3 314,6	27 728,1	0,352	91,9
S	1 175 254	3 176 044	0,309	11 014,7	38 742,9	0,491	106,7
E	507 751	3 683 795	0,358	4 518,6	43 261,5	0,549	112,4
H	549 643	4 233 438	0,412	4 758,4	48 019,8	0,609	115,5
M	639 894	4 873 332	0,474	5 266,8	53 286,6	0,676	121,5
L	430 774	5 304 106	0,516	3 163,0	56 449,6	0,716	136,2
Z	589 839	5 893 945	0,573	3 963,5	60 413,1	0,766	148,8
U	823 265	6 717 210	0,653	5 334,5	65 747,6	0,834	154,3
B	1 132 563	7 849 773	0,763	7 196,3	72 943,9	0,925	157,4
T	1 249 290	9 099 063	0,885	5 427,0	78 370,9	0,994	230,2
A	1 188 126	10 287 189	1,000	496,1	78 867,0	1,000	2 395,0

Lorenzova křivka: obyvatelé krajů, vztaženo k rozloze



Jdi ke grafu okr

Lorenzova křivka: obyvatelé okresů, vztaženo k rozloze



Jdi zpět ke graf

poznámky

- hrubší hodnocení (kraje, nikoliv okresy) znamená **menší** hodnotu Giniho indexu! (obecná vlastnost)
- nezáleží na zvolených jednotkách
- ve všech případech je **pořadí** sčítanců dáno pořadím „hustot“ $y_i^{\text{prům}} = \frac{y_i}{x_i}$ (např. obyvatel/rozloha), tj. $y_1^{\text{prům}} \leq \ldots \leq y_k^{\text{prům}}$
- na svislé ose y jde o podíl na bohatství
- kumulativní součty od nejchudších jsou $Y_j = \sum_{i=1}^{j} y_i$
- na vodorovné ose x jde o umístění v řadě od nejchudších k nejbohatším
- označme kumulativní součty $X_j = \sum_{i=1}^{j} x_i$
- pro zajímavost: $X_k = x_1 + \ldots + x_k$ odpovídá u četností celkovému počtu pozorování n, rozděluje se bohatství Y_k

shrnutí výpočtu v případě vah

(stále předpokládáme $y_1^{\text{prům}} \leq \ldots \leq y_k^{\text{prům}}$)

- kumulativní **součty** $X_j = \sum_{i=1}^{j} x_i, \qquad Y_j = \sum_{i=1}^{j} y_i$
- Lorenzova křivka spojuje body $\left[\frac{X_j}{X_k}; \frac{Y_j}{Y_k}\right]$
- střední diference průměrných počtů obyvatel na km^2 (hustot)

$$\Delta = \frac{1}{X_k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} x_i x_j \left| y_i^{\text{prům}} - y_j^{\text{prům}} \right| = \frac{2}{X_k^2} \sum_{i=2}^{k} \sum_{j=1}^{i-1} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j} \right)$$

$$= \frac{2}{X_k^2} \sum_{i=2}^{k} \sum_{j=1}^{i-1} (x_j y_i - x_i y_j) = \ldots = \frac{2}{X_k^2} \sum_{i=1}^{k-1} (X_i Y_{i+1} - X_{i+1} Y_i)$$

$$G = \frac{\Delta}{2\bar{y}} = \sum_{i=1}^{k-1} \left(\frac{X_i}{X_k} \frac{Y_{i+1}}{Y_k} - \frac{X_{i+1}}{X_k} \frac{Y_i}{Y_k} \right)$$

- při výpočtu G se použijí relativní kumulativní podíly x i y

přehled závislostí

- abychom mohli vyšetřovat závislost, musíme na jedné statistické jednotce měřit aspoň dva znaky
- postupy (i grafické) závisí na měřítcích obou znaků
  - kvalitativní – kvalitativní (vzdělání – pracovní zařazení)
  - kvalitativní – kvantitativní (vzdělání – roční příjem)
  - kvantitativní – kvantitativní (věk – roční příjem)
- zatím popisné charakteristiky a grafy, prokazování závislosti později

kvalitativní – kvalitativní

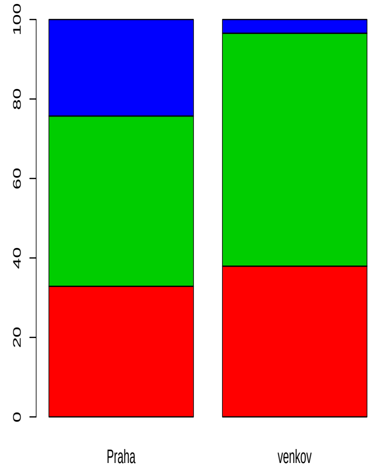
- kvalitativní data – znak v nominálním (ordinálním) měřítku
- hodnoty vyjadřujeme pomocí četností
- dva znaky – četnosti možných **dvojic hodnot** n_{ij} (**sdružené četnosti**)
- zapisujeme do **kontingenční tabulky** [contingency table] [table(x,y)] nebo [xtabs(∼x+y)]
- doplňujeme **marginální četnosti** [marginal frequencies]
  - součty po řádcích a po sloupcích
  - četnosti jednotlivých hodnot každého ze znaků zvlášť [addmargins(table(x,y))]
- oba znaky nula-jedničkové – kontingenční tabulka 2×2, **čtyřpolní tabulka** [fourfold table]

příklad – vzdělání matek

(zobrazení relativních četností v %, pozor na orientaci grafu!)

vzdělání	porodnice Praha	porodnice venkov	celkem
základní	23	11	34
střední	30	17	47
VŠ	17	1	18
celkem	70	29	99

vzdělání	porodnice Praha	porodnice venkov	celkem
základní	32,9 %	37,9 %	34,3 %
střední	42,8 %	58,6 %	47,5 %
VŠ	24,3 %	3,5 %	18,2 %
celkem	100 %	100 %	100 %

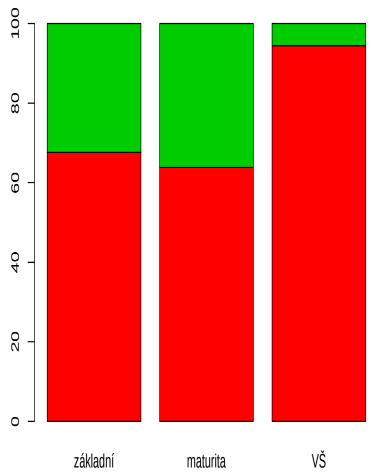


příklad – vzdělání matek

(zobrazení relativních četností v %, pozor na orientaci grafu!)

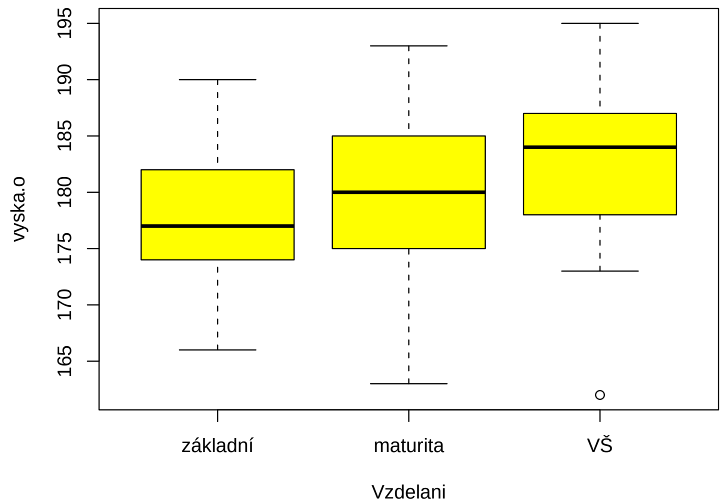
vzdělání	porodnice Praha	porodnice venkov	celkem
základní	23	11	34
střední	30	17	47
VŠ	17	1	18
celkem	70	29	99

vzdělání	porodnice Praha	porodnice venkov	celkem
základní	67,6 %	32,4 %	100 %
střední	63,8 %	36,2 %	100 %
VŠ	94,4 %	6,6 %	100 %
celkem	70,7 %	29,3 %	100 %



kvalitativní – kvantitativní

příklad: závislost výšky otce na vzdělání matky



kvalitativní – kvantitativní

- podle kvalitativní proměnné rozdělíme hodnoty kvantitativní proměnné do dílčích souborů
- porovnáme charakteristiky dílčích souborů (zejména charakteristiky polohy) mezi sebou, pokud se hodně liší, svědčí to pro závislost
- celkový průměr = vážený průměr dílčích souborů
- celkový rozptyl $\doteq$ vážený průměr rozptylů + vážený rozptyl průměrů (přesně jen pro populační rozptyly s n ve jmenovateli)
- snáze jako **rozklad součtu čtverců**

příklad: platy u tří skupin zaměstnanců

skup.	příjem	n_j	$\bar{x}_j$	s_j	s_j^2
žlutí	200 150	2	175,00	35,4	1250,0
modří	80 70 60 60	4	67,50	9,6	91,7
černí	20 20 18 18 15 15 10 10	8	15,75	4,0	16,2
všichni	746	14	53,29	57,7	3334,4

$$\bar{x} = \frac{2 \cdot 175{,}0 + 4 \cdot 67{,}50 + 8 \cdot 15{,}75}{2+4+8} = \frac{746}{14} = 53{,}29$$

$$s^2 = 3334{,}4 > \frac{2 \cdot 1250{,}0 + 4 \cdot 91{,}7 + 8 \cdot 16{,}2}{2+4+8} = 214{,}0$$

- nevážený (nesmyslný) průměr by byl $(175 + 67{,}5 + 15{,}75)/3 = 86{,}08!$
- rozptyl celkem je mnohem větší, než jsou rozptyly ve skupinách
- příčina: nestejné průměry

rozklad součtu čtverců

- velikost kolísání **všech** platů (celková variabilita):

$$\begin{aligned} SST &= (200-53{,}29)^2 + (150-53{,}29)^2 + (80-53{,}29)^2 + \dots \\ &\quad + (10-53{,}29)^2 = 43\ 346{,}86 \end{aligned}$$

- velikost kolísání **uvnitř** skupin:

$$\begin{aligned} SSE &= (200-175)^2 + (150-175)^2 + (80-67{,}5)^2 + \dots \\ &\quad + (10-15{,}75)^2 = 1\ 638{,}5 \end{aligned}$$

- kolísání průměrů (**mezi** skupinami):

$$\begin{aligned} SSA &= 2 \cdot (175-53{,}29)^2 + 4 \cdot (67{,}5-53{,}29)^2 \\ &\quad + 8 \cdot (15{,}75-53{,}29)^2 = 41\ 708{,}36 \end{aligned}$$

- kontrola: 1 638,5+41 708,36=43 346,86

rozklad součtu čtverců obecně

základ tzv. analýzy rozptylu

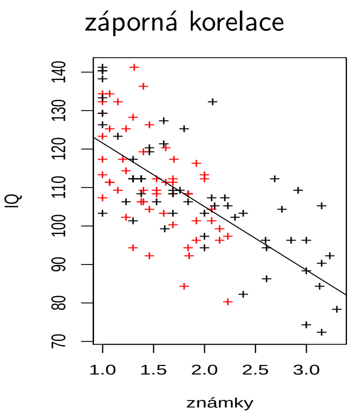
- x_{ij} j-tá hodnota v i-té skupině (plat j-té osoby v i-té skupině)
- n_i počet hodnot v i-té skupině, k počet skupin
- $\bar{x}_{i\bullet}$ průměr v i-té skupině (průměrný plat v i-té skupině)
- $\bar{x}_{\bullet\bullet}$ celkový průměr (průměr všech platů)

$$\begin{aligned} SST &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{\bullet\bullet})^2 \\ &= \sum_{i=1}^{k} n_i (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\bullet})^2 \\ &= SSA + SSE \end{aligned}$$

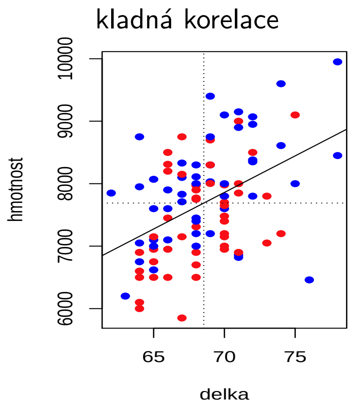
celkový = mezi skupinami + uvnitř skupin (reziduální)

kvantitativní – kvantitativní

[plot(iq∼zn7,data=Iq,col=1+divka,pch="+")]



$r = -0{,}69$



$r = 0{,}45$

popis závislosti spojitých veličin

- (výběrová) **kovariance** [covariance] [cov(vek.o,vek.m)]

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- zřejmě je $s_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = s_x^2$, $s_{yy} = s_y^2$
- (Pearsonův, momentový) **korelační koeficient** [(Pearson, product-moment) correlation coefficient]
- lze zapsat pomocí z-skórů [cor(vek.o,vek.m)]

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

příklad: hmotnost a délka dětí (24. týden věku)

korelační koeficient při změně měřítka

- délka [cm]: $\bar{x} = 68{,}5$ $s_x = 3{,}28$
- hmotnost [g]: $\bar{y} = 7690$, $s_y = 845$
- kovariance [cm · g]: $s_{xy} = 1257$
- korelační koeficient: $r = \frac{1257}{3{,}28\cdot 845} = 0{,}45$
- hmotnost [kg]: $\bar{y} = 7{,}69$ $s_y = 0{,}845$
- kovariance [cm · kg]: $s_{xy} = 1{,}257$
- korelační koeficient: $r = \frac{1{,}257}{3{,}28\cdot 0{,}845} = 0{,}45$
- které charakteristiky závisí na použitém měřítku?

vlastnosti Pearsonova korelačního koeficientu

- vypovídá o směru závislosti
- při $r < 0$ s x rostoucím y v průměru klesá (např. IQ a známky)
- při $r > 0$ s x rostoucím y v průměru roste (např. váha a výška)
- platí $-1 \leq r \leq 1$
- $r = \pm 1$ jedině tehdy, když body $[x; y]$ leží na přímce
- vzájemné nezávislosti x, y odpovídají r blízká nule (upřesníme!)
- nemusí zachytit křivočarou (nelineární) závislost

základní pojmy

- **pokus** – dobře definovaná situace (postup), která končí jedním z řady možných výsledků (vržená kostka spadne na pevnou podložku)
- **náhodný pokus** – pokus, u něhož předem nevíme, který výsledek nastane (která strana kostky padne příště?); předpokládá se stabilita relativních četností možných výsledků
- **náhodný jev** – tvrzení o výsledku náhodného pokusu
- **pravděpodobnost** náhodného jevu A – číselné vyjádření očekávání, že výsledkem náhodného pokusu bude právě A
- racionální představa: při velkém počtu opakování pokusu se relativní četnost jevu blíží k pravděpodobnosti tohoto jevu

klasická pravděpodobnost (Laplace)

- **jistý jev** (nastává vždy) lze rozdělit na M *stejně pravděpodobných* neslučitelných (disjunktních) **elementárních jevů** (symetrie)
- každý jev lze složit z těchto elementárních jevů
- je celkem M_A **příznivých** jevu A (je z nich složen)
- **klasická definice pravděpodobnosti** (metoda výpočtu)

$$\mathrm{P}(A) = \frac{M_A}{M} \qquad \left(= \frac{\#\ \text{příznivých}}{\#\ \text{možných}} \right)$$

- **klasickou pst lze použít jen někdy!** (Sportka, Sazka)
- nelze použít např.:
  - dostuduje resp. nedostuje
  - dostuduje s vyznamenání, dostuduje bez vyznamenání, nedostuduje

příklad: hrací kostka

- idealizovaná symetrická hrací kostka
  - homogenní materiál
  - přesná krychle
  - těžiště uprostřed
  - každá strana má stejnou pravděpodobnost
- A – padne šestka, B – padne sudé číslo
- $M = 6$
- $M_A = 1$, tedy $\mathrm{P}(A) = 1/6$
- $M_B = 3$, tedy $\mathrm{P}(B) = 3/6 = 1/2$

faktoriál

[FAKTORIÁL(n)] [factorial(n)]

- **faktoriál** $n! = n \cdot (n-1) \cdots 2 \cdot 1 \qquad 0! = 1$
- kolika způsoby lze uspořádat za sebou n rozlišitelných prvků
- příklady:
  - $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$
  - $1! = 1$
- kolika způsoby lze uspořádat za sebou 14 krajů ČR: $14! = 14 \cdot 13 \cdot 12 \cdots 2 \cdot 1 = 87\ 178\ 291\ 200 = 8{,}7 \cdot 10^{10}$
- [8.71782912e+10]

počet kombinací

[KOMBINACE(n; k)] [choose(n, k)]

- **kombinační číslo** $\binom{n}{k}$ (čti „n nad k")
- počet k-prvkových podmnožin množiny o n prvcích (tj. nezávisle na pořadí vybraných prvků)

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n\cdot(n-1)\cdots(n-k+1)}{k\cdot(k-1)\cdots 2\cdot 1}$$

- kolika způsoby si mohu z pěti knížek vybrat dvě na dovolenou:

$$\binom{5}{2} = \frac{5!}{2!3!} = \frac{5\cdot 4}{2\cdot 1} = 10$$

- kolika způsoby si z oněch pěti mohu vybrat tři knihy? (10)

příklad: losování otázek (1)

- student *neumí* 5 otázek, *umí* 10 otázek z 15 možných
- losuje se dvojice otázek z oněch 15 otázek
- pravděpodobnost $P(A)$, že student nezná ani jednu z vylosovaných:
- elementární jev: dvojice otázek
  první otázka – 15 možností, druhá jen 14 možností, nezáleží na pořadí, tedy dělit 2 (počet kombinací)

$$M = \binom{5+10}{2} = \binom{15}{2} = \frac{15!}{2!13!} = \frac{15\cdot 14}{2\cdot 1} = 105$$

- příznivé elementární jevy: vylosuje obě z pěti, které neumí

$$M_A = \binom{5}{2}\binom{10}{0} = \frac{5\cdot 4}{2\cdot 1}\cdot 1 = 10 \Rightarrow P(A) = \frac{10}{105} = 9{,}5\ \%$$

příklad: losování otázek (2)

- pravděpodobnost $P(B)$, že zná *právě* jednu otázku

$$M_B = \binom{5}{1}\cdot\binom{10}{1} = 5\cdot 10 = 50 \Rightarrow P(B) = \frac{50}{105} = 47{,}6\ \%$$

- pravděpodobnost $P(C)$, že zná *obě* otázky (*právě dvě*)

$$M_C = \binom{5}{0}\cdot\binom{10}{2} = 1\cdot\frac{10\cdot 9}{2\cdot 1} = 45 \Rightarrow P(C) = \frac{45}{105} = 42{,}9\ \%$$

- pravděpodobnost $P(D)$, že zná *aspoň jednu* otázku

$$M_D = M_B + M_C = 50 + 45 = 95 \Rightarrow P(D) = \frac{95}{105} = 90{,}5\ \%$$

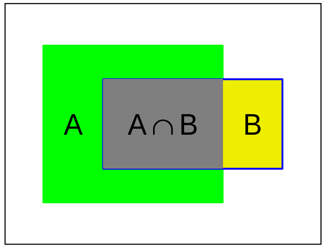
- kontrola: $M_D + M_A = M$

pravidla pro pravděpodobnost (1)

- **sjednocení** jevů $A\cup B$: platí A **nebo** B (**aspoň** jeden z jevů A, B, mohou být pravdivá obě tvrzení)
- **průnik** $A\cap B$: platí A a **současně** B (oba jevy A, B současně)

$$P(A\cup B) = P(A) + P(B) - P(A\cap B)$$

- Vennův diagram



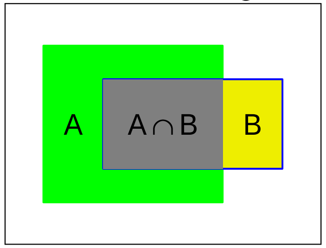
$A\cup B$ = celá vybarvená plocha
$P(A) = 0{,}42$ = zelená + šedivá plocha
$P(B) = 0{,}24$ = žlutá + šedivá plocha
$P(A\cap B) = 0{,}16$ = šedivá plocha
$P(A) + P(B)$ = (zelená + šedivá) + (žlutá + šedivá)
$P(A\cup B) = 0{,}42 + 0{,}24 - 0{,}16 = 0{,}50$

pravidla pro pravděpodobnost (2)

- $\overline{A}$ **jev opačný** k jevu A nastává právě tehdy, když nenastává jev A
$$\mathrm{P}(A) + \mathrm{P}(\overline{A}) = 1$$
- Ω – **jev jistý** nastává vždy, $\mathrm{P}(\Omega) = 1$
- $\emptyset$ – **jev nemožný** nenastává nikdy, je jevem opačným k jevu jistému, $\mathrm{P}(\emptyset) = 0$
- **neslučitelné jevy**: nemohou nastat nikdy současně, navzájem se vylučují; jejich průnikem je jev nemožný; pro neslučitelné jevy platí
$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$

podmíněná pravděpodobnost

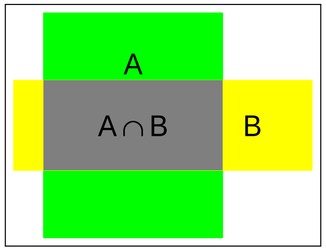
- **podmíněná pravděpodobnost** pravděpodobnost jevu A, když už jev B nastal:
$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$
- Vennův diagram



$\mathrm{P}(B) = 0{,}24 =$ žlutá + šedivá plocha
$\mathrm{P}(A \cap B) = 0{,}16 =$ šedivá plocha
$\mathrm{P}(A|B) =$ šedivá vzhledem k (žlutá + šedivá)
$\mathrm{P}(A|B) = 0{,}16/0{,}24 = 0{,}67$, ale
$\mathrm{P}(A) = 0{,}42$

nezávislost náhodných jevů

- **nezávislé jevy**: výskyt jednoho jevu **neovlivní pravděpodobnost** výskytu druhého
- (definice **nezávislosti** náhodných jevů):
$$\mathrm{P}(A) = \mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \Leftrightarrow \boxed{\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)}$$
- Vennův diagram



$\mathrm{P}(A) = 0{,}60 =$ zelená + šedivá
$\mathrm{P}(B) = 0{,}40 =$ žlutá + šedivá plocha
$\mathrm{P}(A \cap B) = 0{,}24 =$ šedivá plocha
$\mathrm{P}(A|B) =$ šedivá vzhledem k (žlutá + šedivá)
$\mathrm{P}(A|B) = 0{,}24/0{,}40 = 0{,}60$
$\mathrm{P}(A) \cdot \mathrm{P}(B) = \mathrm{P}(A \cap B)$
$\Rightarrow$ A a B jsou nezávislé

idealizovaný příklad

náhodně vybraný student ...

- A – jednička ze statistiky, $\mathrm{P}(A) = 0{,}3$
- B – jednička z matematiky, $\mathrm{P}(B) = 0{,}2$
- $A \cap B$ – jednička z obou předmětů, $\mathrm{P}(A \cap B) = 0{,}1$
- jsou jevy A, B nezávislé? (jsou jedničky ze dvou předmětů nezávislé?) NE, protože $0{,}3 \cdot 0{,}2 \neq 0{,}1$
- jaká je pst jedničky ze statistiky, když už je z matematiky?
$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{0{,}1}{0{,}2} = 0{,}5$$
- pst jedničky z matematiky, když už je ze statistiky: $\mathrm{P}(B|A) = 0{,}1/0{,}3 = 1/3$
- pravděpodobnost, že aspoň jedna jednička:
$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) = 0{,}3 + 0{,}2 - 0{,}1 = 0{,}4$$

rozdělení náhodné veličiny

- **náhodná veličina** – číselně vyjádřený výsledek náhodného pokusu
- **distribuční funkce** $F_X(x)$ náhodné veličiny X určuje pro každé x pravděpodobnost, že náhodná veličina **nepřekročí** číslo x:

$$\boxed{F_X(x) = \mathrm{P}(X \leq x)}$$

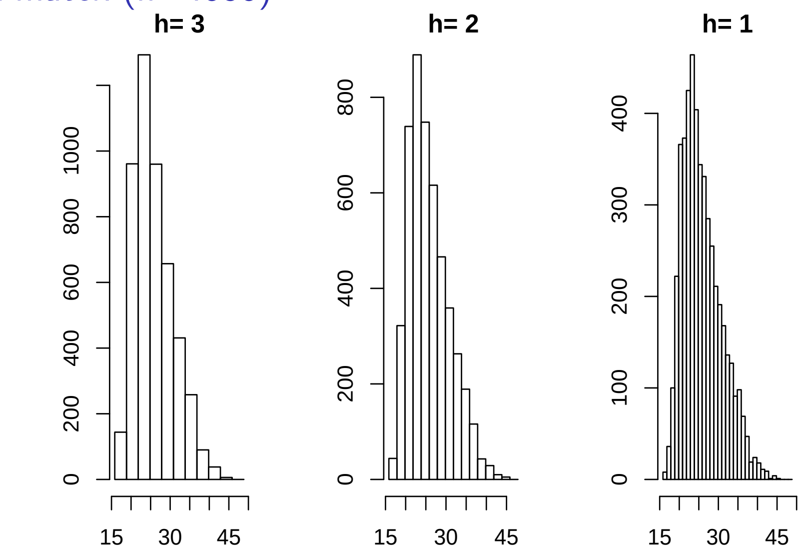
- **diskrétní rozdělení** (pro četnosti) určeno seznamem možných hodnot a jejich pravděpodobnostmi:

$$x_1, x_2, \ldots$$
$$\mathrm{P}(X = x_1), \mathrm{P}(X = x_2), \ldots$$

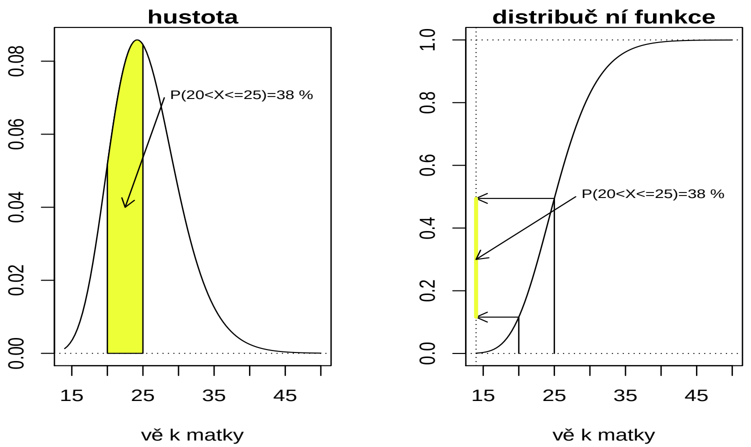
- **spojité rozdělení** (pro spojité měřítko) určeno **hustotou**

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x), \qquad F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$$

věk matek (n=4838)



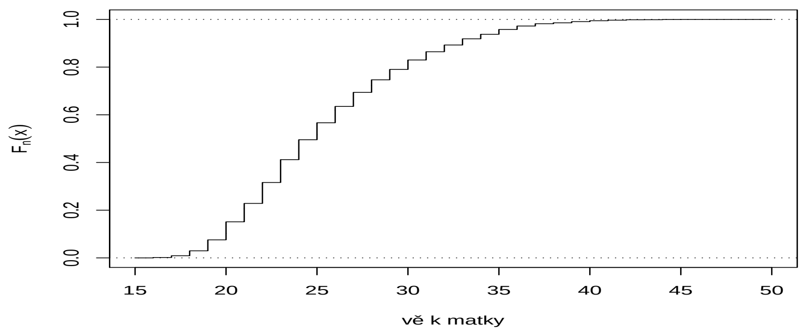
- velká populace, spojitá veličina – intervaly pro třídění mohou být krátké, obálce histogramu **relativních četností** odpovídá v idealizované představě **hustota** $f_X(x)$ [density]
- podobně **kumulativním relativním četnostem** odpovídá **distribuční funkce** [distribution function]



- bezprostředním výběrovým protějškem distribuční funkce je **empirická distribuční funkce**

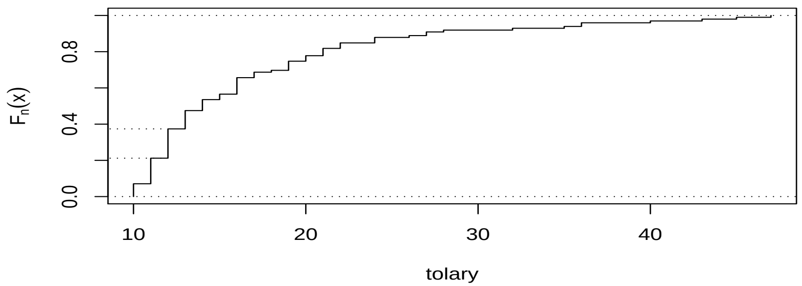
$$\boxed{F_n(x) = \frac{\#(x_i \leq x)}{n}}$$

- $x_1^* < x_2^* < \ldots < x_m^*$ existující různé hodnoty
  $n_1, n_2, \ldots, n_m$ jejich četnosti $(n = \sum_j n_j)$
  $F_n(x)$ je schodovitá funkce, v bodě x_j^* má skok n_j/n



empirická distribuční funkce (tolary)

skoky odpovídají četnostem, např. ve 12 je skok z 0,212 na 0,374 o 16/99=0,162



x_j^*	10	11	12	13	14	15	16	17	18	19	20		
n_j	7	14	16	10	6	3	9	3	1	5	3		
N_j	7	21	37	47	53	56	65	68	69	74	77		
x_j^*	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1
N_j	81	84	87	88	90	91	92	93	95	96	97	98	99

příklad diskrétního rozdělení: známky u zkoušky

X, Y známky ze dvou předmětů

známka k	1	2	3	4
$\mathrm{P}(X=k)$	0,3	0,4	0,2	0,1
$\mathrm{P}(Y=k)$	0,2	0,3	0,3	0,2

- z tabulky *nic* nepoznáme o případné závislosti X, Y
- jak jedním číslem charakterizovat úroveň známek?
- obyčejný průměr možných hodnot by X, Y nerozlišil
- použijme **vážený průměr**, kde vahami známek jsou **pravděpodobnosti možných hodnot**
- dostaneme tak **střední hodnoty** X a Y (**populační průměry**)

$$\mu_X = 1\cdot 0{,}3 + 2\cdot 0{,}4 + 3\cdot 0{,}2 + 4\cdot 0{,}1 = 2{,}1$$
$$\mu_Y = 1\cdot 0{,}2 + 2\cdot 0{,}3 + 3\cdot 0{,}3 + 4\cdot 0{,}2 = 2{,}5$$

charakteristiky rozdělení náhodné veličiny (1)

- **střední hodnota** μ_X náhodné veličiny X (populační průměr)
- je to **vážený průměr možných hodnot**
- vahami jsou pravděpodobnosti hodnot

$$\mu_X = \mathrm{E}\,X = x_1\cdot\mathrm{P}(X=x_1)+x_2\cdot\mathrm{P}(X=x_2)+\ldots = \sum_j x_j\cdot\mathrm{P}(X=x_j)$$

- operátor E (expectation) aplikovaný na náhodnou veličinu X spočítá vážený průměr jejích hodnot, vahami jsou u diskrétního rozdělení pravděpodobnosti těchto hodnot
- pro spojité rozdělení

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

- **střední hodnota funkce** $Y = g(X)$ náhodné veličiny X je vážený průměr **funkčních hodnot**

$$\mathrm{E}\,Y = \mathrm{E}\,g(X) = \sum_k g(x_k)\,\mathrm{P}(X=x_k)$$

resp. pro spojité rozdělení

$$\mathrm{E}\,Y = \mathrm{E}\,g(X) = \int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x$$

- **populační medián** $\tilde{\mu}$ spojitého rozdělení

$$F_X(\tilde{\mu}) = \mathrm{P}(X \le \tilde{\mu}) = 0{,}5$$

$\tilde{x}$ číslo, které dělí možné hodnoty náhodné veličiny na dva stejně pravděpodobné intervaly hodnot větších a menších

(populační) rozptyl náhodné veličiny X

- vážený průměr čtverců vzdáleností možných hodnot od střední hodnoty

$$\sigma_X^2 = \mathrm{E}(X-\mu_X)^2$$
$$= (x_1-\mu_X)^2\,\mathrm{P}(X=x_1) + (x_2-\mu_X)^2\,\mathrm{P}(X=x_2) + \dots$$
$$= \sum_j (x_j-\mu_X)^2\,\mathrm{P}(X=x_j)$$
$$\sigma_X^2 = \mathrm{E}(X-\mu_X)^2 = \int_{-\infty}^{\infty}(x-\mu_X)^2 f_X(x)\mathrm{d}x$$

- **(populační) směrodatná odchylka** odmocnina z (populačního) rozptylu

$$\sigma_X = \sqrt{\sigma_X^2}$$

příklad diskrétního rozdělení: známka u zkoušky

známka k	1	2	3	4	μ	σ^2	σ
$\mathrm{P}(X=k)$	0,3	0,4	0,2	0,1	2,1	0,89	0,943
$\mathrm{P}(Y=k)$	0,2	0,3	0,3	0,2	2,5	1,05	1,025

- jedním číslem charakterizovat kolísání známek (**variabilitu**)
- **(populační) rozptyl** = **vážený průměr čtverců** vzdáleností od střední hodnoty
- vahami jsou pravděpodobnosti

$$\sigma_X^2 = (1-2{,}1)^2\cdot 0{,}3 + (2-2{,}1)^2\cdot 0{,}4$$
$$+ (3-2{,}1)^2\cdot 0{,}2 + (4-2{,}1)^2\cdot 0{,}1 = 0{,}89 \doteq 0{,}943^2$$
$$\sigma_Y^2 = (1-2{,}5)^2\cdot 0{,}3 + (2-2{,}5)^2\cdot 0{,}3$$
$$+ (3-2{,}5)^2\cdot 0{,}2 + (4-2{,}5)^2\cdot 0{,}2 = 1{,}05 \doteq 1{,}025^2$$

vlastnosti střední hodnoty a rozptylu

X, Y – náhodné veličiny, a, b konstanty, $b > 0$

$$\mu_{a+X} = \mathrm{E}(a+X) = a + \mathrm{E}\,X = a + \mu_X$$
$$\mu_{b\cdot X} = \mathrm{E}(b\cdot X) = b\cdot \mathrm{E}\,X = b\cdot\mu_X$$
$$\mu_{X+Y} = \mathrm{E}(X+Y) = \mathrm{E}\,X + \mathrm{E}\,Y = \mu_X + \mu_Y$$

Návrat k průměru

$$\sigma_{a+X}^2 = \sigma_X^2, \qquad \sigma_{a+X} = \sigma_X$$
$$\sigma_{b\cdot X}^2 = b^2\sigma_X^2, \qquad \sigma_{b\cdot X} = |b|\sigma_X$$
$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{X,Y}$$

Návrat k rozptylu

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

$$= (x_1-\mu_X)(y_1-\mu_Y)\,\mathrm{P}(X=x_1, Y=y_1)$$
$$+ (x_1-\mu_X)(y_2-\mu_Y)\,\mathrm{P}(X=x_1, Y=y_2) + \dots$$

(sčítá se přes všechny možné dvojice)

příklad: známky

součet náhodných veličin: $V = X + Y$, $\sigma_V^2 \neq \sigma_X^2 + \sigma_Y^2$

	Y			
X	1	2	3	4
1	0,15	0,10	0,05	0,00
2	0,05	0,15	0,15	0,05
3	0,00	0,05	0,10	0,05
4	0,00	0,00	0,00	0,10

$$\mu_V = 2\cdot 0{,}15 + 3\cdot(0{,}10+0{,}05) + 4\cdot(0{,}05+0{,}15+0{,}00)$$
$$+ 5\cdot(0{,}15+0{,}05) + 6\cdot(0{,}05+0{,}10) + 7\cdot(0{,}05) + 8\cdot 0{,}10$$
$$= \mu_X + \mu_y = 2{,}1 + 2{,}5 = 4{,}6$$
$$\sigma_V^2 = (2-4{,}6)^2\cdot 0{,}15 + (3-4{,}6)^2\cdot 0{,}15 + (4-4{,}6)^2\cdot 0{,}20$$
$$+ (5-4{,}6)^2\cdot 0{,}20 + (6-4{,}6)^2\cdot 0{,}15 + (7-4{,}6)^2\cdot 0{,}05$$
$$+ (8-4{,}6)^2\cdot 0{,}10 = 3{,}24 \neq \sigma_X^2 + \sigma_Y^2 = 0{,}89 + 1{,}05 = 1{,}94$$

příklad: známky

sdružené a marginální rozdělení, kovariance

	Y				
X	1	2	3	4	celkem
1	0,15	0,10	0,05	0,00	0,30
2	0,05	0,15	0,15	0,05	0,40
3	0,00	0,05	0,10	0,05	0,20
4	0,00	0,00	0,00	0,10	0,10
celkem	0,20	0,30	0,30	0,20	1,00

$$\begin{aligned}\sigma_{X,Y} &= (1-2{,}1)\cdot(1-2{,}5)\cdot 0{,}15 + (1-2{,}1)\cdot(2-2{,}5)\cdot 0{,}10\\ &+ (1-2{,}1)\cdot(3-2{,}5)\cdot 0{,}05 + (1-2{,}1)\cdot(4-2{,}5)\cdot 0{,}00\\ &\dots\\ &+ (4-2{,}1)\cdot(3-2{,}5)\cdot 0{,}00 + (4-2{,}1)\cdot(4-2{,}5)\cdot 0{,}10\\ &= 0{,}65\end{aligned}$$

$$\sigma^2_V = 3{,}24 = 0{,}89 + 1{,}05 + 2\cdot 0{,}65 = \sigma^2_X + \sigma^2_Y + 2\cdot\sigma_{X,Y}$$

nezávislost náhodných veličin

- ze **sdruženého** rozdělení $\mathrm{P}(X = x_i, Y = y_j)$ dostaneme **marginální** rozdělení

$$\mathrm{P}(X = x_i) = \sum_j \mathrm{P}(X = x_i, Y = y_j)$$
$$\mathrm{P}(Y = y_j) = \sum_i \mathrm{P}(X = x_i, Y = y_j)$$

- náhodné veličiny X, Y jsou **nezávislé**, jsou-li nezávislé všechny náhodné jevy A, B, kde A je tvrzení o X a B je tvrzení o Y
- k nezávislosti X, Y stačí, když je

$$\mathrm{P}(X = x_i, Y = y_j) = \mathrm{P}(X = x_i)\cdot\mathrm{P}(Y = y_j) \quad \text{všechna } x_i, y_j$$

- jsou-li X, Y nezávislé, potom je nutně $\sigma_{X,Y} = 0$

populační korelační koeficient

modelový protějšek korelace ze str. 90

- kovariance $\sigma_{X,Y}$ modelovým protějškem s_{xy}
- populační korelační koeficient ρ_{XY} definován podobně jako r_{xy} pomocí kovariance $\sigma_{X,Y}$ a směrodatných odchylek σ_X, σ_Y

$$\rho_{XY} = \frac{\sigma_{X,Y}}{\sigma_X\sigma_Y}$$

- jsou-li X, Y nezávislé, pak je nutně $\rho_{XY} = 0$
- vždy platí $\quad -1 \le \rho_{XY} \le 1$
- příklad se známkami:

$$\rho_{XY} = \frac{\sigma_{X,Y}}{\sigma_X\sigma_Y} = \frac{0{,}65}{0{,}943\cdot 1{,}025} \doteq 0{,}67$$

známky jsou závislé

alternativní rozdělení (Bernoulliovo, nula-jedničkové)

nabývá dvou **číselných** hodnot

- diskrétní, s jediným parametrem π (nikoliv Ludolfovo číslo)
- $\mathrm{P}(X = 1) = \pi, \qquad \mathrm{P}(X = 0) = 1 - \pi \qquad (0 < \pi < 1)$
- X – kolikrát v jednom pokusu došlo k události, která má pravděpodobnost π (jen dvě možné hodnoty: 0 nebo 1)
- **střední hodnota** (populační průměr)

$$\mu_X = 1\cdot\mathrm{P}(X = 1) + 0\cdot\mathrm{P}(X = 0) = \pi$$

- (populační) **rozptyl**

$$\begin{aligned}\sigma^2_X &= (1-\mu_X)^2\,\mathrm{P}(X = 1) + (0-\mu_X)^2\,\mathrm{P}(X = 0)\\ &= (1-\pi)^2\cdot\pi + (0-\pi)^2\cdot(1-\pi)\\ &= (1-\pi)^2\pi + \pi^2(1-\pi) = \pi(1-\pi)\end{aligned}$$

binomické rozdělení bi(n, π)

- diskrétní rozdělení s parametry n, π $\qquad (0 < \pi < 1)$
- n **nezávislých** pokusů
- v každém zdar s pravděpodobností π, nezdar s pstí $1 - \pi$
- **celk. počet zdarů** X má binomické rozdělení s parametry n, π
- zapisujeme $X \sim \text{bi}(n, \pi)$
- X je součet n nezávislých náhodných veličin X_i
  (X_i = počet zdarů v i-tém pokusu)
  každé X_i má alternativní rozdělení s parametrem π
- z vlastnosti střední hodnoty součtu náh. veličin: $\boxed{\mu_X = n\pi}$
- z vlastnosti rozptylu součtu **nezávislých** náhodných veličin

$$\boxed{\sigma_X^2 = n\pi(1-\pi)}$$

binomické rozdělení bi(n, π)

- pravděpodobnosti možných hodnot

$$\mathrm{P}(X = k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}, \qquad k = 0, 1, , \dots, n$$

- pst, že v **daných** k pokusech zdar Z, v ostatních nezdar N

$$\underbrace{ZZ \dots Z}_{k}\underbrace{NN \cdots N}_{n-k} \text{ s pstí } \pi^k (1-\pi)^{n-k}$$

- zvolíme k míst pro zdar Z, na ostatních místech nezdar N, počet možností:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 2 \cdot 1}$$

příklad: zkoušky

- C – zdar = udělat zkoušku, $\mathrm{P}(C) = 0{,}8$
- zkoušku dělá $n = 10$ studentů stejně připravených (u všech stejná pravděpodobnost π), studenti neopisují (nezávislost)
- pst, že zkoušku udělá nějakých 9 studentů

$$\mathrm{P}(X = 9) = \binom{10}{9} \cdot 0{,}8^9 \cdot 0{,}2^1 = 10 \cdot 0{,}8^9 \cdot 0{,}2^1 = 0{,}268$$

- pst, že právě jeden student (nějaký) zkoušku neudělá

$$\mathrm{P}(Y = 1) = \binom{10}{1} \cdot 0{,}2^1 \cdot 0{,}8^9 = 10 \cdot 0{,}2^1 \cdot 0{,}8^9 = 0{,}268$$

- pst, že zkoušku udělá **daných** 9 studentů a zbylý ji neudělá:
  $0{,}8^9 \cdot 0{,}2^1 \doteq 0{,}0268$

příklad: kouření

- víme, že mezi dvacetiletými muži je (řekněme) 35 % kuřáků (např. je-li 70 tisíc dvacetiletých, pak je mezi nimi asi 24 500 kuřáků, ale nevíme, kteří to jsou)
- vybereme náhodně 60 dvacetiletých mužů, X – počet kuřáků mezi nimi, tedy $X \sim \text{bi}(60, 0{,}35)$
- populační průměr, rozptyl, směrodatná odchylka

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

- ukázky pravděpodobností možných hodnot
  [BINOMDIST(15;60;0,35;0)] [dbinom(15,60,0.35)]

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

Poissonovo rozdělení Po(λ)

- diskrétní rozdělení (zákon vzácných jevů), $Y \sim \text{Po}(\lambda)$
- Y – počet výskytů jevu ve zvolené časové (prostorové, plošné ...) jednotce
- $\lambda > 0$ – jediný parametr, intenzita výskytu jevu (jak často se v průměru vyskytuje ve zvolené jednotce, konstanta e $\doteq$ 2,718)

$$\text{P}(Y = k) = \frac{\lambda^k}{k!}\text{e}^{-\lambda}, \qquad k = 0, 1, \ldots$$

- střední hodnota a (populační) rozptyl jsou stejné:

$$\mu_Y = \lambda, \qquad \sigma_Y^2 = \lambda$$

- u binomického rozdělení bylo $\mu_X > \sigma_X^2$, zde rovnost

Poissonovo rozdělení Po(λ)

- parametr λ znamená hustotu na jednotku času (plochy ...) (populační průměr počtu případů na jednotku ...)
- změníme-li jednotku, změní se parametr: např. při počítání pravděpodobností toho, kolikrát najdeme případ na trojnásobku původní jednotky (trojnásobné ploše, ve trojnásobném čase ...), bude novým parametrem 3λ
- analogicky pro jiné kladné násobky
- aproximace: $X \sim \text{bi}(n, \pi)$, n velké, π malé ($\mu_X = n \cdot \pi$) pak pravděpodobnosti hodnot X lze aproximovat (přibližně vyjádřit) pomocí pravděpodobností hodnot $Y \sim \text{Po}(n \cdot \pi)$
- Poissonovo rozdělení Po($n \cdot \pi$) aproximuje binomické bi(n, π)

příklady Poissonova rozdělení

- do pasti padá za noc v průměru 8 brouků ($\lambda = 8$)
- s jakou pravděpodobností jich tam ráno najdeme 10?
  [POISSON(10;8;0)] [dpois(10,8)]

$$\text{P}(Y = 10) = \frac{8^{10}}{10!}\text{e}^{-8} = 0{,}099$$

- vezmeme-li past s polovičním obvodem, očekáváme poloviční průměr za noc ($\lambda = 4$)

$$\text{P}(Y = 10) = \frac{4^{10}}{10!}\text{e}^{-4} = 0{,}005$$
$$\text{P}(Y = 5) = \frac{4^{5}}{5!}\text{e}^{-4} = 0{,}156$$

souvislost binomického a Poissonova rozdělení

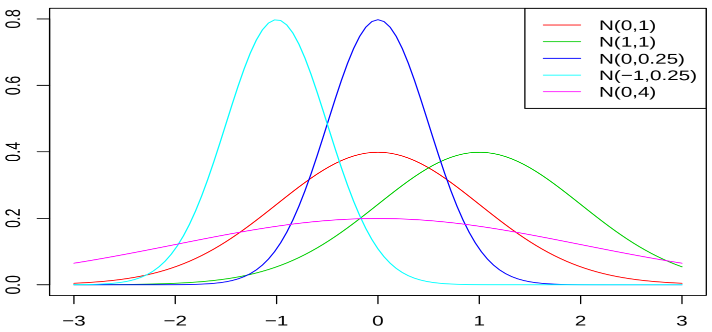
- s jakou pravděpodobností **neudělá** 12 z 50 stejně připravených studentů zkoušku? (pst neúspěchu = 0,2)
- binomické rozdělení bi(50, 0,2)
  [BINOMDIST(12;50;0,2)] [dbinom(12,50,0.2)]

$$\text{P}(X = 12) = \binom{50}{12} \cdot 0{,}2^{12} \cdot 0{,}8^{38} = 0{,}103$$

- Poissonovo rozdělení Po(50 · 0,2)=Po(10)
  [POISSON(12;10;0)] [dpois(12,10)]

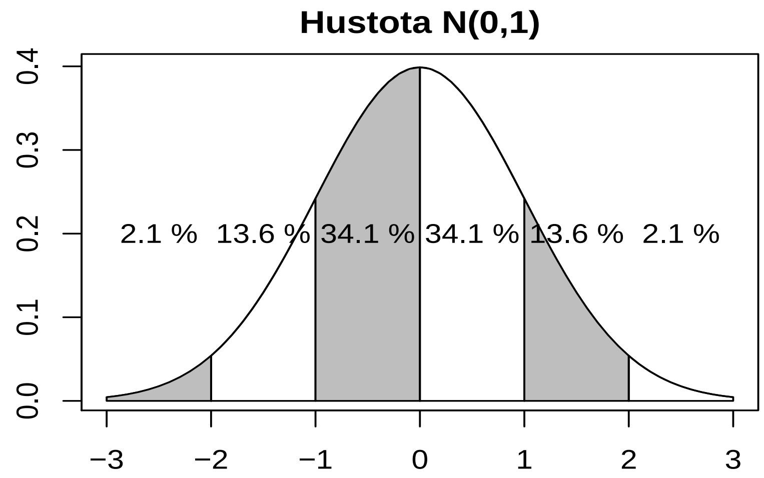
$$\text{P}(Y = 12) = \frac{10^{12}}{12!}\text{e}^{-10} = 0{,}095$$

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



- spojité rozdělení, symetrické okolo střední hodnoty μ
- maximální hodnota hustoty je úměrná $1/\sigma$ $\left(\frac{1}{\sqrt{2\pi\sigma^2}} \doteq \frac{0{,}4}{\sigma}\right)$
- model vzniku: součet velkého počtu nepatrných příspěvků

normované normální rozdělení $Z \sim \mathrm{N}(0, 1)$



příklady pravděpodobností o normálním rozdělení

- pro $X \sim \mathrm{N}(\mu, \sigma^2)$ platí

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

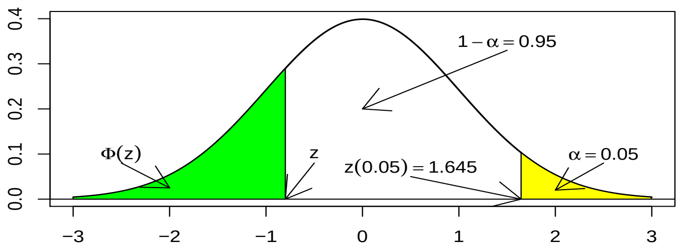
- $\boxed{X \sim \mathrm{N}(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1)}$
- $$\mathrm{P}(|Z| < c) = \mathrm{P}\left(\left|\frac{X-\mu}{\sigma}\right| < c\right) = \mathrm{P}(|X - \mu| < c \cdot \sigma)$$
- tedy

$$\mathrm{P}(|X - \mu| < 1{,}00\ \sigma) = 0{,}683, \text{ tj. } 68{,}3\ \%$$
$$\mathrm{P}(|X - \mu| < 1{,}96\ \sigma) = 0{,}95, \text{ tj. } 95\ \%$$
$$\mathrm{P}(|X - \mu| < 2{,}00\ \sigma) = 0{,}954, \text{ tj. } 95{,}4\ \%$$
$$\mathrm{P}(|X - \mu| < 3{,}00\ \sigma) = 0{,}997, \text{ tj. } 99{,}7\ \%$$

normované normální rozdělení $Z \sim \mathrm{N}(0, 1)$

tabelováno:

- hustota $\varphi(z)$
  [NORMDIST(z;0;1)] [dnorm(z)]
- distribuční funkce $\Phi(z) = \mathrm{P}(Z \leq z)$
  [NORMSDIST(z)] [pnorm(z)]
- **kritické hodnoty** $z(\alpha)$: $\mathrm{P}(Z \leq z(\alpha)) = \Phi(z(\alpha)) = 1 - \alpha$
  [NORMSINV($1 - \alpha$)] [qnorm($1 - \alpha$)]



zajímavé kritické hodnoty

$$
\begin{aligned}
z(0{,}025) &= 1{,}96 \text{ tj. } \mathrm{P}(Z > 1{,}96) = 2{,}5\ \%\\
z(0{,}025) &= 1{,}96 \text{ tj. } \mathrm{P}(Z < -1{,}96) = 2{,}5\ \%\\
z(0{,}025) &= 1{,}96 \text{ tj. } \mathrm{P}(|Z| > 1{,}96) = 5\ \%\\
z(0{,}005) &= 2{,}58 \text{ tj. } \mathrm{P}(Z > 2{,}58) = 0{,}5\ \%\\
z(0{,}005) &= 2{,}58 \text{ tj. } \mathrm{P}(Z < -2{,}58) = 0{,}5\ \%\\
z(0{,}005) &= 2{,}58 \text{ tj. } \mathrm{P}(|Z| > 2{,}58) = 1\ \%\\
z(0{,}050) &= 1{,}64 \text{ tj. } \mathrm{P}(Z > 1{,}64) = 5\ \%\\
z(0{,}050) &= 1{,}64 \text{ tj. } \mathrm{P}(Z < -1{,}64) = 5\ \%\\
z(0{,}050) &= 1{,}64 \text{ tj. } \mathrm{P}(|Z| > 1{,}64) = 10\ \%
\end{aligned}
$$

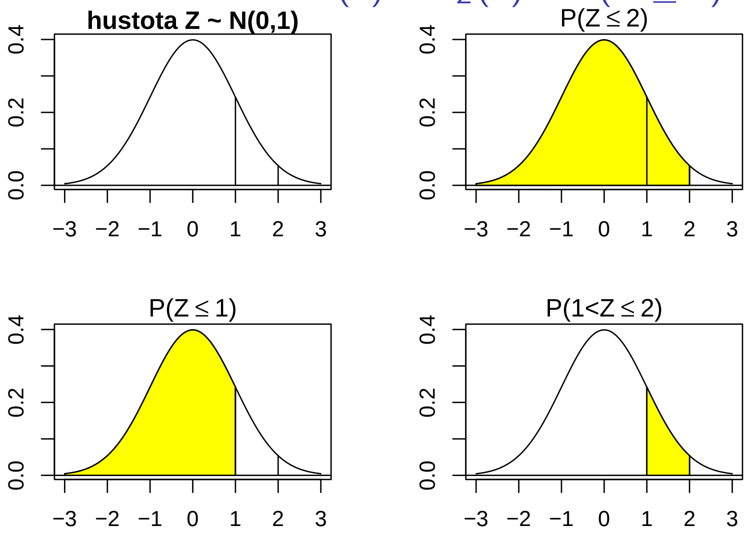
výpočet pravděpodobností pro $Z \sim \mathrm{N}(0,1)$

- u spojitého rozdělení je $\mathrm{P}(X < x) = \mathrm{P}(X \le x)$, tedy i u Z
- $Z \sim \mathrm{N}(0,1), a < b$, pak $\boxed{\mathrm{P}(a < Z < b) = \Phi(b) - \Phi(a)}$
- odvození: jevy $(Z \le a)$ a $(a < Z \le b)$ jsou neslučitelné (tvrzení nemohou platit současně)
  jejich sjednocením je jev $(Z \le b)$, proto

$$
\begin{aligned}
\mathrm{P}(Z \le b) &= \mathrm{P}(Z \le a) + \mathrm{P}(a < Z \le b)\\
\Phi(b) &= \Phi(a) + \mathrm{P}(a < Z \le b)
\end{aligned}
$$

- příklad: $\mathrm{P}(1 < Z < 2) = \Phi(2) - \Phi(1) = 0{,}977 - 0{,}841 = 0{,}136$, jak bylo na obrázku
  [NORMSDIST(2)-NORMSDIST(1)] [pnorm(2)–pnorm(1)]

Postup výpočtu $\mathrm{P}(1 < Z < 2)$ $(Z \sim \mathrm{N}(0,1))$ pomocí tabelované funkce $\Phi(z) = F_Z(z) = \mathrm{P}(Z \le z)$



výpočet pro $X \sim \mathrm{N}(\mu, \sigma^2)$

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

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

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

příklad: $X \sim \mathrm{N}(136{,}1, 6{,}4^2)$ (výšky 10letých hochů v roce 1951)

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

tedy v rozmezí 135 cm až 140 cm bylo asi 35,3 % hochů

pohodlnější možnost

- $X \sim \mathrm{N}(136{,}1, 6{,}4^2)$
- počítáme $\mathrm{P}(134{,}5 < X < 140{,}5)$
- Excel i R nabízejí možnost dosadit skutečné parametry normálního rozdělení
- druhým parametrem je **směrodatná odchylka**
- Excel (nepřehlédněte, že nejde o NORMSDIST!): [NORMDIST(140,5;136,1;6,4;1)-NORMDIST(134,5;136,1;6,4;1)]
- R: (pozor, na vstupu nutně desetinná **tečka**, čárka je oddělovač parametrů) [pnorm(140.5,136.1,6.4)-pnorm(134.5,136.1,6.4)]

populace a výběr

- populaci (základní soubor) charakterizujeme pomocí parametrů rozdělení, případně typu rozdělení
- výsledkem měření na náhodně vybraném prvku populace (základního souboru) je náhodná veličina
- skutečné hodnoty parametrů neznáme
  - chceme parametry odhadnout
  - chceme rozhodnout o platnosti tvrzení (hypotézy) o parametrech nebo o typu rozdělení
- jako výběr si představujeme několik **nezávislých** náhodných veličin se stejným rozdělením (možná s neznámými parametry)
  - parametry odhadujeme na základě výběru
  - o hypotézách rozhodujeme na základě výběru
- příklady
  - střední hodnotu náhodné veličiny (populační průměr) odhadujeme pomocí výběrového průměru
  - rozptyl náhodné veličiny odhadujeme pomocí výběrového rozptylu

chování výběrového průměru z náhodného výběru

- nechť $X_1, X_2, \dots X_n$ jsou nezávislé náhodné veličiny s **libovolným stejným rozdělením** se střední hodnotou μ a rozptylem σ^2, tj. **náhodný výběr** z onoho rozdělení
- **průměr** $X_1, X_2, \dots X_n$: $\boxed{\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i}$
- připomeňme vlastnosti střední hodnoty, Vlastnosti zejména

$$\mu_{X+Y} = \mu_X + \mu_Y, \qquad \mu_{b\cdot X} = b \cdot \mu_X$$

- proto je

$$\mu_{\overline{X}} = \mu_{\frac{1}{n}\cdot\sum_{i=1}^{n} X_i} = \frac{1}{n}\cdot\mu_{\sum_{i=1}^{n} X_i} = \frac{1}{n}\sum_{i=1}^{n}\mu_{X_i} = \frac{1}{n} n\mu = \mu$$

- $\mu_{\overline{X}} = \mu$, tj. $\overline{X}$ je **nestranný odhad** parametru μ

variabilita výběrového průměru

- pro rozptyl **nezávislých** náhodných veličin platí Vlastnosti

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y \qquad \sigma^2_{b\cdot X} = b^2\sigma^2_X$$

- proto je

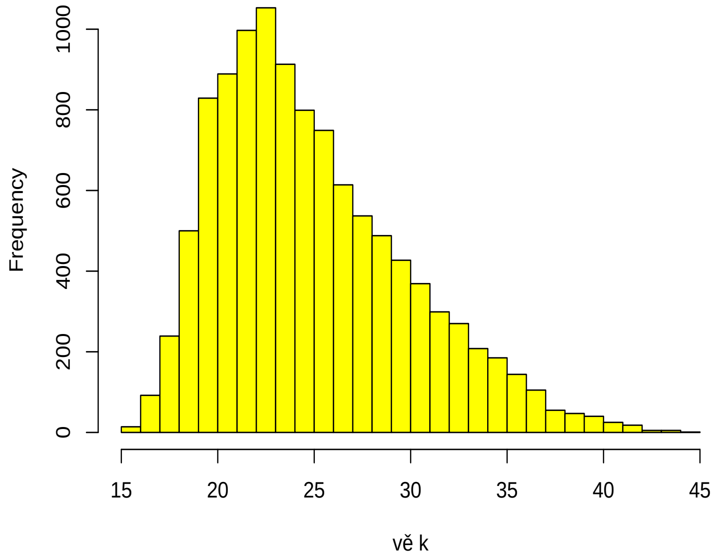
$$\sigma^2_{\overline{X}} = \sigma^2_{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2_{X_i} = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

- průměr $\overline{X}$ má tedy rozptyl n-krát menší, než jednotlivá pozorování
- **střední chyba průměru** = směrodatná odchylka průměru

$$\boxed{\mathrm{S.E.}(\overline{X}) = \frac{\sigma}{\sqrt{n}}}$$

- **střední chyba odhadu** nějakého parametru = směrodatná odchylka tohoto odhadu

příklad: věk matek



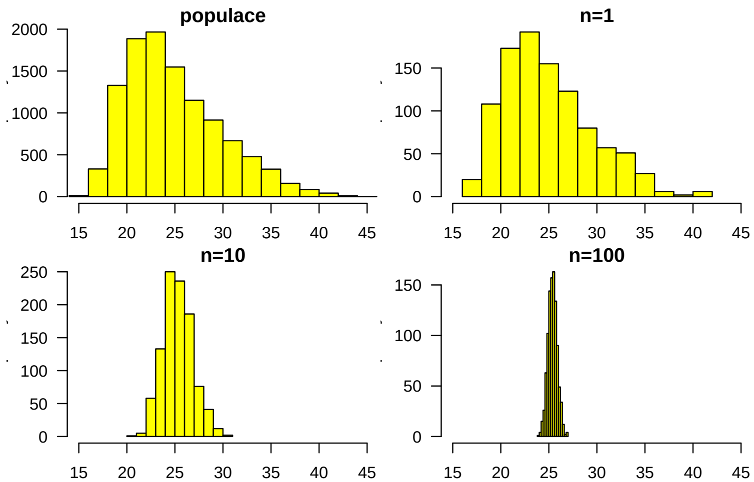
příklad: věk matek

- výjimečný umělý příklad, kdy známe celou populaci
- populace obsahuje 10 916 hodnot
- rozdělení věku je výrazně nesymetrické
- prováděn výběr rozsahu n, vždy spočítán průměr
- výběr Nkrát opakujeme (spočítáno $N = 1000$ průměrů)
- spočítány charakteristiky z N průměrů jako výchozích hodnot, (modře charakteristiky celé populace nebo hodnoty z nich odvozené)

n	průměr	sm. odch.	$\sigma/\sqrt{n}$	šikmost	špičatost
1	25,43	4,62	4,94	0,74	0,29
10	25,35	1,54	1,56	0,28	-0,04
100	25,39	0,48	0,49	0,08	-0,05
populace	$\mu =$25,40	$\sigma =$ 4,94	4,94	0,77	0,19

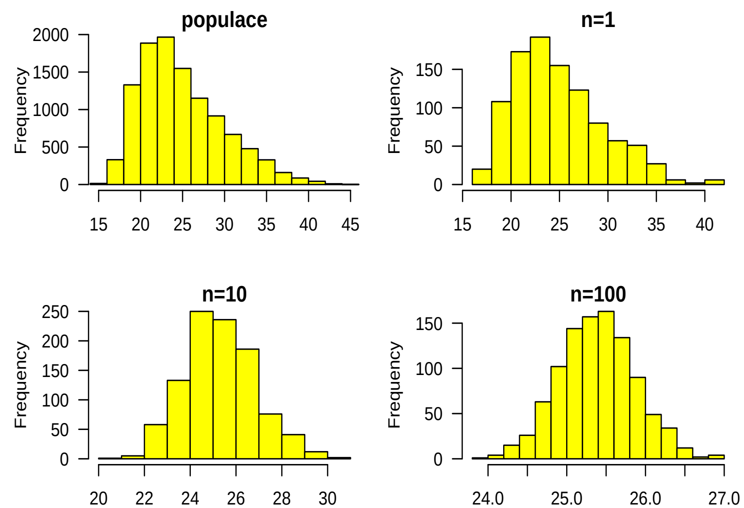
příklad: histogram populace a histogramy průměrů

šířky intervalů stejné, variabilita průměrů s rostoucím n klesá



příklad: histogram populace a histogramy výběrů

šířky intervalů přizpůsobené variabilitě, s rostoucím n se zlepšuje normalita



příklad: shrnutí

- průměry kolísají kolem populačního průměru μ
- směrodatné odchylky s rostoucím $\sqrt{n}$ klesají
- šikmost a špičatost se s rostoucím n blíží k nule
- je naděje, že s rostoucím n je histogram podobnější hustotě normálního rozdělení – projev *centrální limitní věty*

centrální limitní věta (CLV)

- vlastnost součtu nezávislých náhodných veličin se stejným rozdělením (s populačním průměrem μ, popul. rozptylem σ^2)
- průměr je součet dělený počtem sčítanců
  $\Rightarrow$ pro průměr platí CLV také
- standardizovaný součet (průměr) n nezávislých náhodných veličin lze pro velké n aproximovat normálním rozdělením $\mathrm{N}(0,1)$ CLV pro četnosti

$$\boxed{Z = \frac{\sum_{i=1}^{n} X_i - n\cdot\mu}{\sigma\sqrt{n}} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \,\dot\sim\, \mathrm{N}(0,1)}$$

- pro velká n se výběrový průměr chová, jako by šlo o výběr z normálního rozdělení, a to bez ohledu na výchozí rozdělení

$$\boxed{\overline{X} \,\dot\sim\, \mathrm{N}\left(\mu, \sigma^2/n\right)}$$

interval spolehlivosti pro populační průměr μ

- pro nezávislé náhodné veličiny $X_1,\dots,X_n \sim \mathrm{N}\left(\mu,\sigma^2\right)$ platí

$$\overline{X} \sim \mathrm{N}\left(\mu, \sigma^2/n\right)$$

- proto je $$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$
- použijeme kritickou hodnotu

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

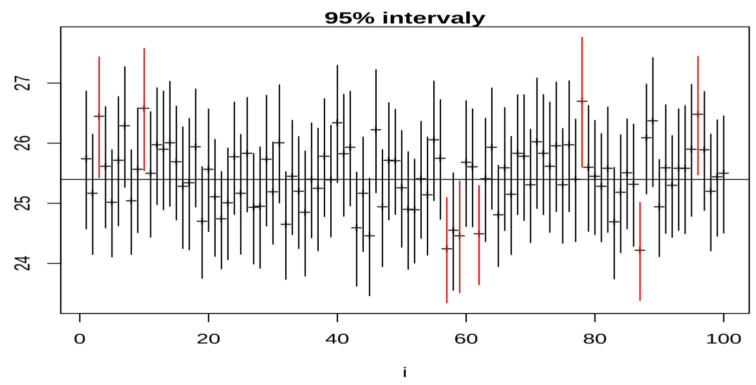
- hodnota neznámého parametru μ je s pravděpodobností $1-\alpha$ pokryta intervalem

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

- lze použít pro velká n i bez požadavku na normální rozdělení X_i

100 intervalů spolehlivosti ($n = 100$, $1-\alpha = 95$ %)

(v 8 případech interval **neobsahuje** μ)



příklad: IQ vysokoškoláků

- u $n = 16$ náhodně vybraných studentů jisté fakulty byla zjištěna hodnota IQ
- metoda měření IQ je konstruována tak, že je $\sigma = 15$
- vyšel průměr $\overline{x} = 110$
- co lze říci o populačním průměru všech studentů oné velké fakulty?
- 95% interval spolehlivosti ($z(0{,}025) = 1{,}96$):

$$(110 - \frac{15}{4} \cdot 1{,}96; 110 + \frac{15}{4} \cdot 1{,}96) = (102{,}65; 117{,}35)$$

- skutečný populační průměr μ (všech studentů oné fakulty) leží s 95% pravděpodobností mezi 102,65 a 117,35
- μ leží s 90% pravděpodobností mezi 103,83 a 116,17

vlastnosti intervalu spolehlivosti pro μ

- délka intervalu roste s požadovanou spolehlivostí
  - 90% interval (103,83; 116,17) má délku 12,34
  - 95% interval (102,65; 117,35) má délku 14,70
- délka intervalu klesá s rostoucím počtem pozorování n
  - pro $n = 16$ má 95% interval (102,65; 117,35) délku 14,70
  - pro $n = 4 \cdot 16 = 64$ má 95% interval (106,325; 113,675) délku 7,35, tedy poloviční ($1/\sqrt{4} = 1/2$)
- kolik potřebujeme pozorování, aby měl 95% interval délku 2δ?

$$\frac{\sigma}{\sqrt{n}} z(\alpha/2) = \delta \qquad \Rightarrow \qquad n = \left(\frac{\sigma}{\delta} z(\alpha/2)\right)^2$$

- v příkladu s IQ požadujeme $\delta = 5$:

$$n = \left(\frac{15}{5} 1{,}96\right)^2 \doteq 35$$

interval spolehlivosti pro μ (neznámé σ)

- neznáme-li σ, nahradíme je pomocí (výběrová směr. odchylka)

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

- interval spolehlivosti pro μ:

$$\boxed{\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha); \overline{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha)\right)}$$

- použití kritické hodnoty $t_{n-1}(\alpha)$ Studentova t-rozdělení místo kritické hodnoty $z(\alpha/2)$ je penalizací za to, že neznámou směrodatnou odchylku σ jsme nahradili jejím odhadem S
- platí totiž $t_{n-1}(\alpha) > z(\alpha/2)$, s rostoucím n se rozdíl zmenšuje
- délku intervalu spolehlivosti určuje zejména střední chyba průměru $S/\sqrt{n}$

příklad: výška postavy

- studenti odhadovali výšku přednášejícího; předpokládejme, že nestranně a nezávisle na sobě
- $n = 22$, $\quad \overline{x} = 172{,}4$, $\quad s_x = 4{,}032$
- z tabulek: $t_{21}(0{,}05) = 2{,}080$

$$(172{,}4 - \frac{4{,}032}{\sqrt{22}} \cdot 2{,}080; 172{,}4 + \frac{4{,}032}{\sqrt{22}} \cdot 2{,}080)$$

$$(170{,}6; 174{,}2)$$

- $\Rightarrow$ skutečná výška je s pravděpodobností 95 % někde mezi 170,7 cm a 174,2 cm

centrální limitní věta pro četnosti

- co říkala CLV? ▸ CLV
- absolutní četnost Y
  - Y – součet nezávislých veličin s alternativním rozdělením
  - $Y = \sum_{i=1}^{n} X_i$
  - populační průměr X_i je π
  - populační rozptyl X_i je $\pi(1-\pi)$
  - $Y \sim \mathrm{bi}(n, \pi)$, proto přibližně $Y \sim \mathrm{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $f = Y/n$
  - f – průměr nezávislých veličin s alternativním rozdělením
  - $f \sim \mathrm{N}(\pi, \pi(1-\pi)/n)$

příklad: počet studentek

- zkušenost: mezi uchazeči o studium bývá 45 % dívek
- s jakou pravděpodobností bude při 500 přihláškách počet dívek mezi 200 a 220 (včetně)?
- $Y \sim \mathrm{bi}(500, 0{,}45)$ má $\mu_Y = 500 \cdot 0{,}45 = 225$, $\sigma_Y^2 = 500 \cdot 0{,}45 \cdot 0{,}55 = 123{,}75$, tedy $\sigma_Y = 11{,}1$

$$\mathrm{P}(200 \leq Y \leq 220) \doteq \Phi\left(\frac{220{,}5 - 225}{11{,}1}\right) - \Phi\left(\frac{199{,}5 - 225}{11{,}1}\right)$$

- hledaná pravděpodobnost je přibližně 33,2 % (přesně 33,3 %)
  [NORMDIST(220,5;225;11,1243;1) – NORMDIST(199,5;225;11,1243;1)]
  [pnorm(220.5,500*0.45,sqrt(500*0.45*0.55))
  – pnorm(199.5,500*0.45,sqrt(500*0.45*0.55))]
  [BINOMDIST(220;500;0,45;1) – BINOMDIST(199;500;0,45;1)]
  [pbinom(220,500,0.45) – pbinom(199,500,0.45)]

interval spolehlivosti pro podíl (pravděpodobnost) π

- π – podíl prvků populace s danou vlastností
- π – pst, s jakou takový prvek vylosujeme
- počet prvků náhodně vybraných s onou vlastností $Y \sim \mathrm{bi}(n, \pi)$
- střední chyba relativní četnosti $Y/n = f$
  $=$ směrodatná odchylka relativní četnosti f
  $=$ odmocnina z rozptylu relativní četnosti f, tedy $\sqrt{\frac{\pi(1-\pi)}{n}}$
- pravděpodobnost π neznáme, odhadneme ji pomocí f
- odtud je přibližný 95% interval spolehlivosti pro π

$$\left(f - 1{,}96 \cdot \sqrt{\frac{f(1-f)}{n}}; f + 1{,}96 \cdot \sqrt{\frac{f(1-f)}{n}}\right)$$

- skutečná pst π je tedy s 95% pstí v uvedeném rozmezí
- existuje přesnější (pracnější) postup

příklad: hody s hrací kostkou

- odhadujeme pravděpodobnost šestky
- kostka A: $n = 100, n_A = 17, f_A = 0{,}17$

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

$$(0{,}10; 0{,}24)$$

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

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

$$(0{,}31; 0{,}51)$$

- důležitý rozdíl: u kostky A patří $1/6 = 0{,}167$ do intervalu spolehlivosti; u kostky B nikoliv; může to něco znamenat?

proč **testování hypotéz**

- připomeňme 95% intervaly spolehlivosti pro šestku u kostek:
  - kostka A: (0,10; 0,24)
  - kostka B: (0,31; 0,51)
- znamená něco, když $1/6 \doteq 0{,}167$ leží či neleží v 95% intervalu spolehlivosti?
- nelze bezpečně poznat, že kostka A není falešná nebo že kostka B je falešná
- intervaly spolehlivosti určily rozmezí, kde by skutečná pravděpodobnost šestky měla být, spolehlivost intervalů je velká, ale omezená
- musíme připustit, že jsme mohli mít smůlu, že se v našich pokusech náhodou realizovaly málo pravděpodobné možnosti, přestože k takové smůle dochází jen zřídka
- potřebujeme **standardizovaná pravidla**, jak rozhodovat

hypotézy a možná rozhodnutí

- možné statistické **hypotézy**
  - **(nulová) hypotéza** H_0: – zjednodušuje situaci porovnávané populace se **ne**liší, vyšetřované znaky jsou **ne**závislé ...
    tedy žádný (tj. **nulový**) rozdíl, žádná (tj. **nulová**) závislost
    zpravidla se snažíme H_0 vyvrátit, abychom věcně něco prokázali
  - **alternativa** H_1: **(alternativní hypotéza)** – opak nulové hypotézy, zpravidla to, co chceme věcně dokázat
  - volba co je H_0 je pevně spojena s testem, nezávisí na nás; volíme H_0, ta nabídne test
- možná **rozhodnutí**
  - **zamítnout** H_0 pokud naše data svědčí proti H_0
  - **nezamítnout** H_0 (přijmout H_0) pokud *není dost důvodů* H_0 zamítnout
- hypotéza – tvrzení o **populaci** (základním souboru)
- rozhodujeme na základě dat z **výběru**
- nelze zaručit bezchybnost rozhodnutí

chyby v rozhodování

- nelze zaručit bezchybnost rozhodnutí, mohou nastat chyby:
  - **chyba 1. druhu**, když zamítneme platnou (pravdivou) hypotézu H_0
  - **chyba 2. druhu**, když nepoznáme, že hypotéza H_0 neplatí a nezamítneme ji (přijmeme ji)
- nechceme příliš často *chybně* zamítat H_0, dělat chybu 1. druhu (tedy falešně něco věcně prokazovat)
- proto se snažíme chybě 1. druhu pokud možno vyvarovat, i když ji nelze vyloučit
- **hladina testu** α = maximální přípustná pravděpodobnost chyby 1. druhu (zpravidla $\alpha = 0{,}05$, tj. $\alpha = 5$ %)
- **síla testu** = pravděpodobnost správného zamítnutí neplatné hypotézy

schéma rozhodování

rozhodnutí	H_0 platí	H_0 neplatí
H_0 zamítnout	chyba 1. druhu (pst $\leq \alpha$) hladina testu	správné rozhodnutí (pst $1-\beta$) síla testu
H_0 nezamítnout (přijmout)	správné rozhodnutí (pst $\geq 1-\alpha$)	chyba 2. druhu (pst β)

- na základě dat volíme rozhodnutí (řádek)
- nevíme, jaká skutečnost (sloupec) platí

klasický postup při rozhodování

- zvolit (nulovou) hypotézu H_0, alternativu H_1
- zvolit hladinu testu α (zpravidla 5 %)
- zvolit metodu rozhodování (který test použít)
- z dat spočítat testovou statistiku T a porovnat ji s tabelovanou kritickou hodnotou (bude ještě: porovnat p-hodnotu s hladinou α)
- **kritický obor** – množina těch výsledků pokusu (např. hodnot T), kdy budeme hypotézu zamítat
- když padne statistika T do **kritického oboru**, pak hypotézu zamítnout (zpravidla, když $T \geq t_0$, kde t_0 je kritická hodnota)

příklad: padá na kostce šestka příliš často?

- chceme na 5% hladině prokázat, že pravděpodobnost šestky na dané kostce je větší, než by měla být (tj. větší než 1/6)
- $H_0 : P(\text{padne šestka}) = 1/6 \qquad (\pi = \pi_0)$
- $H_1 : P(\text{padne šestka}) > 1/6 \qquad (\pi > \pi_0)$
- provedeme $n = 100$ pokusů, Y je počet šestek
- co svědčí pro neplatnost hypotézy? Je to situace, kdy „šestka padá mnohem častěji, než by měla padat za H_0“
- **tvar kritického oboru**: hypotézu zamítat, když $Y \geq y_0$
- za platnosti H_0 má počet šestek Y rozdělení $\text{bi}(n, 1/6)$
- **velikost kritického oboru**: y_0 zvolíme tak, abychom hypotézu za její platnosti zamítali s pravděpodobností nejvýše α, tj.

$$P_0(Y \geq y_0) \leq \alpha$$

příklad: jak zvolit kritickou hodotu y_0?

- některé pravděpodobnosti pro $Y \sim \text{bi}(100, 1/6)$

y_0	20	21	22	23	**24**	25
$P(Y \geq y_0)$	0,220	0,152	0,100	0,063	**0,038**	0,022

- podmínku $P(Y \geq y_0) \leq 0{,}05 = \alpha$ splňuje $y_0 = 24$
- padne-li ve 100 nezávislých hodech kostkou aspoň 24 šestek, budeme na **5% hladině zamítat hypotézu**, že pst šestky je 1/6 **ve prospěch alternativy**, že pst šestky je větší než 1/6 (dáno zvolenou alternativou)
- na kostce A nám padlo 17 šestek, hypotézu **nezamítáme**, to ale neznamená, že bychom hypotézu prokázali
- na kostce B nám padlo 41 šestek, hypotézu **zamítáme**
- pro $\alpha = 10$ % bychom zvolili $y_0 = 22$, bylo by však větší riziko zamítnutí platné hypotézy

příklad: síla testu

- **síla testu** = pst, že hypotézu zamítneme, když ona neplatí
- při 100 hodech hypotézu na 5% hladině zamítáme, je-li $Y \geq 24$
- nechť je ve skutečnosti $\pi = 1/4$, pak hypotézu zamítneme (výsledek pokusu padne do kritického oboru) s pstí

$$P(Y \geq 24) = \sum_{k=24}^{100} \binom{100}{k} \left(\frac{1}{4}\right)^k \left(1 - \frac{1}{4}\right)^{100-k} = 0{,}629$$

[1-BINOMDIST(23;100;1/4;1)] [1-pbinom(23,100,1/4)]

- pro $\pi = 0{,}25$ je tedy síla testu 62,9 %
- pro $\pi = 0{,}3$ je podobně síla testu rovna 92,4 %

rozhodování pomocí p-hodnoty

- p-**hodnota** p je nejmenší α, při kterém H_0 z daných dat ještě zamítáme
- p-hodnota p je za platnosti H_0 spočítaná *pravděpodobnost* výsledků stejně nebo *méně příznivých* pro H_0, než ten, který opravdu nastal
- H_0 zamítáme právě tehdy, když je $\boxed{p \leq \alpha}$
- p-hodnotu počítají moderní počítačové programy
- existují úlohy, kdy se rozhoduje pouze podle p-hodnoty (např. Fisherův exaktní test ve čtyřpolní tabulce)
- statistické rozhodování:
  spočítat k T odpovídající p-hodnotu a porovnat ji s α

příklad: rozhodování pomocí p-hodnoty

- snažíme se prokázat, že šestka padá příliš často ($H_1 : \pi > 1/6$)
- hypotéza $H_0 : \pi = 1/6$, kritický obor: $Y \geq y_0 = 24$
- padlo nám $Y = 17$, proto (psti binomického rozdělení)

$$p = P(Y \geq 17) = \sum_{k=17}^{100} \binom{100}{k} \left(\frac{1}{6}\right)^k \left(1 - \frac{1}{6}\right)^{100-k} = 0{,}506$$

$$= 1 - P(Y \leq 16) \qquad [\text{1-BINOMDIST(16;100;1/6;1)}]$$

- protože 50,6 % > 5 %, hypotézu nemůžeme na 5% hladině zamítnout, nemůžeme tvrdit, že pst šestky je větší než 1/6
- neprokázali jsme však, že by hypotéza platila
- na kostce B: $p = P(Y \geq 41) = 1 - P(Y \leq 40) = 7{,}4 \cdot 10^{-9}$
  hypotézu zamítáme [1-pbinom(40,100,1/6)]

příklad: kostka a oboustranná alternativa

- chceme ověřit, zda je kostka v pořádku
- pokusíme se prokázat, že pst šestky je větší než 1/6 (pak šestka padá příliš často) **nebo** je menší než 1/6 (padá příliš zřídka) (**oboustranná alternativa**)
- H_0 : P(padne šestka) = 1/6 $\quad (\pi = \pi_0)$
- H_1 : P(padne šestka) $\neq$ 1/6 $\quad (\pi \neq \pi_0)$
- *proti* hypotéze svědčí malé *nebo* velké hodnoty Y
- pst chyby 1. druhu α rozdělíme na dvě poloviny:
  $\alpha/2$ pro příliš malé Y, $\alpha/2$ příliš velké Y

příklad: kostka, oboustranná alternativa

y_0	8	9	10	...	24	25	26
$P(Y \leq y_0)$	0,010	0,021	0,043	...	0,978	0,988	0,994
$P(Y \geq y_0)$	0,996	0,990	0,979	...	0,038	0,022	0,012
$P(Y = y_0)$	0,006	0,012	0,021	...	0,016	0,010	0,006

- $\alpha = 0{,}05$, tj. $\alpha/2 = 0{,}025$ (resp. $\alpha = 0{,}1$, tj. $\alpha/2 = 0{,}05$)
- H_0 zamítneme, když bude $Y \leq 9$ *nebo* když bude $Y \geq 25$
- skutečná pst chyby 1. druhu bude $0{,}021 + 0{,}022 = 0{,}043$
- [pbinom(9,100,1/6)+(1-pbinom(24,100,1/6))]
  [BINOMDIST(9;100;1/6;1) + 1-BINOMDIST(24;100;1/6;1)]
- hodnoty v rozmezí 10 až 24 (včetně mezí) nesvědčí proti H_0

oboustranná alternativa (přibližně)

- $\mathrm{H}_0 : \pi = \pi_0$, např. P(padne šestka) $= 1/6$
- $\mathrm{H}_1 : \pi \neq \pi_0$, např. P(padne šestka) $\neq 1/6$
- proti hypotéze svědčí Y hodně daleko od $\mu_Y = n\pi_0$ (počítáme za platnosti hypotézy), tj. rel. četnost $f = Y/n$ daleko od π_0
- zavedeme

$$Z = \frac{Y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{f - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n}$$

- hypotézu zamítneme, bude-li Z daleko od nuly: $|Z| \geq z(\alpha/2)$
- pro $\alpha = 5$ % zamítáme hypotézu, je-li $|Z| \geq 1{,}96$
- $z_A = 0{,}089$ (nezamítneme), $z_B = 6{,}529$ (zamítneme)

změnila se za deset roků výška desetiletých hochů?

- v roce 1951 byla průměrná výška desetiletých hochů 136,1 cm (zjištěno z velkého výběru o tisících měření)
- v roce 1961 bylo změřeno 15 náhodně vybraných desetiletých hochů: 127 130 133 136 136 138 139 139 139 140 141 142 147 149 151
- $\bar{X} = 139{,}13$ cm, $n = 15$
- znamená to, že po deseti letech jsou desetiletí opravdu vyšší?
- co je výška desetiletých hochů? (**populační průměr**)
- stačí k důkazu, že 10 hochů je větších než 136,1 cm a jen 5 menších než 136,1 cm?
- stačí k důkazu, že nový **výběrový** průměr je o 3 cm větší než hypotézou předpokládaný **populační** průměr?

test o střední hodnotě μ normálního rozdělení

- předpokládáme $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, nezávislé
- $\sigma > 0$ odhadneme pomocí $S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- rozptyl $\bar{X}$ odhadneme pomocí S_x^2/n, střední chybu $\bar{X}$ (odmocnina z rozptylu) odhadneme jako $\widehat{\mathrm{S.E.}}(\bar{X}) = S_x/\sqrt{n}$
- $\mathrm{H}_0 : \mu = \mu_0$ (μ_0 známá konstanta)

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

statistika T má za H_0 Studentovo t-rozdělení s $n-1$ st. vol.

- kdy hypotézu H_0 zamítáme (kritický obor):
  - $\mathrm{H}_1 : \mu \neq \mu_0$ (oboustranná alternativa) $|T| \geq t_{n-1}(\alpha)$
  - $\mathrm{H}_1 : \mu > \mu_0$ (jednostranná alternativa) $T \geq t_{n-1}(2\alpha)$
  - $\mathrm{H}_1 : \mu < \mu_0$ (jednostranná alternativa) $T \leq -t_{n-1}(2\alpha)$

souvislost s intervalem spolehlivosti

- připomeňme interval spolehlivosti pro μ

$$\bar{X} - \widehat{\mathrm{S.E.}}(\bar{X}) \cdot t_{n-1}(\alpha) < \mu < \bar{X} + \widehat{\mathrm{S.E.}}(\bar{X}) \cdot t_{n-1}(\alpha)$$

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

- lze přepsat jako

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

- $\mathrm{H}_0 : \mu = \mu_0$ tedy **nezamítneme** na hladině α při oboustranné alternativě, právě když μ_0 leží v $100(1-\alpha)$% intervalu spolehlivosti
- **interval spolehlivosti obsahuje takové hodnoty μ_0, které bychom jako hypotézu nezamítli**

příklad: výšky desetiletých hochů (σ^2 neznámé)

- kritický obor: $\bar{X}$ se příliš liší od μ_0 ve směru zvolené alternativy
- spočítáme [t.test(hosi,mu=136.1,alternative="greater")]

$$T = \frac{139{,}13 - 136{,}1}{6{,}56}\sqrt{15} = 1{,}79$$

- na 5% hladině při jednostranné alternativě $\mu > \mu_0$ hypotézu zamítáme, neboť $t_{14}(0{,}10) = 1{,}76$ ($p = 4{,}7$ %)
- na 5% hladině jsme **prokázali**, že výška desetiletých vzrostla
- 95% int. spolehlivosti pro populační průměr výšek hochů: $(135{,}5; 142{,}8)$
- na 5% hladině při oboustranné alternativě hypotézu nezamítáme, neboť $t_{14}(0{,}05) = 2{,}14$ ($p = 9{,}5$ %)

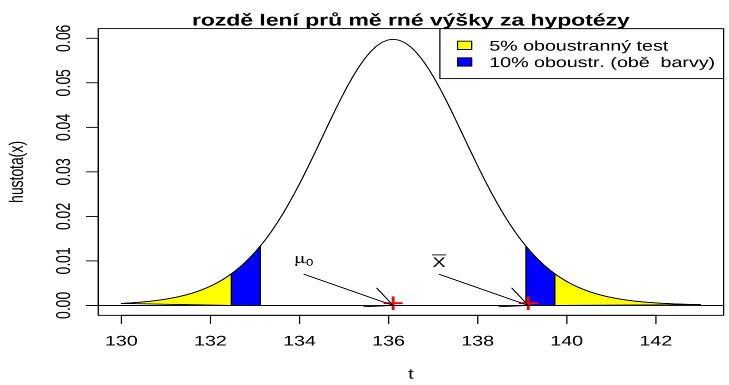
použití Excelu (Analýza dat, Popisná statistika)

přednáška	Excel	hoši
průměr	Stř. hodnota	139,13
střední chyba	Chyba stř. hodnoty	1,693
medián	Medián	139
modus	Modus	139
s	Směr. odchylka	6,56
s^2	Rozptyl výběru	42,98
špičatost	Špičatost	0,006
šikmost	Šikmost	0,090
rozpětí	Rozdíl max-min	24
minimum	Minimum	127
maximum	Maximum	151
součet	Součet	2087
rozsah výběru n	Počet	15
pol. šířka int. spol.	Hladina spol.	3,63

- $139{,}13 - 3{,}63 = 135{,}50$
- $139{,}13 + 3{,}63 = 142{,}76$
- 95% interval spolehlivosti: $(135{,}5; 142{,}8)$
- $\mu_0 = 136{,}1$ je v int. spolehlivosti
- při oboustranné alternativě jsme nezamítli H_0

kritický obor pro $\bar{X}$

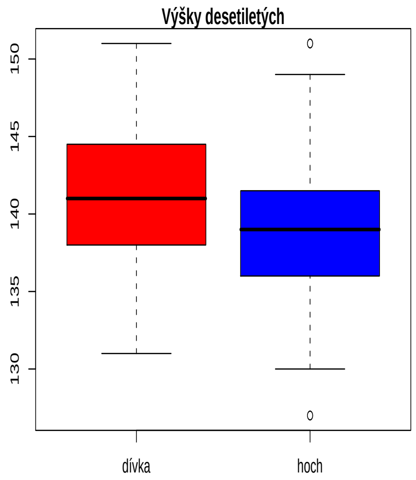
(neznámé σ)



- při jednostr. alternativě $\mu > \mu_0$ je 5% kritický obor označen oběma barvami na pravé straně

porovnání dvou populací (dvouvýběrový t-test)

- příklad: liší se desetileté dívky výškou postavy od desetiletých hochů? (tvrzení o **všech** dětech)
- výšky hochů známe, $\bar{X} = 139{,}13$ cm, $S_x = 6{,}56$, $n_x = 15$
- výšky dívek: 131, 132, 135, 141, 141, 141, 141, 142, 143, 146, 146, 151
- $\bar{Y} = 140{,}83$, $S_y = 5{,}84$, $n_y = 12$



dvouvýběrový *t*-test

- lze předpokládat, že výšky náhodně vybraných hochů mají normální rozdělení

$$X_i \sim \mathrm{N}(\mu_x, \sigma^2), \qquad \text{nezávislé}, \quad i = 1, \dots, n_x$$

- lze předpokládat, že výšky náhodně vybraných dívek mají normální rozdělení

$$Y_i \sim \mathrm{N}(\mu_y, \sigma^2), \qquad \text{nezávislé}, \quad i = 1, \dots, n_y$$

- předpoklad stejných rozptylů bývá splněn, lze jej ověřit
- musí jít o **nezávislé** náhodné výběry, nelze např. vybírat sourozenecké dvojice nebo opakovaně měřit stejnou osobu

porovnání středních hodnot nezávislých výběrů

- $H_0 : \mu_x = \mu_y$ (není rozdíl, **nulová** hypotéza)
  zřejmě totéž jako $\mu_x - \mu_y = 0$ (nulový rozdíl stř. hodnot)
  (hoši a dívky se v deseti letech co do výšky neliší)
- možné alternativy
  - $H_1 : \mu_x \neq \mu_y$ (není-li důvod k jednostranné alternativě)
  - $H_1 : \mu_x > \mu_y$ (bylo cílem dokázat, že hoši jsou větší než dívky)
  - $H_1 : \mu_x < \mu_y$ (bylo cílem dokázat, že hoši jsou menší než dívky)
- rozhodování založeno na porovnání průměrů $\bar{X}$ a $\bar{Y}$; čím více se liší „správným směrem“, tím spíše zamítnout hypotézu
- je třeba porovnat s mírou přesnosti, s jakou rozdíl průměrů $\bar{X} - \bar{Y}$ odhadne skutečný rozdíl populačních průměrů $\mu_x - \mu_y$

odhad σ^2

- k tomu je třeba odhadnout také neznámé σ^2 pomocí

$$S^2 = \frac{1}{n_x + n_y - 2}\left(\sum_{i=1}^{n_x}(X_i - \bar{X})^2 + \sum_{i=1}^{n_y}(Y_i - \bar{Y})^2\right)$$
$$= \frac{n_x - 1}{n_x + n_y - 2}S_x^2 + \frac{n_y - 1}{n_x + n_y - 2}S_y^2$$

  (vážený průměr odhadů rozptylu v obou výběrech)
- výška desetiletých dětí: $n_x = 15$, $n_y = 12$, $\bar{X} = 139{,}13$, $\bar{Y} = 140{,}83$, $S_x^2 = 42{,}98$, $S_y^2 = 33{,}79$, tudíž

$$S^2 = \frac{14}{25} \cdot 42{,}98 + \frac{11}{25} \cdot 33{,}79 = 38{,}94 = 6{,}24^2$$

kritický obor

- o hypotéze $H_0 : \mu_x = \mu_y$ se rozhoduje pomocí

$$T = \frac{\bar{X} - \bar{Y}}{\widehat{\mathrm{S.E.}}(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{S}\sqrt{\frac{n_x n_y}{n_x + n_y}}$$

- $H_1 : \mu_x \neq \mu_y$ zamítáme pokud $|T| \geq t_{n_x+n_y-2}(\alpha)$
- $H_1 : \mu_x > \mu_y$ zamítáme pokud $T \geq t_{n_x+n_y-2}(2\alpha)$
- $H_1 : \mu_x < \mu_y$ zamítáme pokud $T \leq -t_{n_x+n_y-2}(2\alpha)$
- výšky desetiletých: $T = -0{,}70 \Rightarrow$
  $|-0{,}70| < 2{,}06 = t_{15+12-2}(0{,}05)$
- na 5% hladině jsme **neprokázali** rozdíl mezi výškami desetiletých hochů a dívek ($p = 48{,}8$ %)
  [t.test(vyska~Divka,var.equal=TRUE)]
  [TTEST(A14:A28;A2:A13;2;2)]

souvislost s intervalem spolehlivosti

- $\mu_x - \mu_y = \delta$ o kolik se liší populační průměry
- odhadem pro δ je $d = \bar{X} - \bar{Y} = -1{,}7$
- krajní body intervalu spolehlivosti pro δ jsou

$$(\bar{X} - \bar{Y}) \mp \widehat{\mathrm{S.E.}}(\bar{X} - \bar{Y}) \cdot t_{n_x+n_y-2}(\alpha)$$

H$_0$ zamítáme právě tehdy, když nula **není** v int. spol. pro δ

- při porovnání výšek hochů a dívek je 95% interval pro δ

$$\left(-1{,}7 - 6{,}24\sqrt{\frac{1}{15}+\frac{1}{12}} \cdot 2{,}06\,;\; -1{,}7 + 6{,}24\sqrt{\frac{1}{15}+\frac{1}{12}} \cdot 2{,}06\right)$$

$$(-6{,}7; 3{,}3)$$

- nula **je** v intervalu, proto **nezamítáme** $\mathrm{H}_0 : \delta = 0$

shrnutí

- důležité předpoklady
  - nezávislé výběry
  - stejné (populační) rozptyly (lze testovat)
  - normální rozdělení (lze testovat)
- existuje varianta bez předpokladu stejných rozptylů
- pro velká n_x, n_y na normalitě tolik nezáleží (CLV)
- je-li problém s normalitou, lze použít jiný test (Mann-Whittney)

provedení v MS Excelu (stejné rozptyly)

přednáška	Excel	Soubor 1	Soubor 2
průměr	Stř. hodnota	139,133	140,833
rozptyl	Rozptyl	42,981	33,788
rozsah výběru	Pozorování	15	12
spol. odhad rozpt.	Společný rozptyl	38,936	
$\mathrm{H}_0 : \mu_x - \mu_y =$	Hyp. rozdíl stř. hodnot	0	
stupně vol.	Rozdíl	25	
T	t stat	-0,733	
p jednostr. testu	P(T<=t) (1)	0,244	jen někdy!
$t_{n_x+n_y-2}(2\alpha)$	t krit (1)	1,708	
p oboustr. testu	P(T<=t) (2)	0,488	
$t_{n_x+n_y-2}(\alpha)$	t krit (2)	2,060	

při oboustranné alternativě nelze nulovou hypotézu zamítnout

problém nestejných rozptylů

- předpoklad o stejném rozptylu v obou souborech nemusí být ve skutečnosti splněn, lze jej ověřit porovnáním odhadů rozptylu F-testem $F = \frac{S_x^2}{S_y^2}$
- hypotéza $\mathrm{H}_0 : \sigma_x^2 = \sigma_y^2$ se proti $\mathrm{H}_1 : \sigma_x^2 \neq \sigma_y^2$ zamítá, když je

buď $F = \frac{S_x^2}{S_y^2} \geq F_{n_x-1,n_y-1}(\alpha/2)$ nebo $\frac{1}{F} = \frac{S_y^2}{S_x^2} \geq F_{n_y-1,n_x-1}(\alpha/2)$

- vlastně se větší odhad rozptylu dělí menším odhadem, k tomu se musí zvolit odpovídající pořadí stupňů volnosti a hladina
- příklad výšky desetiletých dětí: $F = \frac{42{,}98}{33{,}79} = 1{,}27 < F_{14,11}(0{,}025) = 3{,}36$
- [var.test(vyska~Divka)]

MS Excel: Dvouvýběrový F-test pro rozptyl

přednáška	Excel	Soubor 1	Soubor 2
průměr	Stř. hodnota	139,13	140,83
rozptyl	Rozptyl	42,98	33,79
rozsah	Pozorování	15	12
stupně vol.	Rozdíl	14	11
F	F	1,27	
p	$P(F <= f)$ (1)	0,349	
	F krit (1)	2,739	

pozor Excel pracuje **špatně**: uvádí kritickou hodnotu a p-hodnotu pro jednostrannou alternativu odvozenou z hodnoty statistiky F; při oboustranné alternativě je třeba p-hodnotu vynásobit dvěma ve skutečnosti je $P(F > 1{,}27) = 0{,}349$, takže $p = 2 \cdot 0{,}349 = 0{,}698$ pro oboustrannou alternativu mělo být použito $F_{14,11}(0{,}025) = 3{,}359$

provedení v MS Excelu (nestejné rozptyly)

		Soubor 1	Soubor 2
průměr	Stř. hodnota	139,133	140,833
rozptyl	Rozptyl	42,981	33,788
rozsah	Pozorování	15	12
$H_0 : \mu_x - \mu_y =$	Hyp. rozdíl stř. hodnot	0	
stupně volnosti f	Rozdíl	25	
T	t stat	-0,713	
p jednostr. testu	P(T<=t) (1)	0,241	
$t_f(2\alpha)$	t krit (1)	1,708	
p oboustr. testu	P(T<=t) (2)	0,482	
$t_f(\alpha)$	t krit (2)	2,060	

při oboustranné alternativě nelze nulovou hypotézu zamítnout

párové testy

- není-li předpoklad **nezávislosti** porovnávaných výběrů splněn, nelze dvouvýběrový test (t-test, Wilcoxonův) použít
- typické porušení předpokladu nezávislosti je u párových dat
  - nejde o dva výběry, ale o **výběr dvojic**
  - měření na stejných objektech ve dvou různých časech
  - měření na stejných objektech před zásahem a po něm (ošetření)
  - měření na rodičích (otec a matka), sourozencích ...
- postup
  - spočítají se a hodnotí rozdíly (změny)
  - přejde se tak k úloze s jediným výběrem
  - mají-li rozdíly normální rozdělení, pak párový t-test
- v Excelu **nesmyslně** je párový t-test uveden jako dvouvýběrový párový test (nejde o **dva** výběry)

příklad: výška rodičů

- rozhodnout o tvrzení, že populační průměr výšek otců je právě o 10 cm větší než populační průměr výšek matek
- otcové: $\bar{Y} = 179{,}26$ cm, $s_y = 6{,}78$ cm, $n_y = 99$
  matky: $\bar{Z} = 166{,}97$ cm, $s_z = 6{,}11$ cm, $n_z = 99$
- otcové jsou (ve výběru) v průměru o $\bar{Y} - \bar{Z} = 12{,}29$ cm vyšší
- směrodatná odchylka **rozdílů** výšek je 8,14 (méně, než kdyby byly výšky rodičů nezávislé ... $6{,}78^2 + 6{,}11^2 = 9{,}13^2$)
- **střední chyba** rozdílu průměrů je $8{,}14/\sqrt{99} = 0{,}819$
- rozhodneme podle statistiky [t.test(vyska.o-vyska.m,mu=10)]
  [t.test(vyska.o,vyska.m,paired=TRUE,mu=10)]

$$T = \left|\frac{12{,}29 - 10}{0{,}819}\right| = 2{,}801 > 1{,}984 = t_{98}(0{,}05) \qquad p = 0{,}6\ \%$$

dvouvýběrový Wilcoxonův test (někdy nepřesně Mannův-Whitneyův)

pořadová obdoba dvouvýběrového *t*-testu

- porovnáváme stejný kvantitativní znak ve dvou populacích
- máme dva **nezávislé** výběry z těchto populací
- není třeba předpokládat normální rozdělení, stačí spojité
- nechť $X_1, \dots, X_{n_x}$ a $Y_1, \dots, Y_{n_y}$ jsou **nezávislé** výběry ze spojitého rozdělení (například věk matek, střední délka života mužů při narození ve dvou skupinách zemí, potratovost ...)
- H_0 tvrdí, že obě rozdělení jsou stejná (mezi populacemi není rozdíl, zpravidla nás zajímá, že není rozdíl v mírách polohy)
- specielně to znamená, že **populační mediány** jsou shodné
- postup založen na pořadí bez ohledu na výběr
- idea: kdyby nebyl mezi populacemi rozdíl, byla by takto zjištěná průměrná pořadí v obou výběrech podobná

příklad: potraty na 1000 obyv. (Čechy vers. Morava)

počet potratů na 1000 žen fertilního věku v roce 2003

kraj	Pha	Stč	Jč	Pl	KV	Ús	Lb
potratovost	4,03	4,02	4,11	4,70	5,65	5,80	4,98
pořadí	7	6	8	10	12	13	11
kraj	HK	Par	Vys	JM	Ol	Zl	MS
potratovost	4,33	3,38	3,57	3,70	3,65	3,42	3,87
pořadí	9	1		4	3	2	5

- H_0 : shoda populací (zejm. mediánů), H_1 : neshoda
- nejasné, kam patří kraj Vysočina; vynecháme jej
- průměrné pořadí českých krajů: 77/9=8,56
  $W_x=7+6+8+10+12+13+11+9+1=77$
- průměrné pořadí moravských krajů: 14/4=3,5
  $W_y=4+3+2+5=14$

přibližné rozhodování (n_x, n_y desítky)

- W_x, W_y součty pořadí, W_x standardizujeme

$$Z = \frac{W_x - n_x(n_x + n_y + 1)/2}{\sqrt{n_x n_y (n_x + n_y + 1)/12}}$$

- za nulové hypotézy (není rozdíl mezi populacemi) je použitím centrální limitní věty $Z \dot{\sim} \mathrm{N}(0,1)$
- nulovou hypotézu zamítáme, je-li $|Z| \geq z(\alpha/2)$
- náš příklad: [wilcox.test(potr~Cechy)]

$$Z = \left| \frac{77 - 9 \cdot 14/2}{\sqrt{9 \cdot 4 \cdot 14/12}} \right| = 2{,}16 > 1{,}96 = z(0{,}05/2) \qquad p = 3{,}1\ \%$$

- na 5% hladině jsme prokázali rozdíl

Mannův-Whitneyův test

jiná myšlenka, ale stejný kritický obor jako u dvouvýběrového Wilcoxonova testu

- vyšetříme všechny dvojice X_i, Y_j (je jich $n_x \cdot n_y$)
- za platnosti nulové hypotézy (mezi populacemi není rozdíl):
- asi v polovině případů by mělo být $X_i > Y_j$, v polovině případů obráceně
- zjistíme, kolikrát je opravdu $X_i > Y_j$ (skutečné hodnoty)
- přičteme polovinu případů, kdy je $X_i = Y_j$
- výsledný součet by měla být podobný hodnotě $n_x n_y/2$
- pro zajímavost (R značí tento součet jako W):

$$W = n_x n_y + n_y(n_y + 1)/2 - W_y$$

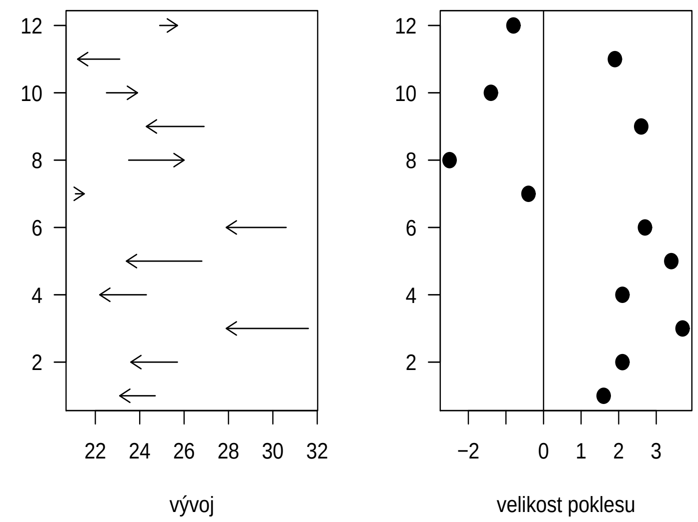
příklad: klesá potratovost? (párový t-test zde nevhodný)

potratů na 100 těhotenství, určitě problém s normalitou

Y_i	Z_i	X_i	R_i^+
24,7	23,1	1,6	4
25,7	23,6	2,1	6
31,6	27,9	3,7	12
24,3	22,2	2,1	7
26,8	23,4	3,4	11
30,6	27,9	2,7	10
21,1	21,5	-0,4	1
23,5	26,0	-2,5	8
26,9	24,3	2,6	9
22,5	23,9	-1,4	3
23,1	21,2	1,9	5
24,9	25,7	-0,8	2

- použijeme údaje z 12 okresů v letech 2000 (Y_i) a 2001 (Z_i)
- hypotéza H_0 : v obou letech potratovost stejná, rozdíly dány náhodným kolísáním; H_1 : potratovost klesá (jednostranná alt.)
- za H_0 by rozdíly měly kolísat **symetricky kolem nuly**
- za H_1 by měly převládat kladné rozdíly, spíše velké
- průměrné pořadí z 8 kladných rozdílů: 8 (součet $W = 64$), průměrné pořadí ze 4 záporných rozdílů 3,5 (součet 14)

příklad: klesá potratovost?



párový Wilcoxonův (Wilcoxon signed rank) test

- nechť $(Y_1, Z_1) \dots, (Y_n, Z_n)$ **nezávislé dvojice**, rozdíly $X_i = Y_i - Z_i$ mají **spojité** rozdělení
- H_0 : Y_i, Z_i mají stejné rozdělení (populace jsou stejné)
- představa: mají-li Y_i, Z_i stejné rozdělení, pak rozdíly $X_i = Y_i - Z_i$ jsou symetricky rozděleny kolem nuly
- postup
  - vyloučit nulové hodnoty X_i (tedy shodné hodnoty Y_i, Z_i), podle toho případně zmenšit n
  - určit pořadí R_i^+ **absolutních hodnot** $|X_i| = |Y_i - Z_i|$
  - určit W, tj. součet pořadí původně kladných hodnot X_i
  - podle W rozhodnout

rozhodování

- na základě centrální limitní věty lze použít

$$Z = \frac{W - \mathrm{E}\,W}{\mathrm{S.E.}(W)} = \frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

- hypotézu o shodě zamítneme, bude-li $|Z| \geq z(\alpha/2)$
- při jednostranné alternativě porovnat Z a $z(\alpha)$
- pro malý počet dvojic (do deseti) raději použít tabulky
- příklad ($W = 64, n = 12$, jinak přesně je $p = 2{,}6$ %)

$$Z = \frac{64 - 12 \cdot 13/4}{\sqrt{12 \cdot 13 \cdot 25/24}} = 1{,}961 > 1{,}645 = z(0{,}05), p = 2{,}5\ \%$$

poznámky k výpočtu

- nezapomenout vyloučit nulové rozdíly
- shodným absolutním hodnotám rozdílům přiřadíme jejich průměrné pořadí
- Excel nám v takovém případě moc nepomůže, protože řeší problém shod nestandardně, např.:

X_i	4	-2	5	2	-6	-4	2	7
$\|X_i\|$	4	2	5	2	6	4	2	7
R_i^+	4,5	2	6	2	7	4,5	2	8
Excel	4	1	6	1	7	4	1	8

- v tabulce patrné nestandardní chování Excelu
- [wilcox.test(pokles,alternative="greater")]

párový znaménkový (sign) test

- hodnotí pouze **počet** kladných a záporných rozdílů, nezáleží na tom, jak jsou rozdíly veliké (slabší test než Wilcoxonův)
- H_0 : Y_i, Z_i mají stejné rozdělení; za hypotézy očekáváme, že počty kladných a záporných X_i jsou podobné
- označme Y počet kladných X_i z celkem n nenulových, za hypotézy $Y \sim \text{bi}(n, 1/2)$
- přibližné rozhodování (centrální limitní věta)

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

- při jednostranné alternativě porovnáme Z a $z(\alpha)$

poznámky

- pro znaménkový test není třeba znát hodnoty Y_i, Z_i, stačí vědět, která z možností $Y_i > Z_i$, $Y_i < Z_i$, $Y_i = Z_i$ nastala
- náš příklad o možném poklesu potratovosti ($n = 12$, $Y = 8$)

$$Z = \frac{2 \cdot 8 - 12}{\sqrt{12}} = 1{,}155, \quad p = \text{P}(Z > 1{,}155) = 0{,}124$$

- při malých hodnotách n (do 30) se doporučuje Yatesova korekce

$$Z_{\text{Yates}} = \frac{|2Y - n| - 1}{\sqrt{n}} \text{sign}(2Y - n)$$

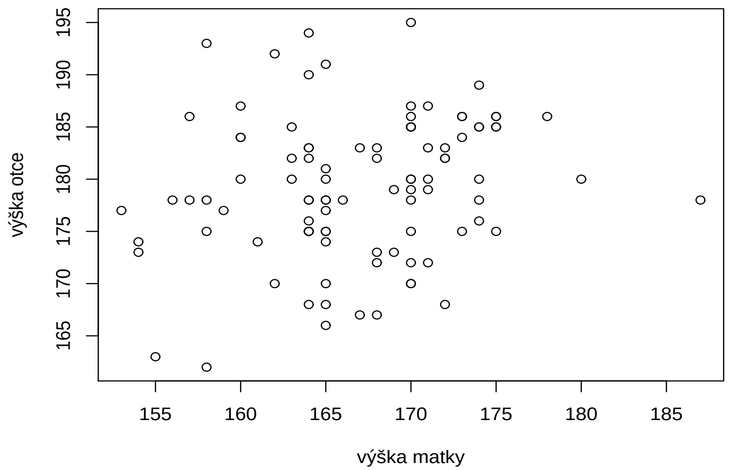
- náš příklad (Yatesova korekce, jiným způsobem přesně p =0,194)

$$Z = \frac{|2 \cdot 8 - 12| - 1}{\sqrt{12}} \cdot 1 = 0{,}866, \quad p = 1 - \Phi(0{,}866) = 0{,}193$$

přehled testů o populačních mírách polohy

rozdělení	normální	jen spojité
populační parametr (o čem je hypotéza)	populační průměr	populační medián
jeden výběr	jednovýběrový t-test	znaménkový nebo Wilcoxonův
výběr dvojic	párový t-test	znaménkový nebo Wilcoxonův
dva nezávislé výběry	dvouvýběrový t-test	Mann-Whitney (dvouvýb. Wilcoxonův)

souvisí spolu výšky rodičů?



prokazování závislosti spojitých veličin

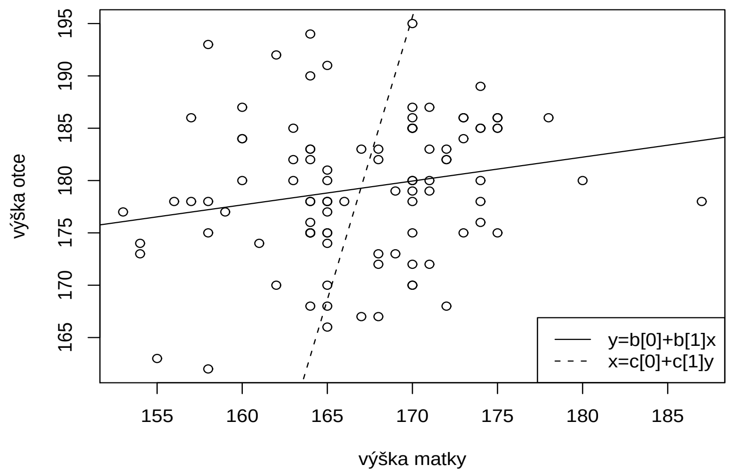
- víme, že pro nezávislé X, Y je $\rho_{X,Y} = 0$
- r_{xy} je odhadem $\rho_{X,Y}$; jak daleko od nuly musí být r_{xy}, abychom na hladině α prokázali závislost X, Y?
- za předpokladu, že X, Y mají normální rozdělení (nebo počet pozorovaných dvojic X_i, Y_i je velký) a **dvojice** (X_i, Y_i) jsou mezi sebou (pro různá i) **nezávislé**, hypotézu nezávislosti zamítáme pokud je $|T| \geq t_{n-2}(\alpha)$, kde

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

příklad: výšky rodičů

- pro $n = 99$ dvojic byl spočítán korelační koeficient $r = 0{,}205$;
- $$T = \frac{0{,}205}{\sqrt{1-0{,}205^2}}\sqrt{97} = 2{,}07 > t_{97}(0{,}05) = 1{,}98$$
- na 5% hladině jsme závislost prokázali
- $t_{97}(0{,}01) = 2{,}63$, tudíž na 1% hladině jsme závislost neprokázali
- výška zpravidla splňuje předpoklad o normálním rozdělení
- [cor.test(∼ vyska.m+vyska.o, data=Kojeni)]
  [CORREL(x;y)] (pouze výpočet korelačního koeficientu)
- není-li normální rozdělení a nemnoho pozorování, raději použít Spearmanův korelační koeficient

příklad: výšky rodičů



Spearmanův korelační koeficient

- místo původních hodnot x_i, y_i používá jejich pořadí R_i, Q_i
- je to vlastně Pearsonův korelační koeficient použitý na pořadí
- výpočet lze upravit, zjednodušit na

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

- vhodný pro nelineární monotonní **závislost**, nevadí odlehlé hodnoty
- při testování nemusí být normální rozdělení
- nezávislost se zamítá, je-li $|r_S| \geq z(\alpha/2)$ (pro n velké), jinak s využitím tabulek

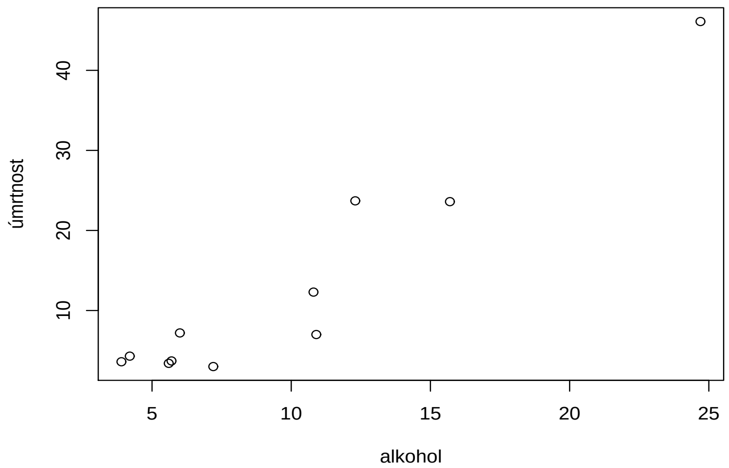
příklad: alkohol a úmrtnost na cirhózu

země	spotřeba	úmrtnost	R_i	Q_i	$R_i - Q_i$
Finsko	3,9	3,6	1	3	-2
Norsko	4,2	4,3	2	5	-3
Irsko	5,6	3,4	3	2	1
Holandsko	5,7	3,7	4	4	0
Švédsko	6,0	7,2	5	7	-2
Anglie	7,2	3,0	6	1	5
Belgie	10,8	12,3	7	8	-1
Rakousko	10,9	7,0	8	6	2
SRN	12,3	23,7	9	10	-1
Itálie	15,7	23,6	10	9	1
Francie	24,7	46,1	11	11	0

$$r_S = 1 - \frac{6}{11 \cdot 120} (2^2 + 3^2 + \ldots) = 0{,}773$$

$r = 0{,}956$ zdánlivě mnohem těsnější závislost!

cirhóza jater a spotřeba alkoholu

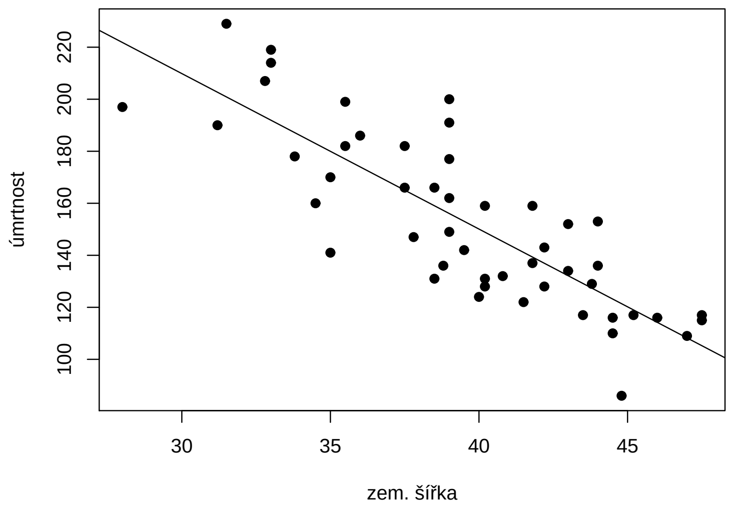


Regrese

- na rozdíl od korelace (síla závislosti) hledáme tvar (způsob) závislosti, zajímá nás také průkaznost závislosti
- snažíme se z daných hodnot **regresorů (nezávisle proměnných, prediktorů)** předpovědět hodnoty **závisle proměnné** (odezvy, vysvětlované proměnné)
- snažíme se variabilitu (kolísání hodnot) odezvy vysvětlit závislostí na kolísajících regresorech
- prvně v tomto smyslu F. Galton (1886) při vyšetřování závislosti výšky potomků na průměrné výšce rodičů
- Pearson, Lee (1903): potomci otců o dva palce vyšších než průměr všech otců byli v průměru jen o palec vyšší než průměr synů; dvoupalcová odchylka se nereprodukovala celá, byl patrný návrat (**regres**) k průměru

příklad: souvisí úmrtnost se zeměpisnou šířkou?

úmrtnost na melanom na 10 000 000 obyvatel v státech USA



regresní přímka

- cíl: chování Y (úmrtnost, mortality) co nejlépe (nejvíce) vysvětlit lineární závislostí na x (zeměpisná šířka, latitude)
- (naše představa, předpoklad:) každé zem. šířce x_i odpovídá jakási střední úmrtnost $\mathrm{E}\,Y_i$, ta závisí na zeměpisné šířce lineárně

$$\mathrm{E}\,Y_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n$$

- obecně předpokládáme, že $Y_1, \ldots, Y_n$ jsou **nezávislé** a

$$Y_i \sim \mathrm{N}\big(\beta_0 + \beta_1 x_i, \sigma^2\big), \qquad i = 1, \ldots, n$$

- parametry β_0, β_1 odhadneme **metodou nejmenších čtverců** minimalizací přes β_0, β_1 součtu čtverců „svislých“ odchylek

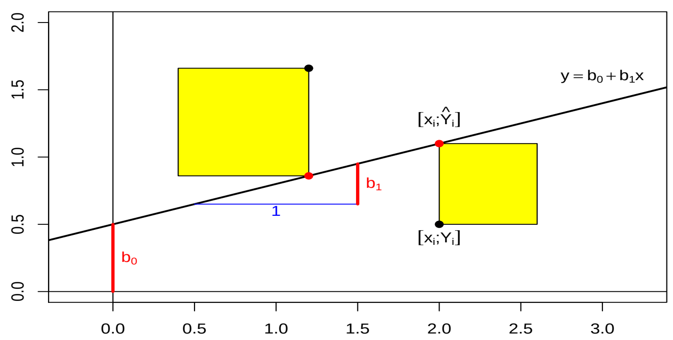
$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- výsledné minimum (pro $\beta_0 = b_0, \beta_1 = b_1$) se nazývá **reziduální součet čtverců**: $S_e = \sum_{i=1}^n (Y_i - b_0 - b_1 x_i)^2$

metoda nejmenších čtverců

na obrázku jsou jen dvě z mnoha pozotování!

odhadovaná závislost:	$y = \beta_0 + \beta_1 \cdot x$	(populace)
odhad závislosti:	$y = b_0 + b_1 \cdot x$	(výběr)
celková plocha čtverců:	$S_e = \sum_{i=1}^n (Y_i - b_0 - b_1 x_i)^2$	(výběr)



náš příklad

[summary(lm(mortality~latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	389,19	23,81	16,34	<0,001
latitude	– 5,98	0,60	– 9,99	<0,001

- odhad závislosti: $\widehat{\text{mortality}}$ = 389,19 – 5,98 latitude
- s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 6 osob na 10 000 000 obyvatel
- na rovníku by úmrtnost měla být 389 jednotek, ale je to extrapolace mimo rozmezí známých hodnot – sotva použitelné
- závislost je průkazná, neboť v řádku pro x (latitude) je p <0,001

obecně

- odhadovaná závislost $y = \beta_0 + \beta_1 x$, odhadnutá $y = b_0 + b_1 x$
- závislost na x prokazujeme testováním hypotézy $\mathrm{H}_0 : \beta_1 = 0$ (pak je y pro všechna x stejné, tedy $y = \beta_0$) pomocí

$$T = \frac{b_1}{\text{S.E.}(b_1)} = \frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- zamítáme H_0 proti oboustr. alternativě, když $|T| \geq t_{n-2}(\alpha)$
- **reziduální součet čtverců – nevysvětlená variabilita** Y
  $S_e = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 x_i))^2$ reziduální součet čtverců
  $s^2 = S_e/(n-2)$ reziduální rozptyl
- **koeficient determinace** ukazuje, jaký **díl variability odezvy** (tj. jaký díl $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$) jsme závislostí vysvětlili

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

náš příklad a tabulka analýzy rozptylu

[anova(lm(mortality~latitude))]

variabilita	st. vol. f	součet čtverců SS	prům. čtverec MS	F	p
model	1	36 464,20	36 464,20	99,797	<0,001
reziduální	47	17 173,07	365,38		
celkem	48	53 637,27			

- kolísání úmrtnosti vysvětlíme závislostí z 68 %, neboť je

$$R^2 = 1 - \frac{17173{,}07}{53637{,}27} = \frac{36464{,}20}{53637{,}27} = 0{,}680$$

interpretace

- předpověď: $\widehat{\text{úmrtnost}}$=389,19-5,98$\cdot$ šířka
- na 30. stupni očekáváme úmrtnost: 389,19 – 5,98 $\cdot$ 30 = 209,9
- na 40. stupni očekáváme úmrtnost: 389,19 – 5,98 $\cdot$ 40 = 150,1
- přechod z 30. stupně na 40. stupeň znamená **v průměru** pokles o 10 $\cdot$ 5,98 = 59,8 úmrtí na 10 000 000 obyvatel
- pokusíme se predikci zlepšit přidáním další nezávisle proměnné

dva regresory $y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot v$

(test přidané informace)

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	401,17	28,04	14,31	<0,001
latitude	– 5,93	0,60	– 9,82	<0,001
longitude	0,15	0,19	0,82	0,418

- pokusíme se přidat zeměpisnou délku
- není průkazné, že by koeficient u longitude byl nenulový (nezamítneme hypotézu, že koeficient je nulový)
- longitude nepřináší další informaci o mortality, kterou bychom už neměli ze známé hodnoty latitude
- $\Rightarrow$ není vhodné přidávat do modelu s latitude také longitude
- koeficient determinace $R^2 = 0{,}684$ (původně 0,680) se téměř nezměnil

podrobnější rozbor – vliv oceánu

- závislost jen pro vnitrozemské státy ($R^2 = 59{,}6$ %):
  [lm(mortality∼latitude,subset=Ocean==0)]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,55	36,70	9,82	<0,001
latitude	– 5,485	0,904	– 6,07	<0,001

- závislost jen pro přímořské státy ($R^2 = 78{,}6$ %):
  [lm(mortality∼latitude,subset=Ocean==1)]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	381,20	24,83	15,35	<0,001
latitude	– 5,491	0,640	– 8,58	<0,001

- směrnice jsou téměř stejné, abs. členy rozdílné
- v obou případech s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 5,5 osob na 10 000 000 obyvatel

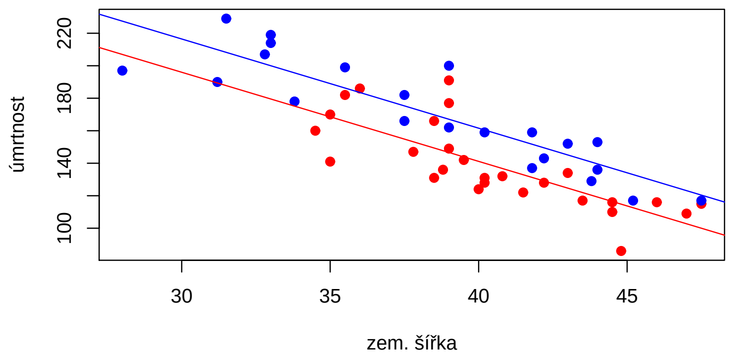
společně vnitrozemské i přímořské státy

[summary(lm(mortality∼Ocean+latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,69	21,50	16,78	<0,001
Ocean	20,43	4,83	4,23	<0,001
latitude	– 5,49	0,53	– 10,44	<0,001

- koeficient determinace $R^2 = 0{,}770$
- při „stěhování“ z vnitrozemí k oceánu po rovnoběžce roste úmrtnost v průměru o 20 osob na 10 milionů obyvatel
- je to ekvivalentní vnitrozemskému stěhování o $20{,}43/5{,}49 = 3{,}72$ stupňů na jih
- na každý stupeň stěhování na sever klesá úmrtnost o 5,5, pokud se nezmění vztah k oceánu

příklad: souvisí úmrtnost se zeměpisnou polohou?



- vnitrozemské státy: y=360,69–5,49 x
  přímořské státy: y=(360,69+20,43)–5,49 x =381,12–5,49 x
- lze ověřit, že přímky mohou být rovnoběžné ($p = 99{,}6$ %)

pozor na interpretaci odhadů (na dalším příkladu)

- závisí procento tuku dospělého muže na jeho výšce? pokud ano, tak s výškou roste nebo klesá?
- závisí na tom, jak se na úlohu díváme, co bereme v úvahu
- $\widehat{\text{fat}} = -\,47{,}68 + 0{,}341\ \text{height}$ $R^2 = 11{,}8$ %
- $\widehat{\text{fat}} = 16{,}55 - 0{,}244\ \text{height} + 0{,}504\ \text{weight}$ $R^2 = 71{,}4$ %
- ve všech případech jsou koeficienty u regresorů na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky (a **nezměněné hmotnosti** pro druhý model)

regrese v MS Excelu 2000, 2003

	Excel 2003	označení
absolutní člen	Hranice	b_0
odhad	Koeficienty	b_i
střední chyba odhadu*	Chyba střední hodnoty	S.E.(b_j)
koeficient (mnohonásobné) korelace	Násobné R	$\sqrt{R^2}$
koeficient determinace	Hodnota spolehlivosti R	R^2
adjustovaný koef. det.	Nastavená hodnota spol. R	R^2_{adj}
resid. směr. odchylka*	Chyba stř. hodnoty	s
počet pozorování	Pozorování	n
počet st. volnosti	Rozdíl	

* pozor, dva **různé** pojmy označeny stejně!

regrese v MS Excelu 2000, 2003

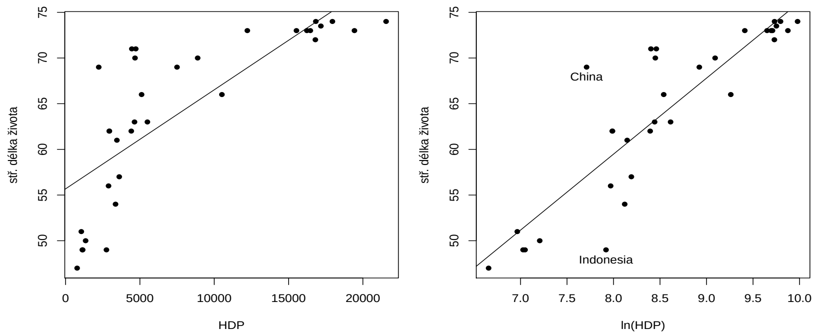
- Pozor na nabízený graf „Graf s rozdělením pravděpodobnosti“: obecně **nevypovídá** o normálním rozdělení, jak bylo asi přání tvůrců programu, bylo by třeba použít místo vysvětlované veličiny některá z reziduí
- Nabízená „Normovaná rezidua“ jsou v regresi zcela nestandardní (z-skóry běžných reziduí)

ověření předpokladů regrese

- reziduum = pozorovaná ∼ vyhlazená hodnota
  [a <− lm(...∼...)] [resid(a)] [rstandard(a)]
- tvar závislosti: rezidua ∼ vyhlazené hodnoty
  [plot(resid(a)∼fitted(a))]
- normální rozdělení: uspořádaná rezidua ∼ očekávané hodnoty
  nikoliv závisle proměnná ∼ něco! (Excel)
  test normality na rezidua [shapiro.test(rstandard(a))]
- stabilita rozptylu:
  odmocnina abs. hodnoty reziduí ∼ vyhlazené hodnoty
  [plot(sqrt(abs(resid(a)))∼fitted(a))]
- v R připraveno [plot(a)]

praktické problémy: transformace

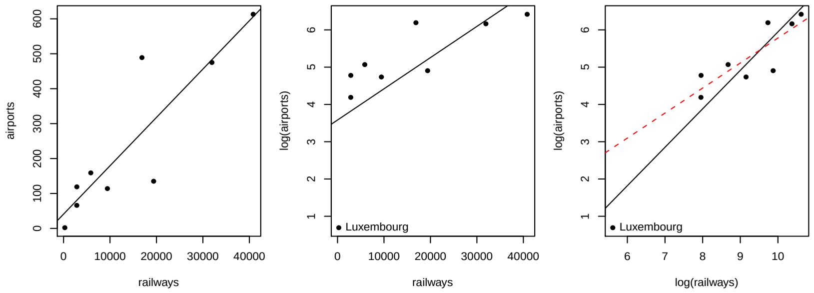
střední délka života ∼ HDP (rok 1992, 33 skupin zemí z celého světa)



- v původním měřítku závislost nelineární
- logaritmování HDP hodně pomohlo, ale ještě jistě jiné vlivy
- log(HDP) vysvětlí téměř 79 % variability střední délky života
- lze identifikovat státy, které se zvlášť vymykají

praktické problémy: zdánlivá závislost

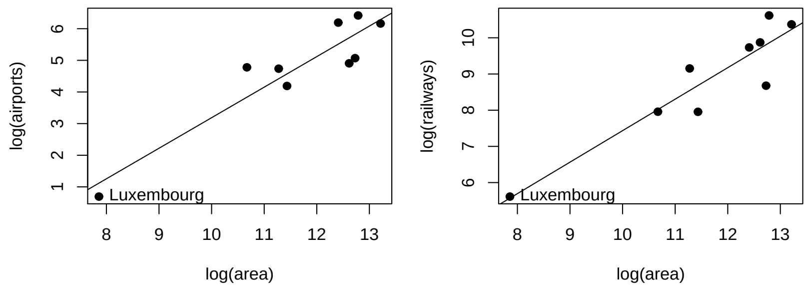
počet letišť ∼ délka železnic v Evropě



- v původním měřítku: $R^2 = 78$ %, $p = 0{,}2$ %
- v logaritmickém měřítku x i y: $R^2 = 87$ %, $p = 0{,}02$ %
- logaritmické měřítko, bez Lucemburska: $R^2 = 69$ %, $p = 1$ %

praktické problémy: zdánlivá závislost

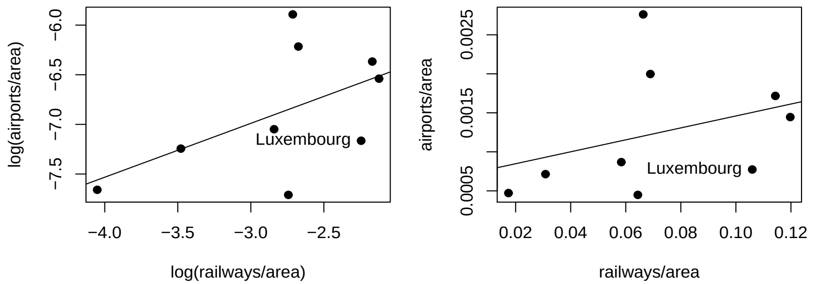
počet letišť resp. délka železnic ∼ velikost země v Evropě



- počet letišť i délka železnic souvisí s velikostí země
- u letišť: $R^2 = 86$ %, $p = 0{,}03$ %
- u železnic: $R^2 = 85$ %, $p = 0{,}04$ %

praktické problémy: zdánlivá závislost

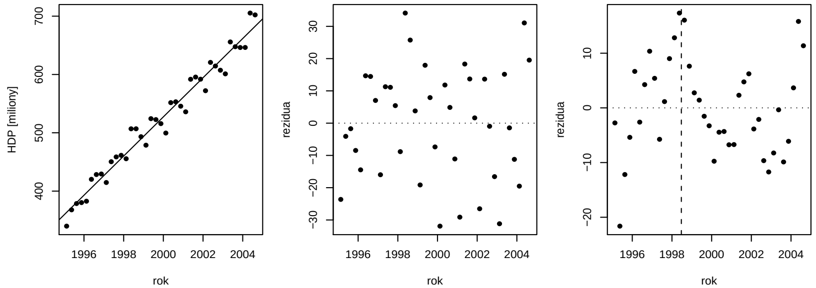
počet letišť a délka železnic ∼ plocha, obojí vztaženo k ploše



- závislost v logaritmech: $R^2 = 28$ %, $p = 14$ %
- závislost v původním měřítku: $R^2 = 12$ %, $p = 36$ %
- relativní počet letišť nesouvisí s relativní délkou železnic

praktické problémy: časová řada

vývoj HDP v ČR – pozorování tvoří časovou řadu



- po sobě jdoucí pozorování nejsou nezávislá
- zdánlivě dobře rozmístěná rezidua po vyrovnání přímkou, periodičnost překryta kvartálním kolísáním (uprostřed)
- na obr. vpravo (model s přihlédnutím ke kvartálům) je patrná závislost po sobě jdoucích reziduí, tedy i hodnot
- na pravém grafu vyznačen okamžik parlamentních voleb 1998

motivační příklad: je výběr reprezentativní?

- bylo provedeno šetření mezi ženami ve věku 18 až 50 roků
- mezi 498 náhodně oslovenými ženami bylo celkem 180 žen svobodných, 239 žen vdaných, 75 žen rozvedených a 4 ovdovělé
- stejné údaje v procentech: 36,14 % svobodných, 47,99 % vdaných, 15,06 % rozvedených, 0,80 % ovdovělých
- je známo, že v celé populaci žen v ČR uvedeného věkového rozpětí je 34,27 % svobodných, 52,03 % vdaných, 12,50 % rozvedených a 1,20 % ovdovělých
- lze výběr považovat za reprezentativní co do stavu?
- odpovídají procenta výběru procentům populace, tj. je výběr **reprezentativní**?
- dostali bychom reprezentativní výběr, kdybychom hledali ženy např. v porodnici?

multinomické rozdělení

- zobecnění binomického rozdělení na k-tici náhodných veličin $X_1, \ldots, X_k$
- parametry $n, \pi_1, \ldots, \pi_k$ $(0 < \pi_j < 1, \quad \pi_1 + \ldots + \pi_k = 1)$
- n **nezávislých** pokusů
- v každém pokusu **právě jeden** z k možných výsledků
  - možné výsledky se musí vylučovat
  - aspoň jeden z možných výsledků musí nastat
- j-tý výsledek nastává s pravděpodobností π_j
- X_j – počet pokusů, v nichž nastal j-tý možný výsledek, tedy nutně

$$X_1 + \ldots + X_k = n$$

příklady multinomického rozdělení

- předvolební průzkum
  - n – počet tázaných
  - π_j – skutečný podíl voličů j-té strany v populaci
  - X_j – počet (četnost) voličů j-té strany ve výběru
- hody hrací kostkou
  - n – počet hodů
  - $\pi_1, \ldots, \pi_6$ – pravděpodobnosti jednotlivých stran kostky
  - $X_1, \ldots, X_6$ – absolutní četnosti jednotlivých stran kostky
- krevní skupiny
  - n=4 (skupiny 0, A, B, AB)
  - $\pi_0, \pi_A, \pi_B, \pi_{AB}$ – pravděpodobnosti skupin 0, A, B, AB
  - X_0, X_A, X_B, X_{AB} – počty osob se skupinami 0, A, B, AB
- půjde o multinomické rozdělení, když pořídíme vzorek vědců (populaci vědců lze definovat), pokud je budeme třídit podle státní příslušnosti?

vlastnosti multinomického rozdělení

- každá jednotlivá složka X_j má binomické rozdělení:

$$X_j \sim \mathrm{bi}(n, \pi_j)$$

- střední hodnota: $\mu_{X_j} = n\pi_j$, rozptyl: $\sigma^2_{X_j} = n\pi_j(1 - \pi_j)$
- (pro zajímavost) kovariance: $\mathrm{cov}(X_j, X_t) = -n\pi_j\pi_t \quad j \neq t$
- náhodné veličiny $X_1, \ldots, X_k$ jsou závislé $(X_1 + \ldots + X_k = n)$
- asymptotická vlastnost **chí-kvadrát** (velká n, $n\pi_j \geq 5\ \forall j$)

$$\boxed{\chi^2 = \sum_{j=1}^{k} \frac{(X_j - n\pi_j)^2}{n\pi_j} \sim \chi^2_{k-1}}$$

- X_j – **empirické četnosti**,
  $n\pi_j$ – **očekávané (teoretické) četnosti**

příklad: hrací kostka A

- test **jednoduché** hypotézy
- $n = 100$ hodů kostkou
- $X_1 = 12,\ X_2 = 21,\ X_3 = 14,\ X_4 = 15,\ X_5 = 21,\ X_6 = 17$
- hypotéza $H_0 : \pi_1 = \ldots = \pi_6 = 1/6$ dá očekávané četnosti $n\pi_1 = \ldots = n\pi_6 = 100/6 = 16{,}67$ (vždy více než 5))
- $$\chi^2 = \frac{(12-16{,}67)^2}{16{,}67} + \ldots + \frac{(17-16{,}67)^2}{16{,}67} = 4{,}16$$
- $$\chi^2 < \chi^2_5(0{,}05) = 11{,}07, \qquad p = 52{,}7\ \%$$
- neprokázali jsme, že by kostka nebyla symetrická
- [chisq.test(c(12,21,14,15,21,17),p=rep(1,6)/6)]

příklad: hrací kostka B (1)

- $n = 100$ hodů kostkou
- $X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15, X_6 = 41$
- hypotéza $H_0 : \pi_1 = \ldots = \pi_6 = 1/6$ dá očekávané četnosti $n\pi_1 = \ldots = n\pi_6 = 100/6 = 16{,}67$
$$\chi^2 = \frac{(15-16{,}67)^2}{16{,}67} + \ldots + \frac{(41-16{,}67)^2}{16{,}67} = 48{,}32$$
- $$\chi^2 > \chi^2_5(0{,}05) = 11{,}07 \qquad p < 0{,}0001$$
- zřejmě je nutno zamítnout hypotézu, že kostka je symetrická
- na 5% hladině jsme prokázali, že není symetrická

příklad: hrací kostka B (2), jiná H_0

- $n = 100$ hodů kostkou
- $X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15, X_6 = 41$
- nulová hypotéza: $\pi_1 = \ldots = \pi_5 = 1/10, \pi_6 = 5/10 = 1/2$
- očekávané četnosti za hypotézy: $n\pi_1 = \ldots = n\pi_5 = 100/10 = 10, n\pi_6 = 100/2 = 50$
$$\chi^2 = \frac{(15-10)^2}{10} + \ldots + \frac{(15-10)^2}{10} + \frac{(41-50)^2}{50} = 12{,}72$$
- $$\chi^2 > \chi^2_5(0{,}05) = 11{,}07 \qquad p = 2{,}6\ \%$$
- zřejmě je nutno zamítnout i tuto hypotézu [chisq.test(c(15,16,7,6,15,41),p=c(1,1,1,1,1,5)/10)]

příklad: hrací kostka B (3) (použít jen část informace)

- $n = 100$ hodů kostkou
- $X_6 = 41$
- nulová hypotéza: $\pi_6 = 5/10 = 1/2$
- hypotéza o psti jediného z možných výsledků (pst šestky) – binomické rozdělení
- dříve jsme určili přibližný 95% interval spolehlivosti pro pravděpodobnost šestky: (0,31; 0,51)
- 1/2 je v tomto intervalu, na 5% hladine **nelze** zamítnout [binom.test(41,100)]

příklad: je výběr reprezentativní?

- provedeme test hypotézy, že pravděpodobnosti čtyř skupin žen jsou rovny procentům v populaci

	svobodné	vdané	rozvedené	ovdovělé	celkem
populace	34,27 %	52,03 %	12,50 %	1,20 %	100 %
výběr	180	239	75	4	498
výběr (rel.)	36,14 %	47,99 %	15,06 %	0,80 %	100 %
oček. čet.	170,69	259,07	62,26	5,99	498
přínos	0,51	1,55	2,61	0,66	5,33

$$\frac{(180-170{,}69)^2}{170{,}69}+\frac{(239-259{,}07)^2}{259{,}07}+\frac{(75-62{,}26)^2}{62{,}26}+\frac{(4-5{,}99)^2}{5{,}99}$$

- výsledná hodnota chí-kvadrát je $\chi^2 = 5{,}34$ ($p = 14{,}9$ %), ale $\chi^2_3(0{,}05) = 7{,}81$
  [chisq.test(c(180,239,75,4),p=c(34.27,52.03,12.50,1.20)/100)]
- neprokázali jsme, že by výběr nebyl reprezentativní, můžeme jej za reprezentativní považovat

test homogenity r výběrů

- například, zda mají kostky A, B stejné šestice psti (ať už je ta šestice pstí jakákoliv)
- $X_{i1}, \dots, X_{ik}$ i-tý výběr z multinomického rozdělení s parametry $n_{i\bullet}, \pi_{i1}, \dots, \pi_{ik}$ $\qquad (i = 1, \dots, r)$
- H_0 : pravděpodobnosti jsou ve všech srovnávaných populacích stejné: $\pi_{i1} = \pi_1, \dots, \pi_{ik} = \pi_k$ (nezávisí na populaci)
- četnosti uspořádáme do kontingenční tabulky
  - n_{ij} – počet j-tých výsledků v i-tém výběru
  - $n_{i\bullet} = \sum_j n_{ij}$ jsou řádkové marginální četnosti (rozsahy výběrů)
  - $n_{\bullet j} = \sum_i n_{ij}$ jsou sloupcové marginální četnosti (četnosti možných výsledků bez ohledu na výběr)
  - $n = \sum_i n_{i\bullet} = \sum_j n_{\bullet j} = \sum_i \sum_j n_{ij}$ je celkový počet pozorování

test homogenity r výběrů

- neznámé pravděpodobnosti π_j odhadneme pomocí sloupcových marginálních relativních četností $n_{\bullet j}/n$
- očekávané četnosti tak budou $o_{ij} = n_{i\bullet}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}$
- empirické četnosti porovnáme s četnostmi očekávanými

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(n_{ij}-o_{ij})^2}{o_{ij}}$$

- χ^2 je mírou neshody skutečných a očekávaných četností
- platí-li hypotéza, má výsledná statistika χ^2-rozdělení s $(r-1)(k-1)$ stupni volnosti
- hypotézu o shodě pravděpodobností v r populacích zamítáme, je-li $\chi^2 \geq \chi^2_{(r-1)(k-1)}(\alpha)$
- je třeba, aby očekávané četnosti byly dost velké, aspoň 5

mají obě kostky stejné šestice pravděpodobností?

- empirické četnosti (kontingenční tabulka)

A	12	21	14	15	21	17	100
B	15	16	7	6	15	41	100
	27	37	21	21	36	58	200

- očekávané četnosti (za hypotézy): 27·100/200=13,5, ...

A	13,5	18,5	10,5	10,5	18	29	100
B	13,5	18,5	10,5	10,5	18	29	100
	27	37	21	21	36	58	200

- $$X^2 = \frac{(12-13{,}5)^2}{13{,}5}+\frac{(21-18{,}5)^2}{18{,}5}+\ldots+\frac{(41-29)^2}{29} = 18{,}13$$
- [chisq.test(matrix(c(12,15,21,16,14,7,15,6,21,15,17,41),2,6))]

$$\chi^2 > 11{,}07 = \chi^2_5(0{,}05), \qquad p = 0{,}3\ \%$$

- hypotézu o shodě pstí na kostkách A a B **zamítáme**

příklad – vzdělání matek

test homogenity

vzdělání	porodnice Praha	venkov	celkem
základní	23	11	34
střední	30	17	47
VŠ	17	1	18
celkem	70	29	99

vzdělání	porodnice Praha	venkov	celkem
základní	24,0	10,0	34
střední	33,2	13,8	47
VŠ	12,7	5,3	18
celkem	70	29	99

$\chi^2 = 6{,}12, \qquad p = 4{,}7\ \%$

- kdyby rozdělení vzdělání bylo všude stejné, očekáváme tři možnosti v poměru 34:47:18 (marg. četnosti!), celkem 99
- pražských 70 matek by stejný poměr dalo při **očekávaných** četnostech 70·34/99=24,0, resp. 70·47/99=33,2 resp. 70·18/99=12,7
- podobně pro matky z venkova dostaneme 9,96, po zaokrouhlení 10,0, pro další četnosti 13,8 resp. 5,3

motivační příklad: předvolební průzkum

test nezávislosti (marginální četnosti jsme nezvolili my)

zprávy TV xyz	strana A	B	celkem
sledoval	**11**	**4**	15
nesledoval	**6**	**9**	15
celkem	17	13	30

zprávy TV xyz	strana A	B	celkem
sledoval	**73 %**	**27 %**	100 %
nesledoval	40 %	60 %	100 %
celkem	57 %	43 %	100 %

zprávy TV xyz	strana A	B	celkem
sledoval	**65 %**	**31 %**	50 %
nesledoval	35 %	69 %	50 %
celkem	100 %	100 %	100 %

- 30 voličů bylo dotázáno, které ze dvou stran dali přednost
- souvisí odpovědi se sledováním večerních zpráv na dané TV stanici?
- znamená něco nestejné zastoupení příznivců stran u těch, kteří sledovali?
- znamenají něco nestejné podíly těch, kteří sledovali mezi příznivci dvou stran?

test nezávislosti kvalitativních znaků

- vyšetřujeme **současně** dva znaky v nominálním měřítku u n nezávislých statistických jednotek
- n_{ij} je počet jednotek, kde je současně i-tá hodnota prvního znaku a j-tá hodnota druhého znaku
- celkem je i-tá hodnota prvního znaku u $n_{i\bullet} = \sum_j n_{ij}$ jednotek, j-tá hodnota druhého znaku u $n_{\bullet j} = \sum_i n_{ij}$ jednotek
- kdyby byly znaky nezávislé, byl by pro každou hodnotu jednoho znaku poměr mezi četnostmi hodnot druhého znaku podobný, proto očekávané četnosti jsou $o_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ (podmíněné psti stejné)
- výpočet χ^2 a jeho hodnocení stejné jako u testu homogenity
- předvolební průzkum: $\chi^2 = 3{,}39 \qquad p = 6{,}5\ \%$

příklad: souvisí plánované těhotenství se vzděláním?

vzdělání	plánované ne	ano	celkem
základní	20	14	34
střední	16	31	47
VŠ	5	13	18
celkem	41	58	99

vzdělání	plánované ne	ano	celkem
základní	58,8 %	**42,1 %**	100 %
střední	34,0 %	**66,0 %**	100 %
VŠ	27,8 %	**72,2 %**	100 %
celkem	41,4 %	**58,6 %**	100 %

- je souvislost mezi odpověďmi o plánovaném těhotenství a vzděláním matek?
- kdyby byly znaky nezávislé, byly by podmíněné pravděpodobnosti pro jednotlivá vzdělání stejné, tedy jejich odhady by byly podobné
- test vlastně porovnává procenta u jednotlivých vzdělání
- chí-kvadrát test porovnává skutečně zjištěné četnosti s tím, jaké četnosti bychom v průměru očekávali, kdyby platila nulová hypotéza

příklad: plánovaná těhotenství

skutečné četnosti (očekávané četnosti)

vzdělání	plánované ne	plánované ano	celkem
základní	20 (14,08)	14 (19,92)	34
střední	16 (19,46)	31 (27,54)	47
VŠ	5 (7,46)	13 (10,54)	18
celkem	41	58	99

- odhad pravděpodobnosti, že má matka základní vzdělání: $\hat{P}(vzdel = zakladni) = 34/99$
- odhad pravděpodobnosti, že jde o plánované těhotenství: $\hat{P}(tehot = plan) = 58/99$
- **jsou-li** vzdělání a plánovanost **nezávislé**, pak $P((vzdel = zakladni) \cap (tehot = plan))$ $= P(vzdel = zakladni) \cdot P(tehot = plan) \doteq (34/99) \cdot (58/99)$
- očekávaný počet matek se základním vzděláním a plánovaným těhotenstvím (**za platnosti nulové hypotézy**) odhadneme: $99 \cdot (34/99) \cdot (58/99) = 34 \cdot 58/99 \doteq 19{,}92$

příklad: plánovaná těhotenství

skutečné četnosti (očekávané četnosti)

vzdělání	plánované ne	plánované ano	celkem
základní	20 (14,08)	14 (19,92)	34
střední	16 (19,46)	31 (27,54)	47
VŠ	5 (7,46)	13 (10,54)	18
celkem	41	58	99

$$\chi^2 = \frac{(20-14{,}08)^2}{14{,}08} + \frac{(14-19{,}92)^2}{19{,}92} + \frac{(16-19{,}46)^2}{19{,}46} + \frac{(31-27{,}54)^2}{27{,}54} + \frac{(5-7{,}46)^2}{7{,}46} + \frac{(13-10{,}54)^2}{10{,}54} = 6{,}68$$

příklad: souvisí plánované těhotenství se vzděláním?

ještě jednou ...

- u každé matky zjišťovány dva znaky: dosažené vzdělání, zda těhotenství plánováno

vzdělání	základní	střední	VŠ	celkem
neplánováno	20 (14,1)	16 (19,5)	5 (7,5)	41
plánováno	14 (19,9)	31 (27,5)	13 (10,5)	58
celkem	34	47	18	99

- kdyby nebyla závislost, očekávali bychom u každého vzdělání v průměru stejné procento plánovaných těhotenství, totiž 58/99=58,6 %
  - u zákl. vzdělání $x/34 = 58/99$ tedy $x = 34 \cdot 58/99 = 19{,}9$
  - u středního vzdělání $x/47 = 58/99$ tedy $x = 47 \cdot 58/99 = 27{,}5$
  - u vysokoškolaček $x/18 = 58/99$ tedy $x = 18 \cdot 58/99 = 10{,}5$
- všechny očekávané četnosti jsou dostatečně velké

$$\chi^2 = 6{,}68 > 5{,}99 = \chi^2_2(0{,}05), \qquad p = 3{,}5\ \%$$

příklad: vzdělání snoubenců

ženich	nevěsta základní	nevěsta střední	nevěsta VŠ	celkem
základní	24	12	3	39
střední	7	24	3	34
VŠ	3	9	15	27
celkem	34	45	21	100

- u 100 náhodně vybraných snoubenců bylo zjištěno vzdělání (základní = základní nebo neúplné střední)
- lze považovat vzdělání snoubenců za nezávislá?
- jsou četnosti dost velké?
- nejmenší očekávané četnost (při nezávislosti): $27 \cdot 21/100 = 5{,}67$

příklad: vzdělání snoubenců

	nevěsta			
ženich	základní	střední	VŠ	celkem
základní	24 (13,2)	12 (17,6)	3 (8,2)	39
střední	7 (11,6)	24 (15,3)	3 (7,1)	34
VŠ	3 (9,2)	9 (12,2)	15 (5,7)	27
celkem	34	45	21	100

- $\chi^2 = 43{,}2 > \chi^2_4(0{,}05) = 9{,}5$, $p < 0{,}1$ %
- na 5 % hladině jsme prokázali závislost
- vzdělání snoubenců nelze považovat za nezávislá
- četnosti na diagonále jsou větší, než očekáváme za nezávislosti
- četnosti daleko od diagonály (velký rozdíl ve vzdělání) jsou menší, než očekáváme za nezávislosti

čtyřpolní tabulka (tabulka 2×2)

speciální případ kontingenční tabulky

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n
- **sílu závislosti** lze měřit ϕ-koeficientem [phi coefficient] (čtyřpolní korelační koeficient)

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

- ϕ je (jako každý korelační koeficient) mezi –1 a 1
- pro

11	4	15
6	9	15
17	13	30

  vyjde

$$\phi = \frac{11 \cdot 9 - 4 \cdot 6}{\sqrt{15 \cdot 15 \cdot 17 \cdot 13}} = 0{,}34$$

příklad: předvolební průzkum

- $\phi > 0$ znamená, že četnosti na hlavní diagonále (indexy 1,1 a 2,2) převládají nad četnostmi na vedlejší diagonále (indexy 1,2 a 2,1)
- v našem příkladu

	strana		
TV XY	A	B	celkem
sledoval	**11**	**4**	15
nesledoval	**6**	**9**	15
celkem	17	13	30

  vychází $\phi = 0{,}34 > 0$
  (tedy kladné), protože je $11{\cdot}9 > 6{\cdot}4$

čtyřpolní tabulka – prokazování závislosti

- chí-kvadrát porovnávající teoretické a očekávané četnosti lze upravit na tvar

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

- nezávislost se na hladině α zamítá, je-li $\chi^2 \geq \chi^2_1(\alpha)$
- příklad (předvolební průzkum)

$$\chi^2 = \frac{30 \cdot (11 \cdot 9 - 4 \cdot 6)^2}{15 \cdot 15 \cdot 17 \cdot 13} = 3{,}39 = 30 \cdot 0{,}34^2$$

- závislost jsme na 5% hladině neprokázali, neboť

$$3{,}39 < 3{,}84 = \chi^2_1(0{,}05), \qquad p = 6{,}5\ \%$$

malé očekávané četnosti ve čtyřpolní tabulce

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

- stále je třeba, aby byly očekávané četnosti dost velké (≥ 5)
- **Yatesova korekce** umožní rozhodnutí i při menších četnostech tím, že zmenší čitatele

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

- nezávislost se zamítá, je-li opět $\chi^2_{\text{Yates}} \geq \chi^2_1(\alpha)$
- **Fisherův exaktní test** počítá přímo p-hodnotu

příklad: souvislost délky kojení a plánování těhotenství

vlastně dva příklady

	Praha a venkov			venkov		
těhot.	neplán	plán.	celkem	neplán.	plán.	celkem
ve 24. t. nekojí	35	36	71	13	9	22
ve 24. t. kojí	6	22	28	1	6	7
celkem	41	58	99	14	15	29

- bez ohledu na místo: $\chi^2 = 6{,}43$, $p = 1{,}1$ %, $\chi^2_{\text{Yates}} = 5{,}33$, $p = 2{,}1$ % (nejm. četnost $41 \cdot 28/99 = 11{,}6$) Fisherův exaktní test: $p = 1{,}3$ %
- venkov: $\chi^2 = 4{,}27$, $p = 3{,}9$ %, $\chi^2_{\text{Yates}} = 2{,}66$, $p = 10{,}3$ % (nejm. četnost $14 \cdot 7/29 = 3{,}4$) Fisherův exaktní test: $p = 8{,}0$ %

Simpsonův paradox

dílčí tabulky mohou ukazovat na závislost jiného směru, než jejich součet

venkov	A	B	celkem
sledoval	34	5	39
nesledoval	28	2	30
celkem	62	7	69

město	A	B	celkem
sledoval	4	29	33
nesledoval	6	35	42
celkem	10	64	74

$$\phi_{\text{venkov}} = -0{,}10 \qquad \phi_{\text{město}} = -0{,}04$$

celkem	A	B	celkem
sledoval	38	34	72
nesledoval	34	37	71
celkem	72	71	143

$$\phi_{\text{celkem}} = 0{,}05$$

- po spojení dvou tabulek se záporným ϕ-koeficientem vyšla tabulka s kladným ϕ-koeficientem

závislost mezi nula-jedničkovým a kvantitativním znakem

- dva nezávislé výběry, např. hoši $X_1, \ldots, X_{n_0}$ a dívky $X_{n_0+1}, \ldots, X_{n_0+n_1}$, vždy normální rozdělení jako pro dvouvýběrový t-test
- otázka: jak silně souvisí sledovaná vlastnost a pohlaví?
- označme pohlaví formálně $Y_i = 0$ pro chlapce a $Y_i = 1$ pro děvčata
- korelační koeficient $r_{X,Y}$ mezi těmito veličinami se dá zapsat také jako

$$r_{\text{bis}} = \frac{\bar{X}_1 - \bar{X}_0}{S}\sqrt{\frac{n_0 n_1}{n(n-1)}}$$

- S je směrodatná odchylka spočítaná bez ohledu na pohlaví, $n = n_0 + n_1$ je celkový počet měření v obou výběrech
- r_{bis} **bodově-biseriální korelační koeficient**

příklad: výška desetiletých

- stejná data jako dvouvýběrový test (data ze str. 179)
- $\bar{X}_0 = 139{,}13, \qquad n_0 = 15$
- $\bar{X}_1 = 140{,}83, \qquad n_1 = 12$
- $S^2 = 38{,}18, \qquad S = 6{,}18$
- $$r_{\text{bis}} = \frac{140{,}83 - 139{,}13}{6{,}18}\sqrt{\frac{15 \cdot 12}{27 \cdot 26}} = 0{,}139$$
- H_0 : nezávislost
- má-li X normální rozdělení, lze použít stejný test, jako u korelačního koeficientu; je to ekvivalentní dvouvýběrovému t-testu (při stejných populačních rozptylech

přehled korelačních koeficientů

- základním je (momentový) Pearsonův
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
- když místo hodnot x_i, y_i dosadíme jejich pořadí R_i, Q_i, dostaneme (pořadový) Spearmanův korelační koeficient
$$r_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$
- je-li jedna z veličin nula-jedničková, vyjde biseriální korelační koeficient r_{bis}
- jsou-li obě veličiny nula-jedničkové, dostaneme ϕ-koeficient (čtyřpolní korelační koeficient)

přehled testů o populačních mírách polohy

rozdělení	normální	spojité
populační parametr (o čem je hypotéza)	populační průměr	populační medián
jeden výběr	jednovýběrový t-test	znaménkový Wilcoxon
výběr dvojic	párový t-test	znaménkový Wilcoxon
dva nezávislé výběry	dvouvýběrový t-test	Mann-Whitney

Přehled I

1. průměr, vážený průměr, medián, kvartily
2. směrodatná odchylka, rozptyl, střední díference
3. z-skóry, šikmost, špičatost
4. geografický střed, geografický medián
5. Giniho koeficient (index)
6. Lorenzova křivka, její konstrukce, souvislost s Giniho indexem
7. (Pearsonův) korelační koeficient r_{xy}
8. střední hodnota a (populační) rozptyl náhodné veličiny
9. nezávislost náhodných veličin, populační korel. koeficient ρ_{XY}
10. alternativní rozdělení s parametrem π, jeho střední hodnota a rozptyl
11. model binomického rozdělení bi(n, π)
12. Poissonovo rozdělení Po(λ), souvislost s binomickým rozdělením
13. normální rozdělení $\mathrm{N}(\mu, \sigma^2)$, význam parametrů

Přehled II

14. střední chyba průměru $\overline{X}_i$ z náhodného výběru $X_1, \dots, X_n$
15. interval spolehlivosti pro střední hodnotu $\mu = \mathrm{E}\,X$ z náhodného výběru $X_1, \dots, X_n$
16. nulová a alternativní hypotéza
17. chyba 1. a 2. druhu
18. hladina testu, síla testu
19. p-hodnota
20. test o pravděpodobnosti π binomického rozdělení bi(n, π)
21. jednovýběrový a párový t-test, předpoklady
22. dvouvýběrový t-test, předpoklady
23. dvouvýběrový Wilcoxonův (Mannův-Whitneyův) test, předpoklady
24. párový Wilcoxonův test, předpoklady

Přehled III

25. Spearmanův korelační koeficient r_S, vztah k Pearsonovu korelačnímu koeficientu r
26. princip metody nejmenších čtverců
27. regresní přímka $y = \beta_0 + \beta_1 x$, prokazování závislosti
28. interpretace koeficientů regresní přímky, predikce pomocí odhadu regresní přímky $y = b_0 + b_1 x$
29. koeficient determinace R^2
30. model multinomického rozdělení s parametry $n, \pi_1, \dots, \pi_k$, souvislost s binomickým rozdělením
31. test dobré shody o parametrech $\pi_1, \dots, \pi_k$ multinomického rozdělení
32. test homogenity/nezávislosti v kontingenční tabulce
33. čtyřpolní tabulka

organizace zkoušení

- zkoušet mohu jen studenty včas zapsané v SIS, a to v PUA (Alb. 6) (případně v B5, Viničná 7, bude-li třeba)
- na začátku zkoušky student musí již mít zápočet (aspoň na jednom z míst: SIS, index)
- každý student dostane vlastní písemné zadání
- výpočty lze provádět v Excelu, v R nebo na vlastní kalkulačce; jiné pomůcky nejsou dovoleny
- student bude mít možnost ústně vysvětlit svůj postup, dále bude odpovídat na dotazy
- budu se ptát na základní věci i mimo písemně položené otázky

několik slov zkoušce

- cílem zkoušení je zjistit, do jaké míry studentka či student zvládl obsah přednášky
- důležité jsou základní pojmy, myšlenkové konstrukce, nikoliv detaily
- u vzorečků je jejich smysl důležitější než symboly
- dám přednost správnému vysvětlení smyslu pomocí nepřesně zvolených slov před nesprávně kombinovanými přesnými termíny (i když na jedničku to pak asi nebude)
- netoužím někoho od zkoušky vyhodit (přidělával bych si práci), ale nechci nikomu ubližovat tím, že by u zkoušky prošel i bez těch nejzákladnějších znalostí