Statistika

(MD360P03Z, MD360P03U)
ak. rok 2013/2014

Karel Zvára

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

(naposledy upraveno 7. ledna 2014)

1. přednáška

- úvod, přehled témat
- co a jak zjišťujeme, měřítka, veličina
- histogram, třídění
- variační řada, pořadí
- medián, kvartily, percentily
- průměr, vážený průměr
- modus

literatura

- K. Zvára: Základy statistiky v prostředí R, Karolinum, Praha 2013 (edice Biomedicínská statistika IV)
  Karolinum (Celetná 18), Lípová 6, Neoluxor (Václ. nám 41)
- K. Zvára: Biostatistika, Karolinum, Praha 1998, 2000, 2001, 2003, 2006, 2008 (pozor na jiné značení)
- Z. Pavlík, K. Kühnl: Úvod do kvantitativních metod pro geografy, SPN Praha, 1981
- slajdy přednášky na adrese
  http://www.karlin.mff.cuni.cz/∼zvara
- může dojít k drobným úpravám slajdů před přednáškou i po ní

cvičení, zápočet, zkouška

- cvičení v počítačové učebně PUA (suterén Albertov 6) nebo v učebně B5 (Viničná 7)
- 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
- znalosti ze cvičení nemusí u zkoušky stačit!
- 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 (populace, základní soubor)
- v souboru je 80 kuřáků a 120 nekuřáků (výběr)
- 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
- dotazy na populaci (základní soubor):
  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)
- přihlížíme 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,pro fyzika přinejmenším ordinální, možná dokonce poměrové)

veličina

- číselně vyjádřený výsledek měření (zjišťování)
- číselné 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, které má vyjádřit **úroveň** výšek, podobně pro dívky)

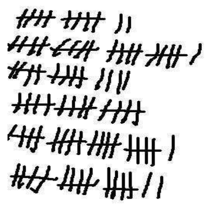
příklad: 100 hodů kostkou

počty puntíků na kostce coby různé obrázky **–** nominální znak

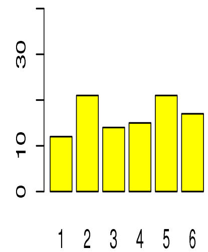
kostka A										kostka B									
4	2	5	6	3	1	1	2	2	2	1	4	6	2	3	2	6	1	5	2
2	4	5	3	1	1	3	5	5	5	5	6	5	5	6	4	2	4	5	6
4	3	2	5	5	5	2	2	5	2	3	6	3	6	5	6	1	3	5	1
2	6	5	5	2	3	6	6	4	6	6	6	2	1	1	2	6	3	2	3
5	4	1	4	2	2	4	5	2	5	4	4	1	6	6	2	6	3	2	6
5	5	3	3	5	3	6	6	6	5	2	6	1	2	6	1	5	5	6	5
3	5	4	5	1	1	4	3	2	4	6	6	5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

zpracování četností (kostka A)

čárkovací metoda, absolutní a relativní četnosti, barplot pro znak v kvalitativním měřítku

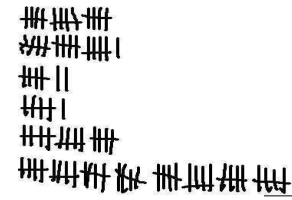


j	n_j	$f_j = n_j/n$
1	12	0,12
2	21	0,21
3	14	0,14
4	15	0,15
5	21	0,21
6	17	0,17
	$n = 100$	1,00

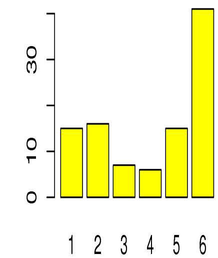


zpracování četností (kostka B)

čárkovací metoda, absolutní a relativní četnosti, barplot pro znak v kvalitativním měřítku



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
	$n = 100$	1,00



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 – **barplot** (nepřesně histogram) (výška obdélníka je úměrná četnosti)
- rozhodování o kvalitě kostky (zda je symetrická) je úlohou **statistické indukce** [inference] – bude později

příklad: věk 99 matek

99 zjištěných hodnot – soubor naměřených hodnot

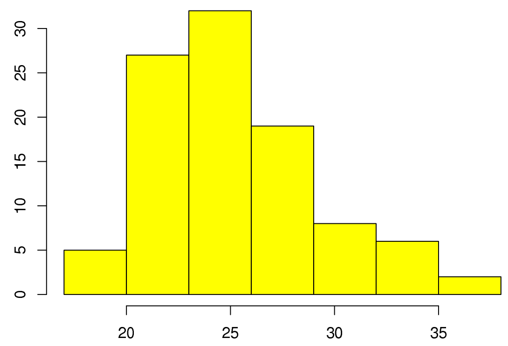
26	35	21	25	27	24	24	30	23	18
35	21	25	26	26	19	29	22	21	27
26	30	28	28	27	29	27	26	21	23
24	21	28	25	34	24	21	28	25	28
22	26	32	22	32	25	21	25	24	32
24	22	31	33	23	30	26	27	25	24
24	23	25	23	26	28	24	25	25	26
28	28	22	23	20	20	21	31	24	21
29	28	26	38	20	23	25	37	33	23
27	23	21	25	21	33	22	29	21	

- spojité hodnoty pouze zaokrouhleny na celá čísla
- Umíme něco užitečného z dat vytáhnout?
- Můžeme si rychle udělat představu?
- Můžeme tyto údaje porovnat s daty 10 roků starými?
- Umíme vhodným číslem charakterizovat úroveň a variabilitu?

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

histogram pro znak v kvantitativním měřítku

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



třídění, třídní četnosti

jak jsme k histogramu dospěli

- 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

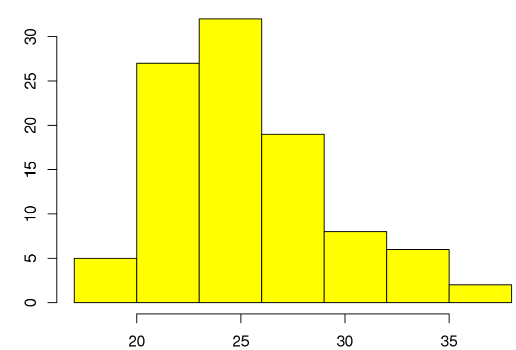
$k = 7$, $n = 99$

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 mírám polohy věku matek

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

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



příklad: tolary

měsíční příjmy 99 osob ve fiktivní měně

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

velmi nepřehledná informace

příklad: tolary

variační řada = hodnoty jsou **uspořádané**

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	

přehlednější informace
budou užitečné četnosti hodnot?

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; na svislou osu se pak 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	

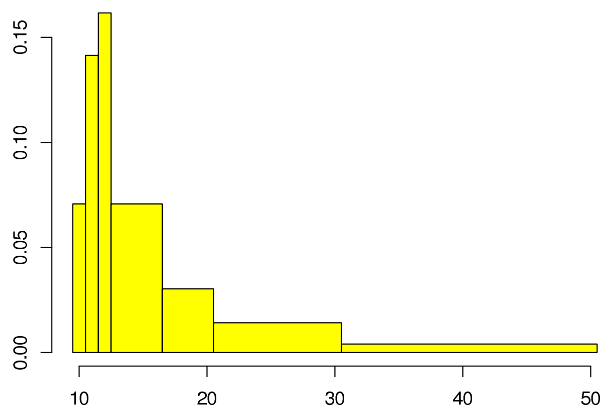
x_j	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1

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

třída	10	11	12	13–16	17–20	21–30	31–50	celk.
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	

příklad (tolary): histogram

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



plocha obdélníka = délka intervalu × hustota

variační řada, pořadí

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

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

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	32	25	27	25	31	23	28
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)

výběrové charakteristiky polohy: medián

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

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

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

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

kvartily, percentily

- **dolní (horní) kvartil** Q_1 (Q_3) [lower (upper) quartile] vyděluje čtvrtinu nejmenších (největších) hodnot od ostatních
- **percentil** [percentile] x_p vyděluje $100p$ % nejmenších hodnot od ostatních
- konkrétní výpočet percentilu může být složitý
  - $100p$ nemusí být celé číslo
  - v datech se mohou čísla opakovat
- výpočet percentilů – mnoho vzorečků (další požadavky)
- kvartil – speciální případ percentilu: $Q_1 = x_{1/4} = x_{0,25}$, $Q_3 = x_{3/4} = x_{0,75}$ `quantile(x,probs=c(1/4,3/4))`
- medián je také percentil, totiž $x_{0,5}$, podobně minimum ($p = 0$) a maximum ($p = 1$)
- `fivenum(x)` podobné příkazu `quantile(x,probs=0:4/4)`
- 1., 5. a 9. **decil** jsou vhodné k popisu rozdělení příjmů

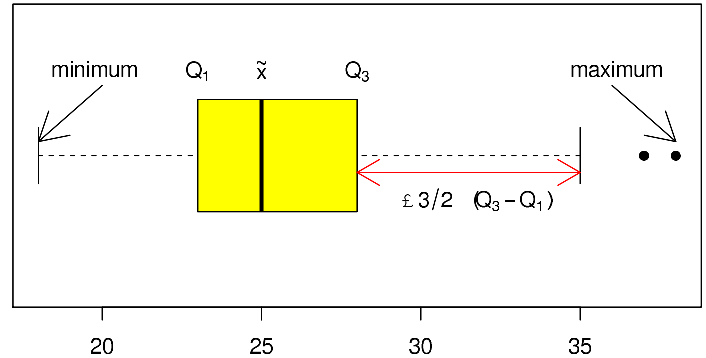
výpočet percentilů (jako v R), jen pro ilustraci

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 $n = 99, p = 0{,}25$ (věk matek) bude
$$k = \lfloor 1 + (99-1)\cdot 0{,}25 \rfloor = \lfloor 1 + 24{,}5 \rfloor = \lfloor 25{,}5 \rfloor = 25$$
$$q = 25{,}5 - 25 = 0{,}5$$
$$Q_1 = x_{0{,}25} = (1-0{,}5)\cdot x_{(25)} + 0{,}5\cdot x_{(26)} = 23$$

krabicový diagram (boxplot)

věk 99 matek



- grafické znázornění mediánu, kvartilů, minima, maxima, případně „odlehlých" pozorování
- tykadlo sahá od kvartilu k co nejvzdálenějšímu pozorování, ale takovému, aby délka tykadla byla nejvýše $\frac{3}{2}(Q_3 - Q_1)$

příklad: tolary

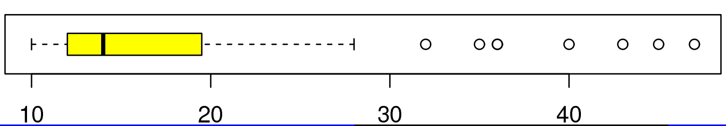
$\tilde{x} = x_{0{,}5} = 14 \qquad x_0 = 10 \qquad x_1 = 47$ tolary rozpětí

$p = 3/4,\ k = \lfloor 1 + 98\cdot 3/4 \rfloor = \lfloor 74{,}5 \rfloor = 74,\ q = 74{,}5 - 74,$

$\Rightarrow Q_3 = (1-0{,}5)\cdot 19 + 0{,}5\cdot 20 = 19{,}5, \qquad Q_1 = 12$

$(Q_3 - Q_1)\cdot 1{,}5 = 7{,}5\cdot 1{,}5 = 11{,}25, \qquad 19{,}5 + 11{,}25 = 30{,}75$

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	



další míry polohy: průměr

- **průměr** [mean] (kdyby bylo všech n hodnot stejných) `mean(x)`
$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- pomocí četností **vážený průměr**: [weighted mean]
$$\bar{x} = \frac{1}{n}(n_1 x_1^* + \ldots + n_k x_k^*) = \frac{1}{n}\sum_{j=1}^{k} n_j x_j^* = \sum_{j=1}^{k} \frac{n_j}{n} x_j^* = \frac{\sum_{j=1}^{k} n_j x_j^*}{\sum_{j=1}^{k} n_j}$$
- obecně **vážený průměr** s vahami $w_1, \ldots, w_k$ hodnot $x_1^*, \ldots, x_k^*$ (váhy nutně nezáporné: $w_j \geq 0$, $\sum_{j=1}^{k} w_j > 0$)
$$\boxed{\bar{x} = \bar{x}_w = \frac{\sum_{j=1}^{k} w_j x_j^*}{\sum_{j=1}^{k} w_j}}$$

příklad: HDP zemí V4 v roce 2010

obyvatelé v tisících, HDP na obyvatele v tisících PPT (standard kupní síly)

země	obyvatel	HDP	součin	podíl obyv.
CZ	10 517,247	19,4	204 034,59	16,33 %
HU	9 976,062	15,8	157 621,78	15,49 %
PL	38 441,588	15,3	588 156,30	59,58 %
SK	5 477,038	18,0	98 586,68	8,50 %
celkem	64 411,935	68,5	1 048 399,35	

- průměr (nevážený): $68{,}5/4 = 17{,}125$
- vážený průměr (vahami počet obyvatel): $1048399{,}35/64411{,}935 \doteq 16{,}276$
- vážený průměr (vahami podíl obyvatel): 16,276
- každý nenulový násobek vah vede ke stejnému váženému průměru
- který průměr vypovídá správně (rozumně)?

modus

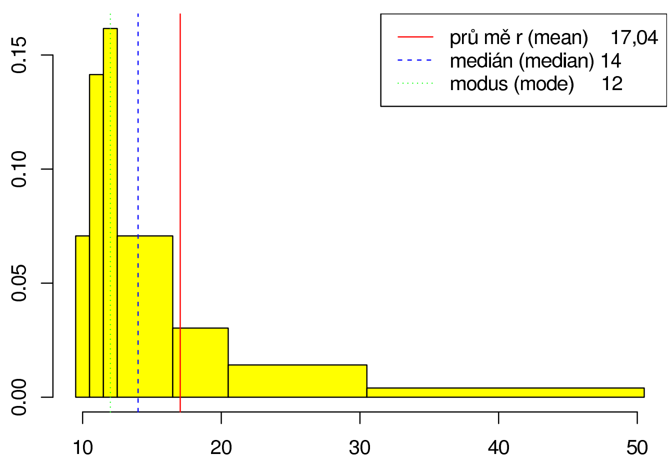
- **modus** $\hat{x}$ [mode] nejčastější hodnota
- modus lze počítat také pro nominální či ordinální měřítko, ale jako míru polohy jej lze interpretovat jen do jisté míry u ordinálního měřítka
- např. příjem v tolarech:

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

- maximální četnost 16 nastává pro příjem 12 tolarů
- modus je tedy $\hat{x} = 12$
- modus nemusí být určen jednoznačně
- v příkladu nejsou chudí, neboť nikdo nemá příjem pod 60 % z mediánu ($0{,}6 \cdot \tilde{x} = 0{,}6 \cdot 14 = 8{,}4$)

příklad (tolary): porovnání tří měr polohy

zpravidla je mean $\geq$ median $\geq$ modus



2. přednáška

- vlastnosti charakteristik polohy
- vlastnosti charakteristik variability
- rozptyl, směrodatná odchylka
- střední odchylka, střední diference
- *z*-skór, standardizace
- šikmost, špičatost
- korelační koeficient
- Giniho koeficient koncentrace
- geografický střed, geografický medián

vlastnosti charakteristik polohy

- charakteristiky (míry) polohy mají měřit úroveň kvantitativního (spojitého) znaku (velký – malý, hodně – málo, ...)
- **posunutí:** změníme-li všechny hodnoty x_i tak, že přidáme ke každé stejnou konstantu a, změní se o tutéž konstantu také charakteristika polohy
- **změna měřítka:** 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
- obecně pro míru polohy $\mu(x)$ platí

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

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

charakteristiky variability

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

$$\sigma(a+x) = \sigma(x), \qquad \text{(srovnej s } \mu(a+x) = a + \mu(x))$$
$$\sigma(b \cdot x) = b \cdot \sigma(x), \quad b > 0, \qquad \text{(srovnej s } \mu(b \cdot x) = b \cdot \mu(x))$$

- **posunutí:** přičtením stejné konstanty a (tj. posunutím) se charakteristika variability nezmění (nezávisí na poloze)
- **změna měřítka:** vynásobení kladnou konstantou b 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 $s^2_{a+b\cdot x} = b^2 \cdot s^2_x$)

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

- nechť $x_1 = 1, x_2 = 3, x_3 = 8$ (tedy $n = 3$), 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
- polovina průměrného čtverce vzájemné závislosti:

$$s_x^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2$$

- **směrodatná odchylka** (standardní 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: mají-li intervaly délku h, odečti $h^2/12$)

příklad – tolary

- rozpětí: $R = 47 - 10 = 37$
- kvartilové rozpětí: $R_Q = 19{,}5 - 12 = 7{,}5$
- rozptyl

$$s^2 = \frac{1}{98}\left((10^2 + 10^2 + \ldots + 45^2 + 47^2) - 99 \cdot \left(\frac{1687}{99}\right)^2\right)$$
$$= \frac{1}{98}\left((7 \cdot 10^2 + 14 \cdot 11^2 + \ldots + 45^2 + 47^2) - 99 \cdot \left(\frac{1687}{99}\right)^2\right)$$
$$= 65{,}080 \doteq 8{,}067^2$$

- směrodatná odchylka je 8,067

Var. řada tolary

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,"-")))`

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

příklad pro `x <- c(1, 3, 8)`

- **střední odchylka**

$$d = (|1-3| + |3-3| + |8-3|)/3 = 7/3 \doteq 2{,}33$$

`mean(abs(x-median(x)))`

- **střední diference** tabulka rozdílů:

$$\begin{matrix} 0 & -2 & -7 \\ 2 & 0 & -5 \\ 7 & 5 & 0 \end{matrix}$$

$$\Delta = \frac{1}{3 \cdot 3}(0+2+7+2+0+5+7+5+0) = \frac{28}{9} \doteq 3{,}11$$
$$\Delta = \frac{2}{3 \cdot 3}((8-1)+(8-3)+(3-1)) = \frac{2(7+5+2)}{9} \doteq 3{,}11$$

`Delta = mean(abs(outer(x,x,"-")))`

normované charakteristiky rozptýlenosti

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

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

- **(Giniho) koeficient koncentrace** `reldist::gini(x)`

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

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-skór nezávisí na posunutí ani na změně měřítka
- z-skóry jsou bezrozměrné ⇒ 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)`

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

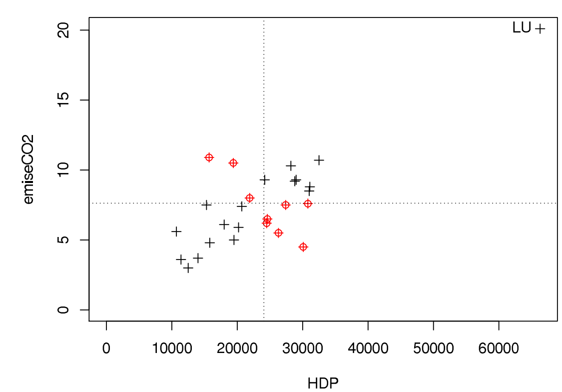
- někdy se počítají odhady populační šikmosti a špičatosti jinak (Excel: s_x jinak, Fisherovo g_1, g_2 – pro zajímavost)

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

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

příklad: souvisí emise CO_2 s HDP?

údaje o zemích EU z roku 2010



lze charakterizovat **sílu závislosti** číslem?

měření síly závislosti

- měříme dva znaky v kvantitativním měřítku: $(x_i, y_i), i = 1, \dots, n$ (např. HDP jako x_i, emise CO_2 jako y_i
- závislosti na fyzikálním měřítku se vyhneme použitím z-skórů
- (výběrový) **korelační koeficient**

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

- klasicky se definuje $r_{xy} = \frac{s_{xy}}{s_x s_y}$, kde

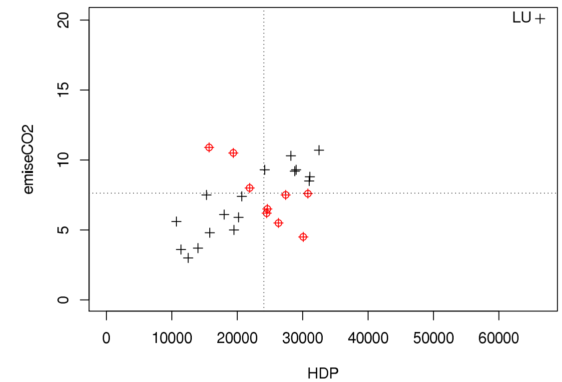
$$s_{xy} = \frac{1}{n-1} \sum_{x=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

je (výběrová) **kovariance**

- platí $-1 \leq r \leq 1$

příklad: souvisí emise CO_2 s HDP?

údaje o zemích EU z roku 2010



$r = 0{,}79$ (bez Lucemburska jen $r = 0{,}52$)

charakteristiky polohy v geografii/demografii

- místo x můžeme označovat měřené hodnoty jako y, princip pojmů je stejný, označení je jen konvence
- často známe jen průměry a četnosti v dílčích souborech: průměry se označí jako y_j^*, četnosti opět n_j
- 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))
```

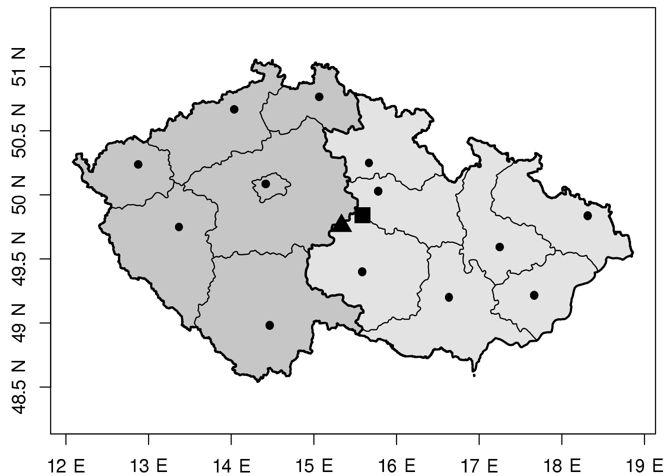
$$\frac{51{,}1 + 47{,}8}{2} = 49{,}4$$

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 stejné velikosti
  - hodnocená vlastnost určí velikost objektů (např. počet obyvatel územní jednotky)
  - uspořádání hodnocení znaků dáno zvolenou geografickou vlastností (např. zeměpisnou délkou)

příklad: geografický střed obyvatel ČR

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



příklad: geografický střed obyvatel ČR

použijeme jen údaje o krajích ČR v roce 2006

	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	Moravskoslezský	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

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

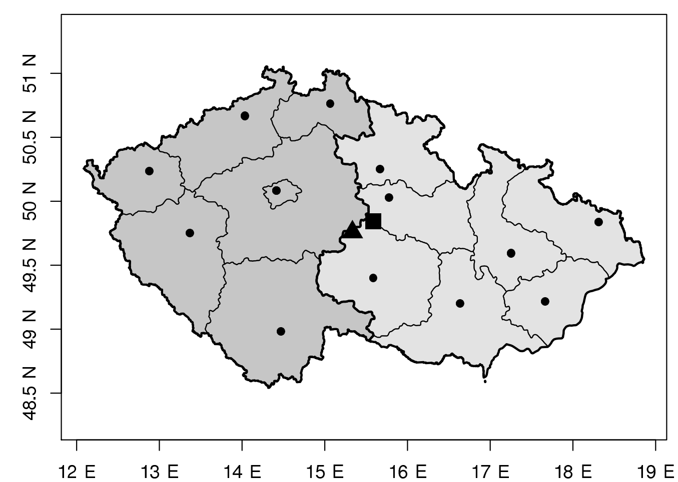
příklad: geografický medián

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)



3. přednáška

- Giniho koeficient koncentrace
- Lorenzova křivka
- Theilův index

Giniho koeficient koncentrace

(místo x nyní píšeme y)

- **Giniho koeficient koncentrace** 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 ($0 = y_{(1)} = \ldots = y_{(n-1)} < y_{(n)} = 100$), pak je

$$\bar{y} = \frac{100}{n} \qquad \Delta = \frac{2(n-1)100}{n^2}$$

$$G = \frac{2(n-1)100}{n^2} \cdot \frac{n}{2 \cdot 100} = \frac{n-1}{n} \xrightarrow[n\to\infty]{} 1$$

- Lorenzova křivka bude jemnějším nástrojem

příklad: rozloha lesů zemí V4 v tisících ha

- CZ 2657 HU 2039 PL 9319 SK 1938
- průměrná rozloha je $\bar{y} = 3988{,}25$ (tisíce ha)
- jednotlivé diference:

	CZ	HU	PL	SK
CZ	0	618	−6662	719
HU	−618	0	−7280	101
PL	6662	7280	0	7381
SK	−719	−101	−7381	0

- střední diference (tisíce ha)

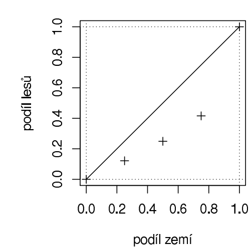
$$\Delta = \frac{0 + 618 + 6662 + \ldots + 7381 + 0}{4 * 4} = 2845{,}125$$

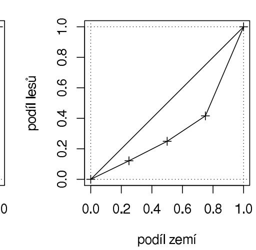
- Giniho koeficient (ten je bezrozměrný!)

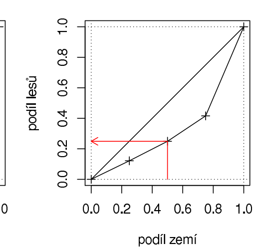
$$G = \frac{\Delta}{2\bar{y}} = \frac{2845{,}125}{2 \cdot 3988{,}25} = 0{,}357$$

příklad: podíl z celkové rozlohy lesů zemí V4

- CZ 16,7 % HU 12,8 % PL 58,4 % SK 12,1 %
- totéž v pořadí o nejmenšího podílu
  SK 12,1 % HU 12,8 % CZ 16,7 % PL 58,4 %
- postupné součty (kumulativní relativní četnosti)
  SK 12,1 % HU 24,9 % CZ 41,6 % PL 100, %
- polovina na lesy nejchudších (SK, HU) má čtvrtinu lesů







Lorenzova křivka

- vodorovná osa: postupné načítání jednotek od nejchudších, jako díl celku
- svislá osa: postupné načítání bohatství (části zdroje) od nejchudších, jako díl celku
- zajímá nás plocha nad touto lomenou čarou a pod úhlopříčkou jednotkového čtverce
- plocha měří nerovnoměrnost rozdělení nějakého zdroje
- kdyby dostala každá jednotka stejně, bude velikost plochy nulová
- kdyby všechno dostala jediná z n jednotek, lomená čára bude nulová až do $(n-1)/n$; pro $n \to \infty$ je $(n-1)/n \to 1$, plocha $=$ dolní trojúhelník
- **dvojnásobek** této plochy ($=$ Giniho koeficient koncentrace) porovnává tuto plochu s plochou dolního trojúhelníku

Lorenzova křivka, shrnutí konstrukce

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

- variační řada: $0 \le y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$ `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 $[j/n; Y_{(j)}/(\sum_{i=1}^{n} y_i)$, $0 \le j \le n$, j/n – díl populace $Y_{(j)}/(\sum_{i=1}^{n} y_i)$ – díl bohatství
- `n = length(y)`
- `Y = c(0,cumsum(sort(y)))`
- `plot((0:n)/n,Y/sum(y),pch=3,asp=1)`
- `lines((0:n)/n,Y/sum(y))`
- `lines(0:1,0:1); abline(h=0:1,v=0:1,lty=3)`

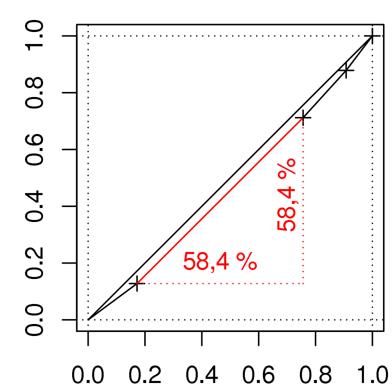
příklad: plochy lesů V4 s přihlédnutím k rozloze zemí

země	rozloha	lesy	zalesnění	rel. rozloha	rel. lesy
CZ	78 865	26 570	33,7 %	15,1 %	16,7 %
HU	89 608	20 390	22,8 %	17,2 %	12,8 %
PL	304 255	93 190	30,6 %	58,4 %	58,4 %
SK	48 105	19 380	40,3 %	9,2 %	12,1 %
celkem	520 833	159 530	30,6 %	100,0 %	100,0 %

- Nakolik jsou nerovnoměrně rozmístěny lesy na území V4?
- každému čtverečnímu km přidělíme příslušný díl, např. v CZ je to 0,337 km^2, v HU podobně 0,228 km^2)
- v HU tak přibude 89 608krát hodnota 0,228, v PL 304 255krát hodnota 0,306 atd.
- při pravidelném přidávání splynou jednotlivé body pro danou zemi v úsečku
- průmět úsečky na osu x = relativní rozloha; průmět na osu y = relativní plocha lesů; pořadí dáno zalesněním

příklad: Lorenzova křivka hustoty zalesnění

body: postupně se načítá rel. velikost zemí (osa X) a rel. plocha lesů (osa Y) v pořadí HU, PL, CZ, SK (např. PL má 58,4 % celkové rozlohy i 58,4 % plochy lesů)



co jsme zjistili (zobecnění příkladu)

- neznáme plochu lesa na jednotlivých čtverečních km
- známe průměrnou plochu lesa na km^2 v každé zemi ($y_i^{\text{prům}}$)
- průměrnou plochu opakujeme tolikrát, kolik km^2 má daná země (četnost), tj. vážíme počtem km^2 (x_i)
- známe tedy celkovou plochu lesů v každé zemi ($y_i = x_i y_i^{\text{prům}}$)
- pořadí zemí dáno průměrnou plochou lesa na km^2 (hustotou lesa) jednotlivých zemí ($y_1^{\text{prům}} \le y_2^{\text{prům}} \le \ldots \le y_k^{\text{prům}}$)
- přírůstky souřadnic bodů:
  - vodorovně: relativní rozloha dané země (mezi všemi zeměmi)
$$\frac{x_i}{\sum_\ell x_\ell}$$
  - svisle: relativní plocha lesů (mezi všemi lesy)
$$\frac{y_i}{\sum_\ell y_\ell} = \frac{x_i y_i^{\text{prům}}}{\sum_\ell x_\ell y_\ell^{\text{prům}}}$$

orientačně shrnutí výpočtu v případě vah (Lorenz, Gini)

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

- kumulativní **součty**
$X_j = \sum_{i=1}^j x_i, X_0 = 0, \quad Y_j = \sum_{i=1}^j y_i = \sum_{i=1}^j x_i y_i^{\text{prům}}, Y_0 = 0$
- Lorenzova křivka spojuje body $\left[\frac{X_j}{X_k}; \frac{Y_j}{Y_k}\right], \qquad j = 0, 1, \ldots, k$
- 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

poznámky

- nezáleží na zvolených fyzikálních jednotkách (např. km vers. ha)
- ve všech případech je **pořadí** sčítanců dáno pořadím „hustot“ $y_i^{\text{prům}} = \frac{y_i}{x_i}$ (např. lesy/rozloha), tj. $y_1^{\text{prům}} \le \ldots \le y_k^{\text{prům}}$
- na svislé ose jde o podíl stat. jednotky na bohatství
- na vodorovné ose jde o podíl velikosti stat. jednotky s daným bohatstvím mezi všemi jednotkami
- hrubší hodnocení (např. kraje, nikoliv okresy) znamená **menší** hodnotu Giniho koeficientu! (obecně, lze dokázat)
- velikost poklesu Giniho koeficientu po hrubším hodnocení není snadné vyjádřit (vysvětlit)

příklad: příjmy 14 osob ve 3 skupinách, Giniho koeficient

skupina	příjem y_i	n_i	průměr	Gini
A	200 150	2	175,00	0,07142857
B	80 70 60 60	4	67,50	0,06481481
C	20 20 18 18 15 15 10 10	8	15,75	0,1309524
celk.	746	14	53,29	0,5090004

skupinové průměry: $G = 0{,}485$, původní data: $G = 0{,}509$

```
prijem=c(200,150,...)
Skup = factor(c(rep("A",2),rep("B",4),rep("C",8)))
ni=table(Skup)
prumery = tapply(prijem,Skup,mean)
require(reldist)
gini(prijem)     #0.5090004
gini(prumery,ni)     #0.4848717
tapply(prijem,Skup,gini)
```

Použití průměrů (včetně jejich četností) snížilo G

Theilův index

- $y_1, \ldots, y_n$ bohatství jednotlivých subjektů (např. jednotky)

$$T = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{\bar{y}} \ln\left(\frac{y_i}{\bar{y}}\right)$$

- vážený průměr hodnot $\ln(y_i/\bar{y})$, váhy $y_i/\bar{y}$, součet vah je n
- měří nerovnoměrnost (variabilitu) rozdělení bohatství
- souvisí s pojmem Shannonovy entropie (nejistota v teorii informace, diverzita)

$$S = -\sum_{i=1}^{n} \frac{y_i}{\sum_\ell y_\ell} \ln\left(\frac{y_i}{\sum_\ell y_\ell}\right)$$

- maximální hodnota entropie $S_{\max}$ je pro $y_1 = y_2 = \ldots = y_n$
- lze dokázat, že $T = S_{\max} - S$ a $T \le \ln(n)$
- čím větší nerovnoměrnost, tím větší T (stejně jako Giniho koeficient)

příklad: příjmy 14 osob

skupina	příjem y_i	n_i	průměr
A	200 150	2	175,00
B	80 70 60 60	4	67,50
C	20 20 18 18 15 15 10 10	8	15,75
celk.	746	14	53,29

$$T = \frac{1}{18}\left(\frac{200}{53{,}29}\ln\left(\frac{200}{53{,}29}\right) + \cdots + \frac{10}{53{,}29}\ln\left(\frac{10}{53{,}29}\right)\right) = 0{,}450$$

```
prijem=c(200,150,80,70,60,60,20,20,18,18,15,15,10,10)
library(ineq)
Theil(prijem)
```

Co kdybychom znali jen průměry a počty hodnot ve skupinách?

Theilův index po skupinách

- y_{ij} příjem j-tého v i-té skupině s n_i jedinci, $i = 1, \ldots, k$
- celkem $n = \sum_{i=1}^{k} n_i$ jedinců
- $\bar{y}_i = (1/n_i)\sum_{j=1}^{n_i} y_{ij}$ průměr v i-té skupině
- $\bar{y} = (1/n)\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = (1/n)\sum_{i=1}^{k} n_i\bar{y}_i$ celkový průměr
- T Theilův index spočítaný ze všech n hodnot
  T_i Theilův index *uvnitř* i-té skupiny,
  T^B Theilův index variability *mezi* skupinami
  (jednotlivé hodnoty nahradíme dílčími průměry)

$$T_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{y_{ij}}{\bar{y}_i} \ln\left(\frac{y_{ij}}{\bar{y}_i}\right) \qquad T^B = \frac{1}{n}\sum_{i=1}^{k} n_i \frac{\bar{y}_i}{\bar{y}} \ln\left(\frac{\bar{y}_i}{\bar{y}}\right)$$

- platí $T = T^B + T^W$, kde T^W je vážený průměr T_i

$$T^W = \sum_{i=1}^{k} \frac{n_i\bar{y}_i}{n\bar{y}} T_i$$

příklad: příjmy 14 osob ve 3 skupinách

skupina	příjem y_i	n_i	průměr	T_i
A	200 150	2	175,00	0,010239
B	80 70 60 60	4	67,50	0,007421
C	20 20 18 18 15 15 10 10	8	15,75	0,030270
celk.	746	14	53,29	0,450220

$$T^B = \frac{1}{14}\left(2\cdot\frac{175{,}00}{53{,}29}\ln\left(\frac{175{,}00}{53{,}29}\right) + 4\cdot\frac{67{,}50}{53{,}29}\ln\left(\frac{67{,}50}{53{,}29}\right) + 8\cdot\frac{15{,}75}{53{,}29}\ln\left(\frac{15{,}75}{53{,}29}\right)\right) = 0{,}437618$$

$$T^W = \frac{350}{746}0{,}010239 + \frac{270}{746}0{,}007421 + \frac{126}{746}0{,}030270 = 0{,}012602$$

$T^B/T = 0{,}437618/0{,}450220 = 0{,}972$, tudíž nerovnoměrnost příjmů je z 97,2 % dána nerovnoměrností mezi skupinami

poznámky

- Theilův index T lze rozložit na součet dvou složek
  - index průměrů T^B, který charakterizuje nerovnoměrnost (různost) dílčích (skupinových) průměrů
  - index T^W, který charakterizuje průměrnou vnitřní nerovnoměrnost uvnitř skupin
- nutně platí nerovnost $T^B \leq T$, tj. nerovnoměrnost mezi skupinami nemůže být větší, než celková nerovnoměrnost
- podobně celkový Giniho koeficient nemůže být větší, než Giniho koeficient průměrů, ale jejich rozdíl nelze tak snadno vyjádřit, jako v případě Theilova indexu je T^W (vážený průměr dílčích indexů)

4. přednáška

- pravděpodobnost
- podmíněná pravděpodobnost
- náhodná veličina
- střední hodnota
- rozptyl
- nezávislost
- korelace

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
- **pravděpodobnost** by tedy měla mít **stejné hlavní vlastnosti** jako **relativní četnost**

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)

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

- **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 $\mathrm{P}(B) = 3/6 = 1/2$
- POZOR NA NESPRÁVNOU INTERPRETACI:
  - celkem stokrát hodíme kostkou ($n = 100$)
  - dvacetkrát padne šestka ($n_A = 20$)
  - poměr $\frac{n_A}{n} = \frac{20}{100} = 0{,}2$ je jen **odhad** pravděpodobnosti jevu A, že padne šestka, nikoliv pravděpodobnost sama

pomůcky k výpočtu pravděpodobnosti: 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

pomůcky k výpočtu pravděpodobnosti: 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)
- kolika způsoby mohu uložit pět knížek do knihovny? (120, záleží na pořadí!)

příklad: losování otázek (1)

- student *umí* 5 otázek, *neumí* 10 otázek z 15 možných
- losuje se dvojice otázek z oněch 15 otázek
- pravděpodobnost $\mathrm{P}(A)$, že student **nezná** ani jednu z vylosovaných:
- elementární jev: dvojice otázek
  první otázka – 15 možností, druhá jen 14 možností,
  ale 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 deseti, které neumí

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

příklad: losování otázek (2)

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

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

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

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

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

$$M_D = M_B + M_C = 50 + 10 = 60 \Rightarrow P(D) = \frac{60}{105} = 57{,}1\ \%$$

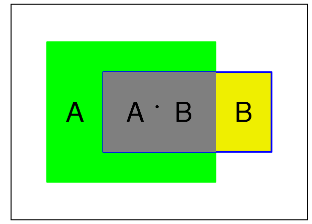
- kontrola: $M_D + M_A = M$

pravidla pro pravděpodobnost (1)

- **sjednocení** jevů $A \cup B$: platí A **nebo** B (**alespoň** 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 (plocha odpovídá pravděpodobnosti)



$A \cup B$ = 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

$$P(A) + P(\overline{A}) = 1$$

- Ω – **jev jistý** nastává vždy, $P(\Omega) = 1$
- $\emptyset$ – **jev nemožný** nenastává nikdy, je jevem opačným k jevu jistému, $P(\emptyset) = 0$
- **neslučitelné jevy**: nemohou nastat nikdy současně, navzájem se vylučují; jejich průnikem je jev nemožný; pro neslučitelné jevy platí

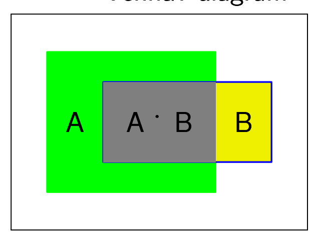
$$P(A \cup B) = P(A) + P(B)$$

podmíněná pravděpodobnost

- **podmíněná pravděpodobnost** pravděpodobnost jevu A, když už jev B nastal:

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

- Vennův diagram



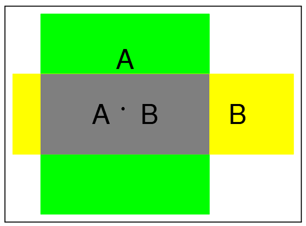
$P(B) = 0{,}24$ = žlutá + šedivá plocha
$P(A \cap B) = 0{,}16$ = šedivá plocha
$P(A|B)$ = šedivá vzhledem k (žlutá + šedivá)
$P(A|B) = 0{,}16/0{,}24 = 0{,}67$, ale
$P(A) = 0{,}42$
vyšlo $P(A|B) > P(A)$
jindy může vyjít také $P(A|B) < P(A)$

nezávislost náhodných jevů

- **nezávislé jevy**: výskyt jednoho jevu **neovlivní pravděpodobnost** výskytu druhého
- (definice **nezávislosti** náhodných jevů A, B, $\mathrm{P}(B) > 0$):

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

- 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, $\mathrm{P}(A \cap B) = 0{,}1$
- pravděpodobnost, že je aspoň jedna jednička:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) = 0{,}3 + 0{,}2 - 0{,}1 = 0{,}4$$

- 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: $\mathrm{P}(B|A) = 0{,}1/0{,}3 = 1/3$

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:

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

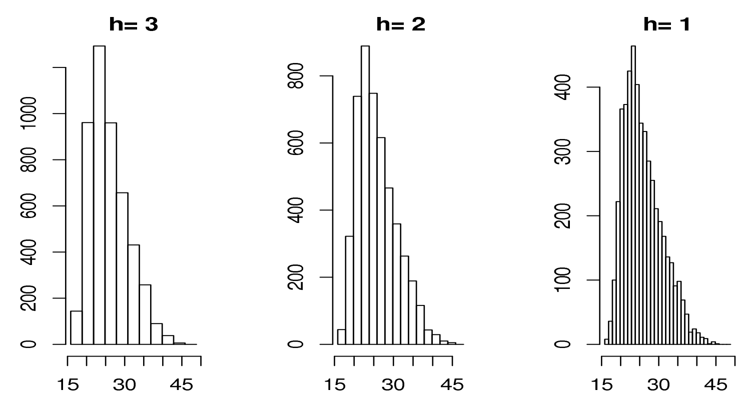
($F_X(x)$ je neklesající, zprava spojitá)
- **diskrétní rozdělení** (pro četnosti) určeno seznamem možných hodnot a jejich pravděpodobnostmi:

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

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

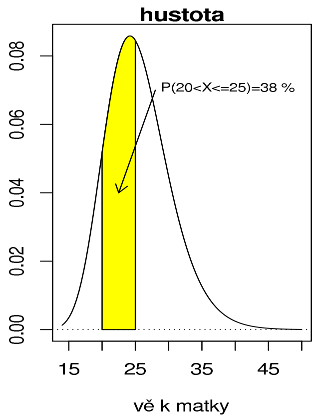
$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x), \qquad F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

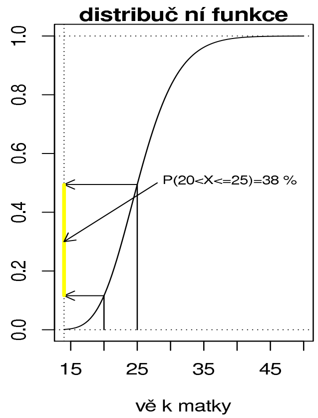
věk matek (n=4838)



představme si histogram založený na věku v hodinách a na rostoucím počtu matek, obálka bude docela hladká

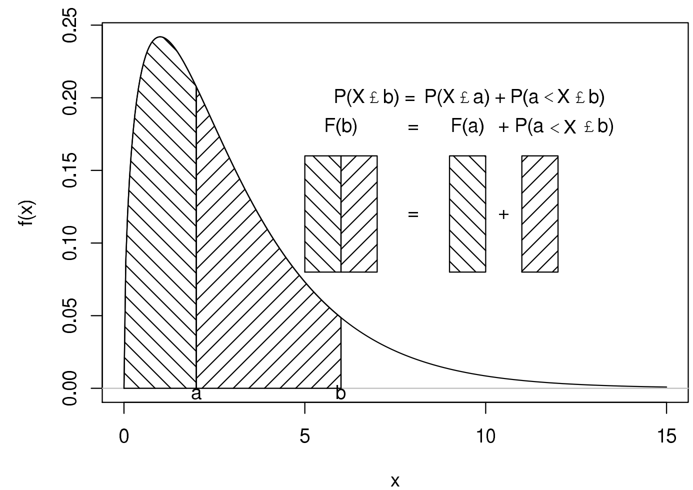
- velká populace, spojitá veličina – intervaly pro třídění mohou být krátké, obálce histogramu **relativních četností** odpovídá v idealizované představě **hustota** $f_X(x)$ [density]
- podobně **kumulativním relativním četnostem** odpovídá **distribuční funkce** [distribution function]





distribuční funkce v bodech $a < b$

jevy $X \leq a$ a $a < X \leq b$ jsou neslučitelné, jejich sjednocení dá jev $X \leq b$



použití distribuční fce (obecně)

- $F(x)$ je pravděpodobnost, že náhodná veličina X je menší než x (nebo stejná): $F(x) = \mathrm{P}(X \leq x)$
- je-li $a < b$, pak náhodný jev $(X \leq a)$ je **podjev** náhodného jevu $(X \leq b)$, proto je pak

$$F(a) = \mathrm{P}(X \leq a) \leq \mathrm{P}(X \leq b) = F(b)$$

(distribuční funkce je **neklesající**)
- náhodný jev $(X \leq b)$ je sjednocení disjunktních jevů $(X \leq a)$ a $(a < X \leq b)$, proto platí

$$\mathrm{P}(X \leq b) = \mathrm{P}(X \leq a) + \mathrm{P}(a < X \leq b)$$

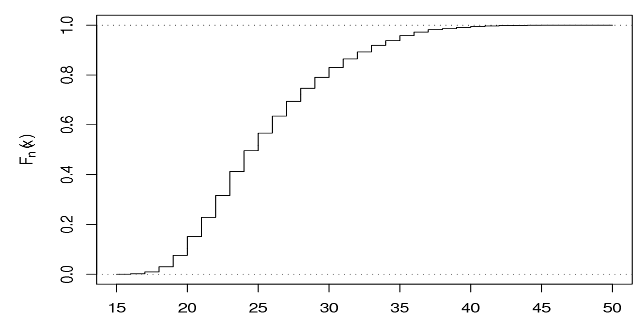
což je totéž, jako

$$F(b) = F(a) + \mathrm{P}(a < X \leq b)$$
$$\mathrm{P}(a < X \leq b) = F(a) - F(b)$$

- bezprostředním **výběrovým protějškem** distribuční funkce (jejím odhadem) je **empirická distribuční funkce**

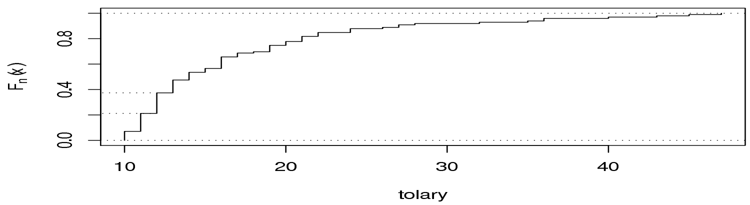
$$\boxed{F_n(x) = \frac{\#(x_i \leq x)}{n}}$$

- $x_1^* < x_2^* < \ldots < x_m^*$ 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: počty bodů na hrací kostece

- n-krát hodíme symetrickou kostkou
- počet bodů na symetrické kostce Y je náhodná veličina
- 1, 2, ..., 6 jsou možné hodnoty
- každá hodnota má pravděpodobnost 1/6
- $n = 100$, četnosti např. 13, 14, 15, 21, 14, 23
- $\bar{y} = (13\cdot 1 + 14\cdot 2 + 15\cdot 3 + 21\cdot 4 + 14\cdot 5 + 23\cdot 6)/100 \doteq 3{,}78$
- vážený průměr hodnot 1, 2, ..., 6, vahami jsou relativní četnosti 0,13, 0,14, 0,15, 0,21, 0,14, 0,23
- každá relativní četnost odhaduje pravděpodobnost 1/6
- nahraďme náhodné relativní četnosti odpovídajícími nenáhodnými pravděpodobnostmi
- dostaneme nenáhodnou **střední hodnotu** náh. veličiny Y

$$\mu_Y = \mathrm{E}\,Y = \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 2 + \frac{1}{6}\cdot 3 + \frac{1}{6}\cdot 4 + \frac{1}{6}\cdot 5 + \frac{1}{6}\cdot 5 = \frac{21}{6} = 3{,}5$$

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 od 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
- u diskrétního rozdělení jsou vahami 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

$$\begin{aligned}\sigma_X^2 &= \mathrm{E}(X-\mu_X)^2\\ &= (x_1-\mu_X)^2\,\mathrm{P}(X=x_1) + (x_2-\mu_X)^2\,\mathrm{P}(X=x_2) + \dots\\ &= \sum_j (x_j-\mu_X)^2\,\mathrm{P}(X=x_j)\end{aligned}$$

$$\sigma_X^2 = \mathrm{E}(X-\mu_X)^2 = \int_{-\infty}^{\infty}(x-\mu_X)^2 f_X(x)\mathrm{d}x$$

- **(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

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

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

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

(vzpomeň si: $(a+b)^2 = a^2 + 2ab + b^2$)

Návrat k rozptylu

$$\begin{aligned}\sigma_{XY} &= \mathrm{E}(X-\mu_X)(Y-\mu_Y)\ \textbf{kovariance}\ X, Y\\ &= (x_1-\mu_X)(y_1-\mu_Y)\,\mathrm{P}(X=x_1, Y=y_1)\\ &\quad + (x_1-\mu_X)(y_2-\mu_Y)\,\mathrm{P}(X=x_1, Y=y_2) + \dots\end{aligned}$$

(sčítá se přes všechny možné dvojice)

5. přednáška

- (populační) kovariance a korelace
- binomické rozdělení
- normální rozdělení
- populace a výběr
- CLV (centrální limitní věta)
- výběrový průměr
- interval spolehlivosti pro stř. hodnotu
- CLV pro četnosti
- interval spolehlivosti pro pravděpodobnost

příklad: známky ($V = X + Y$, $\sigma_V^2 \neq \sigma_X^2 + \sigma_Y^2$)

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

vlastnosti náhodné veličiny V:

$$\begin{aligned}\mu_V &= 2\cdot 0{,}10 + 3\cdot(0{,}10+0{,}10) + 4\cdot(0{,}10+0{,}15+0{,}00)\\ &\quad + 5\cdot(0{,}10+0{,}05) + 6\cdot(0{,}05+0{,}10) + 7\cdot(0{,}05) + 8\cdot 0{,}10\\ &= 4{,}6 = \mu_X + \mu_y = 2{,}1 + 2{,}5\\ \sigma_V^2 &= (2-4{,}6)^2\cdot 0{,}10 + (3-4{,}6)^2\cdot 0{,}20 + (4-4{,}6)^2\cdot 0{,}25\\ &\quad + (5-4{,}6)^2\cdot 0{,}15 + (6-4{,}6)^2\cdot 0{,}15 + (7-4{,}6)^2\cdot 0{,}05\\ &\quad + (8-4{,}6)^2\cdot 0{,}10 = 3{,}04 \neq \sigma_X^2 + \sigma_Y^2 = 0{,}89 + 1{,}05 = 1{,}94\end{aligned}$$

příklad: známky, výpočet kovariance

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

sdružené pravděpodobnosti

$$\begin{aligned}\sigma_{XY} &= (1-2{,}1)\cdot(1-2{,}5)\cdot 0{,}10 + (1-2{,}1)\cdot(2-2{,}5)\cdot 0{,}10\\ &\quad + (1-2{,}1)\cdot(3-2{,}5)\cdot 0{,}10 + (1-2{,}1)\cdot(4-2{,}5)\cdot 0{,}00\\ &\quad + \dots\\ &\quad + (4-2{,}1)\cdot(3-2{,}5)\cdot 0{,}00 + (4-2{,}1)\cdot(4-2{,}5)\cdot 0{,}10\\ &= 0{,}55\\ \sigma_V^2 &= 3{,}04 = 0{,}89 + 1{,}05 + 2\cdot 0{,}55 = \sigma_X^2 + \sigma_Y^2 + 2\cdot\sigma_{X,Y}\end{aligned}$$

modelové pojmy a jejich empirické protějšky

- pravděpodobnost $\mathrm{P}(A)$ vers. relativní četnost n_A/n
- střední hodnota μ_X vers. (výběrový) průměr $\bar{x}$
- (populační) rozptyl σ_X^2 vers. (výběrový) rozptyl s_x^2
- (populační) směrodatná odchylka σ_X vers. (výběrová) směrodatná odchylka
- (populační) kovariance σ_{XY} vers. (výběrová) kovariance s_{xy} (viz slajd 49)

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

- (populační) korelační koeficient $\rho_{X,Y}$ (teprve bude) vers. (výběrový) korelační koeficient r_{xy} (viz slajd 49)

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

nezávislost náhodných veličin

- ze **sdruženého** rozdělení $\mathrm{P}(X = x_i, Y = y_j$ vždy můžeme spočítat **marginální** rozdělení (viz např. slajd 101)

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

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

- 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

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

tj. z marginálních rozdělení lze obnovit sdružené rozdělení

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

příklad: známky jsou známky X a Y nezávislé?

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

zřejmě je například

$$\mathrm{P}(X = 1, Y = 1) = 0{,}10 \neq \mathrm{P}(X = 1) \cdot \mathrm{P}(Y = 1) = 0{,}3 \cdot 0{,}2 = 0{,}06$$

nebo

$$\mathrm{P}(X = 3, Y = 1) = 0{,}00 \neq \mathrm{P}(X = 3) \cdot \mathrm{P}(Y = 1) = 0{,}2 \cdot 0{,}2 = 0{,}04$$

náhodné veličiny X, Y **nemohou být nezávislé**, proto jsou **závislé**

populační korelační koeficient

- kovariance σ_{XY} je modelovým protějškem výb. kovariance s_{xy}
- podobně jako výběrový korelační koeficient r_{xy} je **populační korelační koeficient** ρ_{XY} definován pomocí kovariance σ_{XY} a směrodatných odchylek σ_X, σ_Y

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \qquad \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

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

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{0{,}55}{0{,}943 \cdot 1{,}025} \doteq 0{,}569$$

známky jsou nutně závislé, neboť $\rho_{XY} \neq 0$

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**

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

binomické rozdělení bi(n, π)

- zapisujeme $Y \sim \mathrm{bi}(n, \pi)$
- diskrétní rozdělení s parametry n, π $\quad (0 < \pi < 1)$
- model binomického rozdělení (důležité)
  - n **nezávislých** pokusů
  - v každém zdar s pravděpodobností π, nezdar s pstí $1 - \pi$
  - **celk. počet zdarů** Y má binomické rozdělení s parametry n, π
- Y je součet n nezávislých náhodných veličin $X_1, X_2, \dots, X_n$
  (X_i = počet zdarů v i-tém pokusu)
  každé X_i má alternativní rozdělení s parametrem π
- z vlastnosti střední hodnoty součtu náh. veličin: $\boxed{\mu_Y = n\pi}$
- z vlastnosti rozptylu součtu **nezávislých** náhodných veličin

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

binomické rozdělení bi(n, π)

- pravděpodobnosti možných hodnot `dbinom(k,n,p)`

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

- 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: 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, Y – počet kuřáků mezi nimi, tedy $Y \sim \mathrm{bi}(60, 0{,}35)$
- populační průměr, rozptyl (směrodatná odchylka):

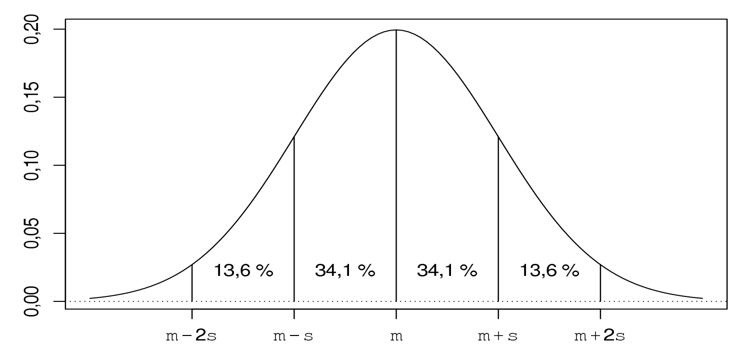
$$\mu_Y = 60 \cdot 0{,}35 = 21, \qquad \sigma_Y^2 = 60 \cdot 0{,}35 \cdot 0{,}65 = 13{,}65 = (3{,}7)^2$$

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

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

normální (Gaussovo) rozdělení N(μ, σ^2)

graf hustoty N(μ, σ^2) pro $\sigma = 2$



- 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

příklady pravděpodobností o normálním rozdělení

- pro $X \sim \mathrm{N}(\mu, \sigma^2)$ platí
  $\boxed{\mu_X = \mathrm{E}\,X = \mu}$ $\quad \boxed{\sigma_X^2 = \mathrm{E}(X-\mu_X)^2 = \sigma^2}$
- $\boxed{X \sim \mathrm{N}(\mu, \sigma^2) \qquad \Rightarrow \qquad Z = \dfrac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)}$
  $\Phi(z) = \mathrm{P}(Z \le y)$ je distribuční funkce $\mathrm{N}(0,1)$
- $$\mathrm{P}\left(|Z| < c\right) = \mathrm{P}\left(\left|\frac{X-\mu}{\sigma}\right| < c\right) = \mathrm{P}\left(|X-\mu| < c\cdot\sigma\right)$$
- tedy

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

poznámky

- Y má **logaritmicko normální rozdělení**, když $\log(Y)$ má normální rozdělení (koncentrace, hustoty …)
- zajímavé kvantily: `qnorm(0.975)`

$$z(0{,}975) = 1{,}96 \text{ tj. } \mathrm{P}(Z > 1{,}96) = 2{,}5\ \%$$
$$z(0{,}975) = 1{,}96 \text{ tj. } \mathrm{P}(Z < -1{,}96) = 2{,}5\ \%$$
$$z(0{,}975) = 1{,}96 \text{ tj. } \mathrm{P}(|Z| > 1{,}96) = 5\ \%$$
$$z(0{,}995) = 2{,}58 \text{ tj. } \mathrm{P}(Z > 2{,}58) = 0{,}5\ \%$$
$$z(0{,}995) = 2{,}58 \text{ tj. } \mathrm{P}(Z < -2{,}58) = 0{,}5\ \%$$
$$z(0{,}995) = 2{,}58 \text{ tj. } \mathrm{P}(|Z| > 2{,}58) = 1\ \%$$
$$z(0{,}950) = 1{,}64 \text{ tj. } \mathrm{P}(Z > 1{,}64) = 5\ \%$$
$$z(0{,}950) = 1{,}64 \text{ tj. } \mathrm{P}(Z < -1{,}64) = 5\ \%$$
$$z(0{,}950) = 1{,}64 \text{ tj. } \mathrm{P}(|Z| > 1{,}64) = 10\ \%$$

příklad: mladí volejbalisti

- předpoklad: výška desetiletých chlapců: $\mathrm{N}(141{,}5, 7^2)$
- Jaký díl populace desetiletých chlapců má výšku aspoň 149 cm, aby o ně měl zájem trenér volejbalu?
  (předpokládáme, že výšku určujeme s přesností na centimetr)

$$\mathrm{P}(X > 148{,}5) = 1 - \mathrm{P}(X \le 148{,}5)$$
$$= 1 - \mathrm{P}\left(\frac{X-141{,}5}{7} \le \frac{148{,}5-141{,}5}{7}\right) = 1 - \mathrm{P}\left(\frac{X-141{,}5}{7} \le 1\right)$$
$$= 1 - \Phi(1) = 1 - 0{,}841 = 0{,}159$$

- Jaký díl této populace má výšku v rozmezí od 142 do 148 cm?

$$\mathrm{P}(141{,}5 < X < 148{,}5) = \mathrm{P}(X < 148{,}5) - \mathrm{P}(X < 141{,}5)$$
$$= 0{,}841 - 0{,}5 = 0{,}341$$

(Proč je $\mathrm{P}(X < 141{,}5) = 0{,}5$?)

6. přednáška

- populace a výběr
- reprezentativnost a rozsah výběru
- chování výběrového průměru
- centrální limitní věta
- interval spolehlivosti

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), tj. měření téže vlastnosti na různých objektech
  - 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

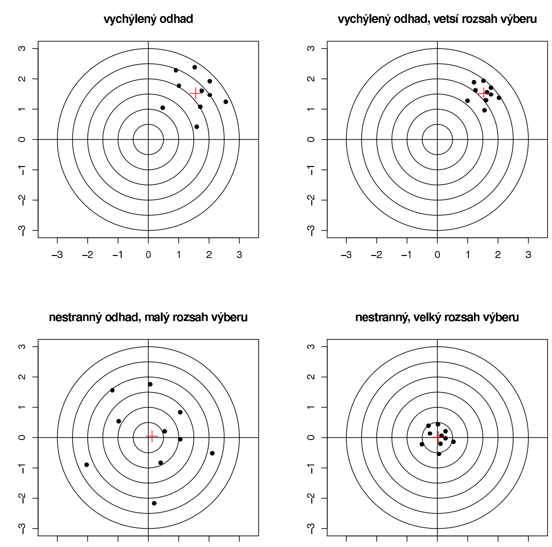
příklady

- volební preference
  - populace: všichni oprávnění voliči (jejich preferovaná strana)
  - výběr: respondenti, kteří odpověděli (odpovědi respondentů)
  - hodnocená vlastnost: podíly voličů jednotlivých stran
  - hodnocený parametr: $\pi_1, \dots, \pi_k$ psti jednotlivých stran
  - odhad parametru: relativní četnosti voličů jednotlivých stran ve výběru
- výšky desetiletých hochů
  - populace: všichni desetiletí hoši (jejich výšky)
  - výběr: změření desetiletí hoši (jejich výšky)
  - hodnocený znak: výška postavy náhodně vybraného chlapce
  - hodnocený parametr: **populační průměr** μ
  - odhad parametru: výběrový průměr, interval spolehlivosti pro μ
- **závěr:** při opakovaném pořízení výběru dostaneme po každé jiný odhad, odhad je tedy **náhodný**

vlastnosti výběru

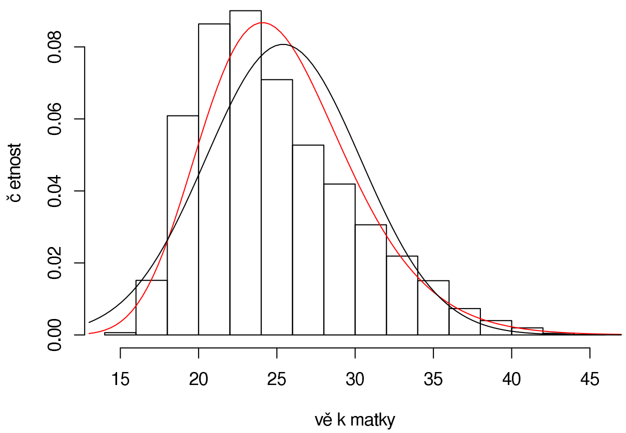
- **reprezentativnost výběru**
  - schopnost reprezentovat celou populaci
  - ve vlastnostech, které mohou souviset s daným šetřením, má složení výběru zhruba odpovídat složení populace
  - např. podíl žen, podíl vysokoškoláků, podíl důchodců ...
  - není-li výběr reprezentativní, jsou odhady vychýlené, nejsou nestranné, odhadují něco jiného, než chceme
  - např. reprezentační mužstvo jistě není reprezentativním výběrem organizovaných fotbalistů
- **rozsah výběru**
  - počet vyšetřovaných (např. dotazovaných) jednotek
  - ovlivní variabilitu odhadů, jejich kolísání
  - neovlivní reprezentativnost výběru či nestrannost odhadů
- reprezentativnost a rozsah výběru jsou různé vlastnosti
- dobrý střelec má všechny zásahy v terči blízko sebe (malá variabilita odhadu)
- pokud vzduchovka zanáší, i dobrý střelec střílí mimo střed terče (vychýlení)

vlastnosti výběru jako střelba do terče



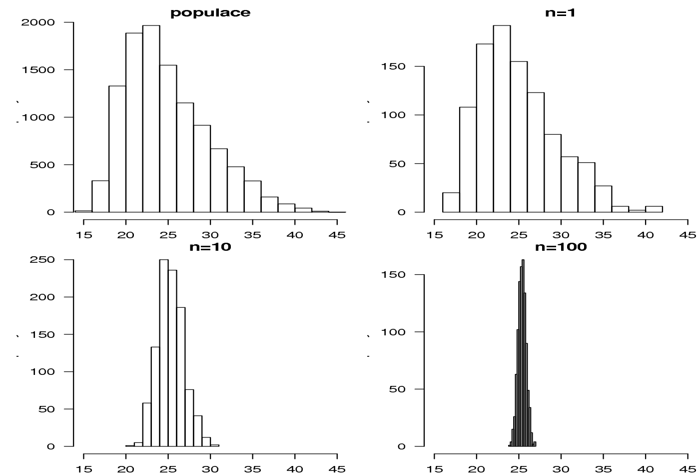
příklad: věk matek

černě hustota normálního rozdělení, červeně hustota logaritmickonormálního rozdělení



příklad: histogram populace a histogramy průměrů

šířky intervalů stejné, variabilita průměrů s rostoucím n klesá



chování výběrového průměru z náhodného výběru

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

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

- proto pro střední hodnotu výběrového průměru platí

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

- $\mathrm{E}\,\overline{X} = \mu_{\overline{X}} = \mu$, 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

$$\boxed{\mathrm{S.E.}(\overline{X}) = \frac{\sigma}{\sqrt{n}}} \qquad \text{Standard Error (of Mean, S.E.M.)}$$

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

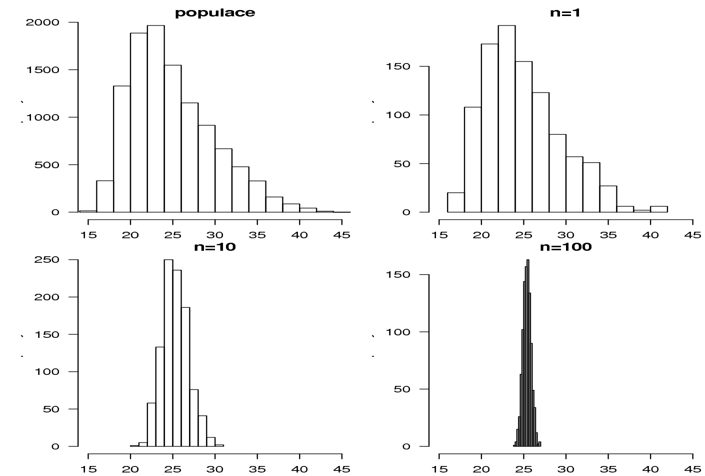
příklad: věk matek

- 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 B-krát opakujeme (spočítáno $B = 1000$ průměrů)
- spočítány charakteristiky z B průměrů jako výchozích hodnot, (modře charakteristiky celé populace nebo hodnoty z nich odvozené)

n	průměr	sm. odch.	$\sigma/\sqrt{n}$
1	25,43	4,62	4,94
10	25,35	1,54	1,56
100	25,39	0,48	0,49
populace	$\mu =$25,40	$\sigma = 4{,}94$	4,94

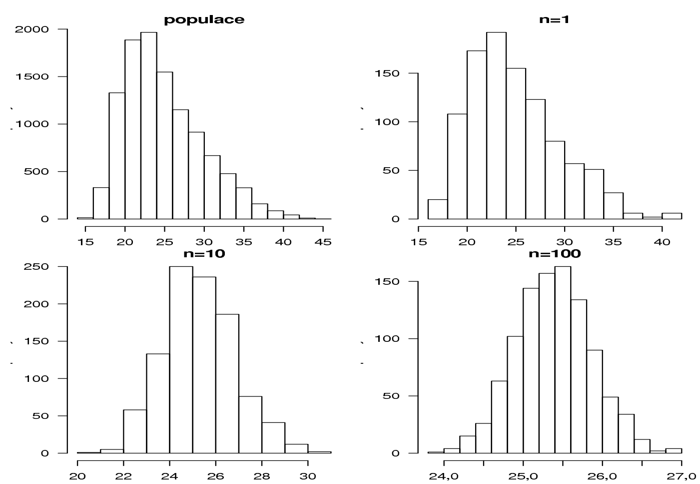
příklad: histogram populace a histogramy průměrů

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

- spočítány charakteristiky z $B = 1000$ 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

- průměry kolísají kolem populačního průměru μ
- směrodatné odchylky s rostoucí hodnotou $\sqrt{n}$ 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 N(0,1)

▸ CLV pro četnosti

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

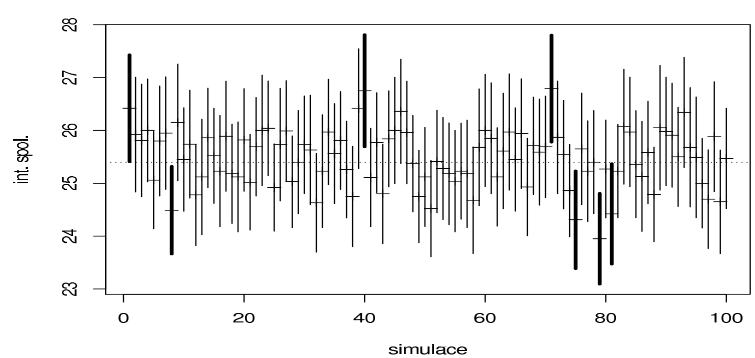
- pro velká n se výběrový průměr chová, jako by šlo o výběr z normálního rozdělení, a to bez ohledu na výchozí rozdělení

$$\boxed{\overline{X} \overset{\cdot}{\sim} \mathrm{N}(\mu,\sigma^2/n)}$$

interval spolehlivosti pro populační průměr μ

- pro nezávislé náhodné veličiny $X_1,\dots,X_n \sim \mathrm{N}(\mu,\sigma^2)$ platí
$$\overline{X} \sim \mathrm{N}(\mu,\sigma^2/n)$$
- proto je $$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$
- použijeme kvantil rozdělení N(0,1) (neboť $Z \sim \mathrm{N}(0,1)$)
$$\mathrm{P}\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| < z(1-\alpha/2)\right) = 1-\alpha$$
- hodnota neznámého parametru μ je s pravděpodobností $1-\alpha$ pokryta intervalem
$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}}z(1-\alpha/2); \overline{X} + \frac{\sigma}{\sqrt{n}}z(1-\alpha/2)\right)$$
- vzhledem k centrální limitní větě lze použít pro velká n i bez požadavku na normální rozdělení X_i

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



$\mu = 25{,}398$ je populační průměr věku 10 916 matek
v 7 případech interval **nepřekrývá** μ
(realizace náhodné veličiny s rozdělením bi(100, 0,05))

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 ($z(0{,}975) = 1{,}96$):
$$\left(110 - \frac{15}{4}\cdot 1{,}96; 110 + \frac{15}{4}\cdot 1{,}96\right) = (102{,}65; 117{,}35)$$
- 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?

$$\frac{\sigma}{\sqrt{n}} z(1-\alpha/2) = \delta \qquad \Rightarrow \qquad n = \left(\frac{\sigma}{\delta} z(1-\alpha/2)\right)^2$$

- 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 výběrové směr. odchylky

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

- interval spolehlivosti pro μ:

$$\boxed{\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha/2); \overline{X} + \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha/2)\right)}$$

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

příklad: výška postavy

- studenti odhadovali výšku přednášejícího; předpokládejme, že nestranně a nezávisle na sobě
- $n = 22, \qquad \overline{x} = 172{,}4, \qquad s_x = 4{,}032$
- z tabulek: $t_{21}(0{,}975) = 2{,}080$

$$(172{,}4 - \frac{4{,}032}{\sqrt{22}} \cdot 2{,}080; 172{,}4 + \frac{4{,}032}{\sqrt{22}} \cdot 2{,}080)$$
$$(170{,}6; 174{,}2)$$

- $\Rightarrow$ skutečná výška je s pravděpodobností 95 % někde mezi 170,6 cm a 174,2 cm
- podobně $(170{,}9; 173{,}9)$ je 90% interval spolehlivosti resp. $(170{,}0; 174{,}8)$ je 99% interval spolehlivosti
- Jak byste reagovali na tvrzení přednášejícího, že měří 174 cm?

centrální limitní věta pro četnosti

- co říkala CLV? ▸ CLV
- mějme n nezávislých opakování pokusu, kde sledovaný jev (zdar) nastane s pravděpodobností π
- absolutní četnost Y
  - Y – součet nezávislých veličin X_i 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 podle CLV pro velká n platí přibližně

$$Y \sim \mathrm{N}(n\pi, n\pi(1-\pi))$$

- relativní četnost $f = \frac{Y}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - 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 $\sigma_Y = 11{,}1$

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

- 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: $n = 100$, $Y_A = 17$, $f_A = 0{,}17$

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

(0,10; 0,24)

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

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

(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?

7. přednáška

- statistická indukce
- testování hypotéz
- p-hodnota
- test o podílu (o psti) π

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), tj. měření téže vlastnosti na různých objektech
  - 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

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

proč **testování hypotéz**

- V příkladu o IQ studentů jisté fakulty jsme dostali 95% interval spolehlivosti pro populační průměr tvaru (102,65; 117,35). Jak rozhodnout o pravdivosti tvrzení, že studenti této fakulty jsou inteligentnější, než běžná populace?
- V příkladu o výšce přednášejícího jsme dostali 95% interval spolehlivosti pro populační průměr tvaru (170,6; 174,2). Jak hodnotit prohlášení přednášejícího, že měří 174 cm?
- potřebujeme **standardizovaná pravidla**, jak rozhodovat
- nulová hypotéza – tvrzení o **populaci** (základním souboru)
- rozhodujeme na základě dat z **výběru**
- nelze zaručit bezchybnost rozhodnutí

hypotézy a možná rozhodnutí

- možné statistické **hypotézy**
  - **(nulová) hypotéza** H_0: – zjednodušuje situaci, porovnávané populace se **neliší**, vyšetřované znaky jsou **nezávislé** . . . (u soudu: obviněný je nevinný) tedy žádný (tj. **nulový**) rozdíl, žádná (tj. **nulová**) závislost
  - zpravidla se snažíme H_0 vyvrátit, abychom věcně něco prokázali (výjimkou je ověřování předpokladů či test dobré shody)
  - **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 (u soudu: obviněného odsoudit)
  - **nezamítnout** H_0 (přijmout H_0) pokud *není dost důvodů* H_0 zamítnout (není dost důvodů k odsouzení)

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, která skutečnost (který 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 (ještě pohodlnější bude: 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?

Franta si svoji kostku možná nějak vylepšil

- 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 v nich
- 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 $Y > y_0$
- 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.

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

příklad: jak zvolit kritickou hodotu y_0?

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

y_0	19	20	21	22	**23**	24
$\mathrm{P}_0(Y > y_0)$	0,220	0,152	0,100	0,063	**0,038**	0,022

- podmínku $\mathrm{P}_0(Y > y_0) \le 0{,}05 = \alpha$ splňuje $y_0 = 23$
- padne-li ve 100 nezávislých hodech kostkou více než 23 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 $y_0 = 21$, 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 $Y > 23$
- nechť je ve skutečnosti $\pi = 1/4$, pak hypotézu zamítneme (výsledek pokusu padne do kritického oboru) s pstí

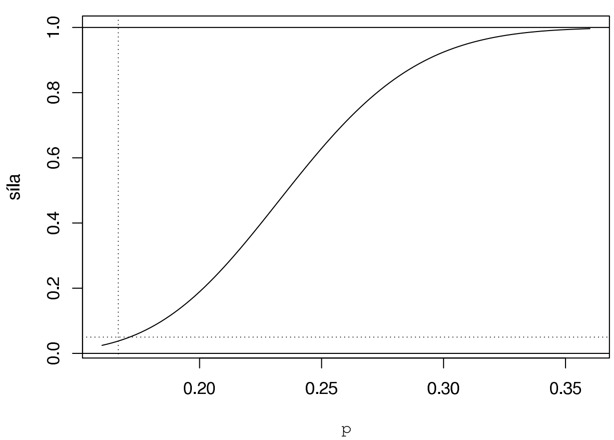
$$\mathrm{P}(Y > 23) = \sum_{k=24}^{100} \binom{100}{k} \left(\frac{1}{4}\right)^k \left(1 - \frac{1}{4}\right)^{100-k} = 0{,}629$$

`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 %

příklad: síla testu

závislost síly testu nulové hypotézy $\mathrm{H}_0 : \pi = 1/6$ proti jednostranné alternativě $\mathrm{H}_1 : \pi > 1/6$



rozhodování pomocí p-hodnoty

- ***p*-hodnota** p je nejmenší α, při kterém z daných dat nulovou hypotézu H_0 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 $\boxed{p \le \alpha}$
- 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
- nulovou hypotézu zamítnout, je-li $\boxed{p \le \alpha}$

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: $Y > y_0 = 23$
- na kostce A padlo 17 šestek, proto (psti binomického rozdělení)

$$p = P_0(Y \geq 17) = \sum_{k=17}^{100} \binom{100}{k} \left(\frac{1}{6}\right)^k \left(1 - \frac{1}{6}\right)^{100-k} = 0{,}506$$

$$= 1 - P_0(Y \leq 16)$$ `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 padlo 41 šestek

$$p = P_0(Y \geq 41) = 1 - P_0(Y \leq 40) = 7{,}4 \cdot 10^{-9}$$

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) ≠ 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	9	10	11	...	23	24	25
$P_0(Y < y_0)$	0,010	0,021	0,043	...	0,937	0,62	0,978
$P_0(Y > y_0)$	0,979	0,957	0,922	...	0,038	0,022	0,012
$P_0(Y = y_0)$	0,012	0,021	0,035	...	0,025	0,016	0,010

- $\alpha = 0{,}05$, tj. $\alpha/2 = 0{,}025$ (resp. $\alpha = 0{,}1$, tj. $\alpha/2 = 0{,}05$)
- H_0 zamítneme, když bude $Y < 10$ *nebo* když bude $Y > 24$
- 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)

- $H_0 : \pi = \pi_0$, např. P(padne šestka) = 1/6
- $H_1 : \pi \neq \pi_0$, např. P(padne šestka) ≠ 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: $|Z| \geq z(1 - \alpha/2)$
- pro $\alpha = 5$ % zamítáme hypotézu, je-li $|Z| \geq 1{,}96$
- $Y_A = 17$, $z_A = 0{,}089$, $p = 92{,}9\%$ (nezamítneme)
  95% int- spol. (0,10; 0,24) překrývá hodnotu $1/6 \doteq 0{,}167$
- $Y_B = 41$, $z_B = 6{,}529$, $p < 0{,}01\%$ (zamítneme)
  95% int- spol. (0,31; 0,51) nepřekrývá hodnotu $1/6 \doteq 0{,}167$

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 (výšky v cm): 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 skutečnost, že nový **výběrový** průměr je o 3 cm větší než hypotézou předpokládaný **populační** průměr?

test o střední hodnotě μ normálního rozdělení jednovýběrový t-test

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

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

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

souvislost s intervalem spolehlivosti

- připomeňme interval spolehlivosti pro μ

$$\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha/2) < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha/2)$$

$$\bar{X} - \widehat{\mathrm{S.E.}}(\bar{X}) \cdot t_{n-1}(1-\alpha/2) < \mu < \bar{X} + \widehat{\mathrm{S.E.}}(\bar{X}) \cdot t_{n-1}(1-\alpha/2)$$

- lze přepsat jako

$$|T| = \left| \frac{\bar{X} - \mu}{S}\sqrt{n} \right| < t_{n-1}(1-\alpha/2)$$

- $\mathrm{H}_0 : \mu = \mu_0$ 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

- máme $\mathrm{H}_0 : \mu = \mu_0 = 136{,}1$, $\mathrm{H}_1 : \mu > \mu_0 = 136{,}1$
- kritický obor: $\bar{X}$ se příliš liší od μ_0 ve směru zvolené alternativy
- spočítáme

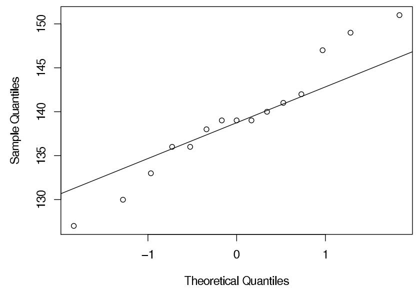
```
t.test(hosi,mu=136.1,alternative="greater")
```

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

- na 5% hladině při jednostranné alternativě $\mu > \mu_0$ hypotézu zamítáme, neboť $t_{14}(0{,}95) = 1{,}76$ ($p = 4{,}7$ %)
- na 5% hladině jsme **prokázali**, že výška desetiletých vzrostla
- 95% int. spolehlivosti pro populační průměr výšek hochů: jednostranný: $(136{,}2; \infty)$, oboustranný $(135{,}5; 142{,}8)$
- na 5% hladině při oboustranné alternativě bychom hypotézu nezamítli, neboť $t_{14}(0{,}975) = 2{,}14$ ($p = 9{,}5$ %)

ověření normality

- nulová hypotéza H_0 : data mají normální rozdělení
- graficky: představu dá **normální diagram** (probability plot): porovná ideální představu na ose x se skutečností na ose y
- Shapirův-Wilkův test hodnotí normální diagram




```
qqnorm(vysky);
qqline(vysky)

shapiro.test(vysky)
W = 0.9663
p-value = 0.7998
```

H_0 : data mají normální rozdělení jsme nezamítli ($p > 0{,}05$), normální rozdělení můžeme **předpokládat**

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

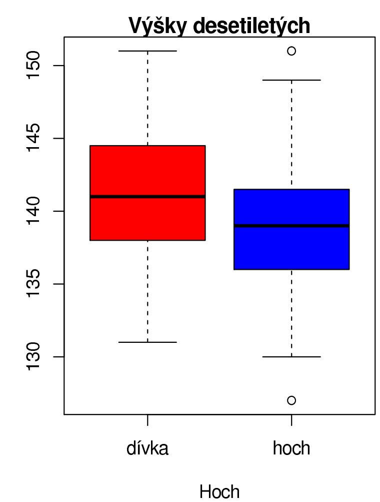
- 139,13 – 3,63 $= 135{,}50$
- 139,13 + 3,63 $= 142{,}76$
- 95% interval spolehlivosti: (135,5; 142,8)
- $\mu_0 = 136{,}1$ je v int. spolehlivosti
- při oboustranné alternativě jsme nezamítli H_0

8. přednáška

- dvouvýběrový t-test
- Mannův-Whitneyův test

porovnání dvou populací (dvouvýběrový t-test)

- příklad: liší se desetileté dívky výškou postavy od desetiletých hochů? (tvrzení o **všech** dětech)
- výběr 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
výšky desetiletých dětí

- 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

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

- μ_x (resp. μ_y) charakterizuje výšky všech chlapců (dívek)
- předpoklad stejných rozptylů bývá splněn, lze jej ověřit
- musí jít o **nezávislé** náhodné výběry, nelze např. vybírat sourozenecké dvojice nebo opakovaně měřit stejnou osobu

dvouvýběrový *t*-test

- $\mathrm{H}_0 : \mu_x = \mu_y$ (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
  - $\mathrm{H}_1 : \mu_x \neq \mu_y$ (není-li důvod k jednostranné alternativě)
  - $\mathrm{H}_1 : \mu_x > \mu_y$ (cílem dokázat, že hoši jsou větší než dívky)
  - $\mathrm{H}_1 : \mu_x < \mu_y$ (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

- 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 $\mathrm{H}_0 : \mu_x = \mu_y$ se rozhoduje pomocí

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

- $\mathrm{H}_1 : \mu_x \neq \mu_y$ zamítáme pokud $|T| \geq t_{n_x+n_y-2}(1 - \alpha/2)$
- $\mathrm{H}_1 : \mu_x > \mu_y$ zamítáme pokud $T \geq t_{n_x+n_y-2}(1 - \alpha)$
- $\mathrm{H}_1 : \mu_x < \mu_y$ zamítáme pokud $T \leq -t_{n_x+n_y-2}(1 - \alpha)$
- výšky desetiletých: $t = -0{,}70 \Rightarrow$
  $|-0{,}70| < 2{,}06 = t_{15+12-2}(0{,}975)$
- na 5% hladině jsme **neprokázali** rozdíl mezi výškami desetiletých hochů a dívek ($p = 48{,}8$ %)

`t.test(vyska~Hoch,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

$$(\bar{X} - \bar{Y}) \mp \widehat{\mathrm{S.E.}}(\bar{X} - \bar{Y}) \cdot t_{n_x+n_y-2}(1-\alpha/2)$$

$H_0 : \delta = 0$ (tj. $\mu_x = \mu_y$) zamítáme právě tehdy, když nula **není** v int. spol. pro δ

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

$$\left(-0{,}7 - 6{,}24\sqrt{\frac{1}{15}+\frac{1}{12}} \cdot 2{,}06;\ -0{,}7 + 6{,}24\sqrt{\frac{1}{15}+\frac{1}{12}} \cdot 2{,}06\right)$$

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

- nula **je** v intervalu, proto **nezamítáme** $H_0 : \delta = 0$

shrnutí

- důležité předpoklady
  - nezávislé výběry (dáno způsobem sběru dat)
  - stejné (populační) rozptyly (lze testovat)
  - normální rozdělení (lze testovat)
- existuje varianta bez předpokladu stejných rozptylů (Welch) (bude následovat)

`t.test(vyska~Hoch,var.equal=FALSE)`

`TTEST(A14:A28;A2:A13;2;3)`

- pro velká n_x, n_y na normalitě tolik nezáleží (CLV)
- je-li problém s normalitou, lze použít jiný test (dvouvýběrový Wilcoxonův, též Mannův-Whittneyův)

problém nestejných rozptylů

- předpoklad o stejném rozptylu v obou souborech nemusí být ve skutečnosti splněn, lze jej ověřit porovnáním odhadů rozptylu F-testem založeným na $F = \frac{S_x^2}{S_y^2}$
- hypotéza $H_0 : \sigma_x^2 = \sigma_y^2$ se proti $H_1 : \sigma_x^2 \neq \sigma_y^2$ zamítá, když je

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

- vlastně se větší odhad rozptylu dělí menším odhadem, k tomu se musí zvolit odpovídající pořadí stupňů volnosti a hladina
- příklad výšky desetiletých dětí: $F = \frac{42{,}98}{33{,}79} = 1{,}27 < F_{14,11}(0{,}975) = 3{,}36$
- `var.test(vyska~Hoch)` `FTEST()` (???)

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 $F_{14,11}(0{,}975) = 3{,}359$

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	
$t_f(1-\alpha)$	t krit (1)	1,708	
p oboustr. testu	P(T<=t) (2)	0,482	
$t_f(1-\alpha/2)$	t krit (2)	2,060	

při oboustranné alternativě nelze nulovou hypotézu zamítnout

dvouvýběrový Wilcoxonův test (někdy nepřesně Mannův-Whitneyův)

pořadová obdoba dvouvýběrového *t*-testu

- porovnáváme stejný kvantitativní (spojitý) 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: kostky cukru

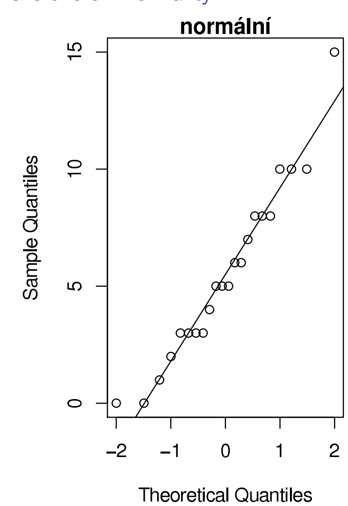
- porovnáváme spotřebu cukru u 43letých mužů s rizikovým faktorem pro ICHS se spotřebou cukru 43letých mužů bez tohoto faktoru
- spotřebu (hrubě) měříme udaným počtem kostek cukru běžně spotřebovaných za den
- dvouvýběrový *t*-test potřebuje normální rozdělení
- Shapirův-Wilkův test:

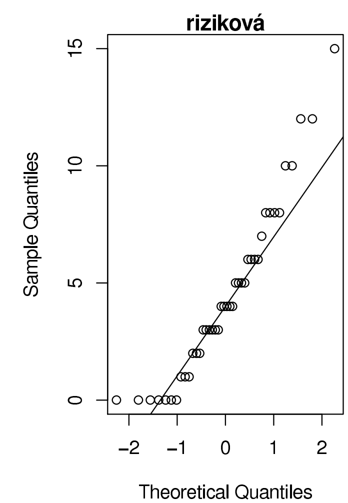
```
data: normální W = 0.9557, p-value = 0.4074
data: riziková W = 0.9246, p-value = 0.008537
```

- dvouvýběrový *t*-test není vhodný, předpoklad o normálním rozdělení není udržitelný

příklad: kostky cukru

grafické ověření normality





příklad: kostky cukru

- dvě skupiny mužů středního věku (riziková, normální)
- spojitý znak měříme hrubě počtem kostek cukru na den

kostek	norm.	rizik.	celkem	pořadí	R(norm.)	R(rizik.)
0	2	7	9	5,0	10,0	35,0
1	1	3	4	11,5	11,5	34,5
2	1	3	4	15,5	15,5	46,5
3	4	6	10	22,5	90,0	135,0
4	1	5	6	30,5	30,5	152,5
5	3	4	7	37,0	111,0	148,0
6	2	4	6	43,5	87,0	174,0
7	1	1	2	47,5	47,5	47,5
8	3	4	7	52,0	156,0	208,0
10	3	2	5	58,0	174,0	116,0
12	0	2	2	61,5	0,0	123,0
15	1	1	2	63,5	63,5	63,5
celkem	22	42	64		796,5	1283,5

dvouvýběrový Wilcoxonův test

- normální: $n_x = 22$, součet pořadí $W_x = 796{,}5$ (průměr $\frac{796{,}5}{22} = 36{,}2$)
- riziková: $n_y = 42$, součet pořadí $W_y = 1283{,}5$ (průměr $\frac{1283{,}5}{42} = 30{,}50$)
- `wilcox.test(normální,riziková)`
  `W = 543.5, p-value = 0.2495`
- statistika W má jiný význam (Mannův-Whitneyův test)
  - uvažuje všechny dvojice mužů (normální, riziková), tj. (X_i, Y_j), $i = 1, \dots, n_x, \quad j = 1, \dots, n_y$
  - není-li mezi dvěma populacemi, dokud pocházejí výběry (platí-li H_0) rozdíl, mělo by asi v polovině případů být $X_i > Y_j$ a v polovině případů $X_i < Y_j$
  - statistika W udává počet dvojic $X_i > Y_j$ zvětšený o polovinu počtu dvojic $X_i = Y_j$
- rozhodnutí dvouvýběrového Wilcoxonova testu a Mannova-Whitneyova testu jsou totožná

příklad: IQ dvojčat (**párový t-test**)

datový soubor twins v balíčku alr3

rodiče	82	90	91	115	115	129	131
pěstouni	82	80	88	108	116	117	132
rozdíl	0	10	3	7	−1	12	-1

- dvojice jednovaječných dvojčat vychovávaných odděleně
- Závisí IQ na způsobu výchovy?
- **párová závislost**, nejde o dvouvýběrový test
- $H_0 : \mu_r = \mu_p$, tj. $\mu_{\text{rozdíl}} = 0$
- jednovýběrový t-test použijeme na rozdíly
- $t = 2{,}097$, $p = 0{,}081$
- rozdíl jsme neprokázali
- 95 % int. spol. pro rozdíl $\mu_r = \mu_p$: $(-0{,}7; 9{,}3)$
- použití jednovýběrového t-testu na rozdíl párově (možná) závislých pozorování se nazývá **párový t-test**
- předpokládá se, že **rozdíly** jsou nezávislé a **mají normální rozdělení**

příklad: IQ dvojčat (**znaménkový test**)

rodiče	82	90	91	115	115	129	131
pěstouni	82	80	88	108	116	117	132
rozdíl	0	10	3	7	−1	12	-1

- čtyři hodnoty z šesti nenulových jsou kladné (znaménko nuly nemá smysl)
- Kdyby nezáleželo na tom, kdo dvojče vychovává, byla by pravděpodobnost kladného znaménka rovna 1/2, použijeme test o pravděpodobnosti v binomickém rozdělení.
- (slajd 156:)
  $$Z = \frac{Y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{f - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n}$$
- $$z = \frac{4 - 6/2}{\sqrt{6/4}} = 0{,}816, \qquad p = 0{,}414$$
- `binom.test(4,6,p=0.5)` je založen na přesném testu: $p = 0{,}6875$

znaménkový test obecně

- $X_1, X_2, \dots, X_n$ nezávislé, stejně rozdělené, spojité rozdělení
- $\mathrm{H}_0 : \mathrm{P}(X_i \leq x_0) = 1/2$, tj. x_0 je populační medián
- vyškrtněme pozorování shodná s x_0, upravme n
- označme Y počet hodnot X_i menších než x_0, platí-li H_0, pak $Y \sim \mathrm{bi}(n, 1/2)$
- rozhodneme o H_0 v binomickém rozdělení $\mathrm{bi}(n, \pi)$, podle které je $\pi = 0{,}5$
- u párově závislých pozorování použijeme jako X_i rozdíly hodnot ve dvojicích

jednovýběrový (párový) Wilcoxonův test

(Wilcoxon signed rank test)

- $X_1, X_2, \dots, X_n$ nezávislé, stejně rozdělené, spojité rozdělení, které je **symetrické** kolem x_0
- $\mathrm{H}_0 : x_0 = a$, kde a je daná hodnota, nejčastěji nula
- podobně jako u znaménkového testu se vyloučí hodnoty $X_i = a$, upraví se n
- určí se pořadí R_i^+ absolutních hodnot $|X_i - a$, sečtou se jen ta, kde je $X_i > a$

$$W = \sum_{i: X_i > a} R_i^+$$

- rozhoduje se podle W
  `wilcox.test(rodiče,pěstouni,paired=TRUE)`

příklad: IQ dvojčat (**znaménkový test**)

rodiče	82	90	91	115	115	129	131
pěstouni	82	80	88	108	116	117	132
rozdíl	0	10	3	7	−1	12	−1
|rozdíl|	—	10	3	7	1	12	1
R_i^+	—	5	3	4	1,5	6	1,5

- $W = 5 + 3 + 4 + 6 = 18 \qquad p = 14{,}1\ \%$
- `wilcox.test(rodiče,pěstouni,paired=TRUE)`
- kdybychom předpokládali normální rozdělení, použili bychom `t.test(rodiče,pěstouni,paired=TRUE)`

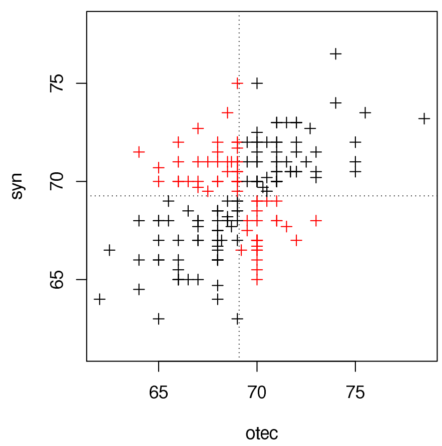
9. přednáška

- (Pearsonův) korelační koeficient
- Spearmanův korelační koeficient
- regrese
- metoda nejmenších čtverců
- testy o regresní přímce
- ověření předpokladů

příklad: výška otce a syna

data: Galton (1886), Hanley (2004), údaje v angl. palcích (1 palec = 25,4 mm)

souvisí spolu výška otce a výška jeho dospělého syna? $r = 0{,}505$



$$r_{xy} = \frac{\sum_{i=1}^n (X_i-\bar X)(Y_i-\bar Y)}{\sqrt{\sum_{i=1}^n (X_i-\bar X)^2 \sum_{i=1}^n (Y_i-\bar Y)^2}} = \frac{1}{n-1}\sum_{i=1}^n \left(\frac{X_i-\bar X}{S_X}\right)\left(\frac{Y_i-\bar Y}{S_Y}\right)$$

prokazování závislosti spojitých veličin

- považujme dvojice výšek X, Y (otec,syn) za náhodný výběr z populace dvojic (otec,syn)
- H_0 : náhodné veličiny X, Y jsou nezávislé
- víme, že pro nezávislé X, Y je $\rho_{XY} = 0$
- r_{xy} je odhadem ρ_{XY}; 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 $|T| \geq t_{n-2}(1-\alpha/2)$, kde

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

- pro (otec,syn) vyšlo $t = 7{,}65$, $p \doteq 10^{-12}$, závislost jsme prokázali (normalitu výšek lze předpokládat)

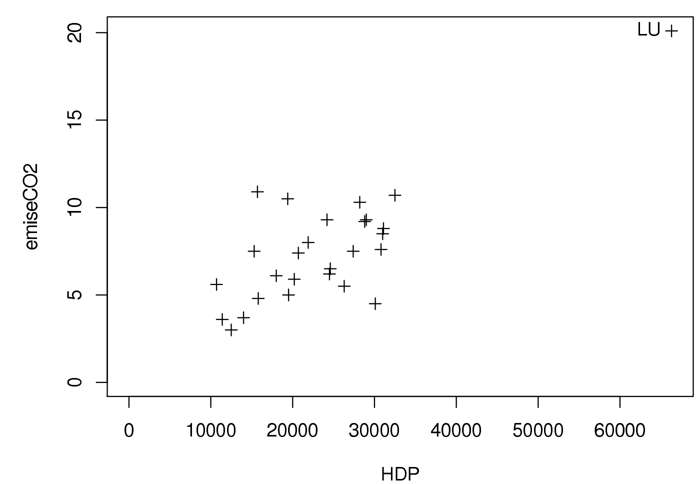
Spearmanův korelační koeficient

- nelze-li předpokládat normalitu (nebo nejsou stovky dvojic), použijeme **Spearmanův korelační koeficient** r_S
- místo původních X_i, Y_i použíjeme jejich pořadí R_i, Q_i
- r_S je vlastně Pearsonův korelační koeficient použitý na pořadí
- výpočet lze upravit (zjednodušit) na

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

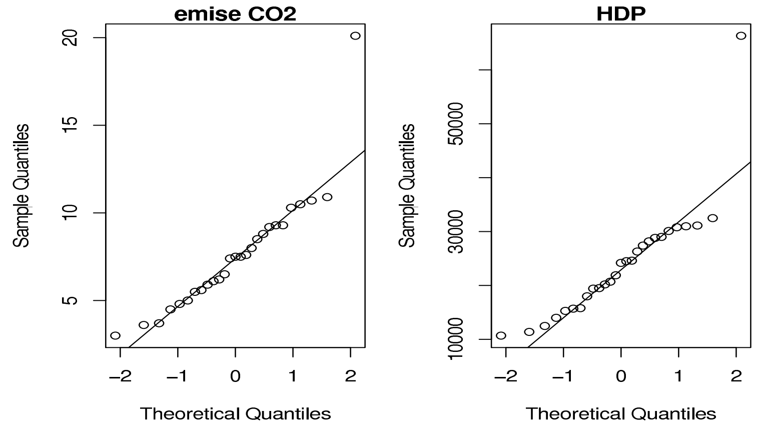
- vhodný také pro nelineární monotonní **závislost**
- nevadí odlehlé hodnoty
- při testování nemusí být normální rozdělení
- nezávislost se zamítá, je-li $|r_S|\sqrt{n-1} \geq z(\alpha/2)$ (pro n velké), jinak s využitím tabulek

příklad: emise CO_2 a HDP v EU



příklad: emise CO_2 a HDP v EU

ověření normálního rozdělení CO_2 a HDP, Shapirův-Wilksův test




```
W = 0.8513, p-value = 0.001224
W = 0.7936, p-value = 0.0001061
```


příklad: emise CO_2 a HDP v EU

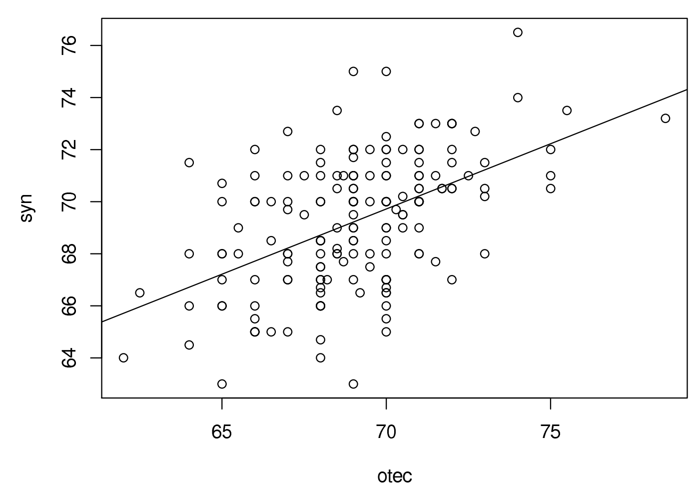
- Spearmanův korelační koeficient:
  `cor.test(~emiseCO2+HDP,data=EU2010,method="spearman")`
  $r_S = 0{,}549$, což při $n = 27$ vede k $p = 0{,}003$
- nesprávně použitý (Pearsonův) korelační koeficient:
  `cor.test(~emiseCO2+HDP,data=EU2010)`
  $r_S = 0{,}795$, což při $n = 27$ vede k $p < 0{,}001$
- když vynecháme **odlehlou hodnotu** Lucemburska:
  `cor.test(~emiseCO2+HDP,data=EU2010,subset=země!="LU")`
  dostaneme $r = 0{,}516$, což při $n = 26$ vede k $p = 0{,}007$
- po vyloučení dat o Lucembursku normalitu nezamítneme
- u podobných dat (charakteristiky států, územních jednotek) je předpoklad normálního rozdělení zpravidla nepoužitelný

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 výška syna s výškou otce?

upravená Galtonova data, údaje v palcích



regresní přímka

- cíl: chování Y (výška syna) co nejlépe (nejvíce) vysvětlit lineární závislostí na x (výška otce)
- (naše představa, předpoklad:) každé výšce otce x_i odpovídá jakási střední výška syna E Y_i, ta závisí na výšce otce lineárně

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

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

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

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

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

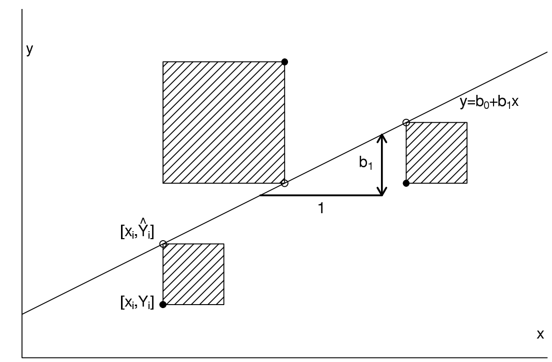
- výsledné minimum (pro $\beta_0 = b_0, \beta_1 = b_1$) se nazývá **reziduální součet čtverců**: $S_e = \sum_{i=1}^{n}(Y_i - b_0 - b_1 x_i)^2$

metoda nejmenších čtverců

na obrázku jsou pouze tři pozorování!

směrnicový tvar rovnice přímky – návod, jak k x spočítat y (souřadnice bodu na přímce)

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(syn~otec,data=GaltonSyn))
```

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	34,652	4,527	7,654	<0,001
otec	0,501	0,065	7,651	<0,001

- odhad závislosti: $\widehat{\text{syn}} = 34{,}652 + 0{,}501$ otec
- s každým palcem výšky otce roste výška syna v průměru zhruba o půl palce
- jiné vyjádření: $(\widehat{\text{syn}} - \overline{\text{syn}}) = 0{,}501\,(\text{otec} - \overline{\text{otec}})$
- Vezměme otce, jejichž výška se liší o jeden palec od průměrné výšky otců. Očekáváme, že průměrná výška jejich synů se od průměrné výšky synů bude lišit jen o půl palce.
- odchylka od průměru se reprodukuje jen z poloviny (regrese k průměru)
- závislost je průkazná, neboť v řádku pro x (otec) 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

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

- zamítáme H_0 proti oboustranné alternativě, když $|T| \geq t_{n-2}(1 - \alpha/2)$
- regresní přímka prochází těžištěm $(\bar{x}; \bar{Y})$
- vyrovnané (vyhlazené) hodnoty:

$$\hat{Y}_i = b_0 + b_1 x_i = \bar{Y} + b_1(x_i - \bar{x})$$

rezidua

- **rezidua**

$$u_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i) = (Y_i - \bar{Y}) - b_1(x_i - \bar{x})$$

- **reziduální součet čtverců**: nevysvětlená variabilita Y

$$S_e = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 x_i))^2$$

- **reziduální rozptyl**: odhad σ^2

$$s^2 = S_e/(n-2)$$

- **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(syn~otec,data=GaltonSyn))
```

variabilita	st. vol. f	součet čtverců SS	prům. čtverec MS	F	p
model	1	279,11	279,107	58,532	<0,001
reziduální	171	815,41	4,768		
celkem	172	1 094,52			

- kolísání výšek synů vysvětlíme závislostí z 25,5 %, neboť je

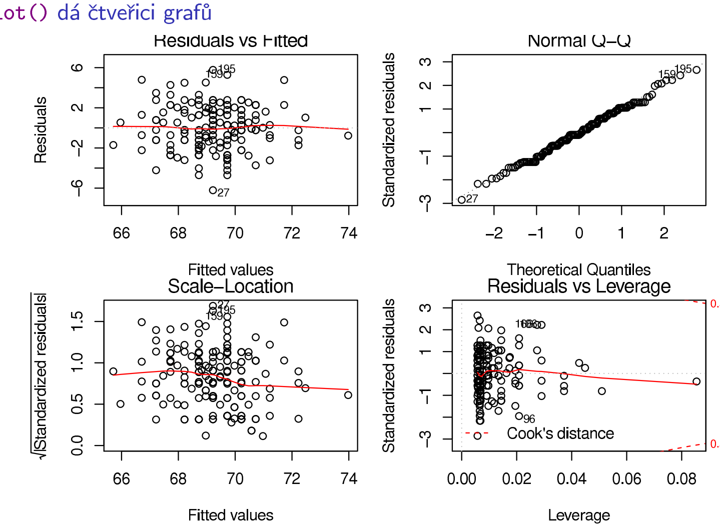
$$R^2 = 1 - \frac{815{,}41}{1094{,}52} = \frac{279{,}11}{1094{,}5} = 0{,}255$$

ověření splnění předpokladů

- ověření **nelze provést** jen hodnocením hodnot závisle proměnné $Y_1, \ldots, Y_n$ (např. nemají stejné střední hodnoty)
- využívají se zejména rezidua $u_i = Y_i - \hat{Y}_i$ (znaménkem opatřené svislé vzdálenosti pozorování od přímky)
- rychlé předběžné grafické ověření pomocí funkce `plot(a)`, kde je `a = lm(Y~x)`
- ověření normality: Shapirův-Wilkův test použitý na rezidua `shapiro.test(rstandard(a))`
- stabilita rozptylu: Breuschův-Paganův test `library(lmtest); bptest(a)`
- nezávislost **po sobě jdoucích** pozorování: Durbinův-Watsonův test (např. časová řada má pozorování navzájem závislá) `library(lmtest); dwtest(a)`

příklad: výška syna a otce

`plot()` dá čtveřici grafů

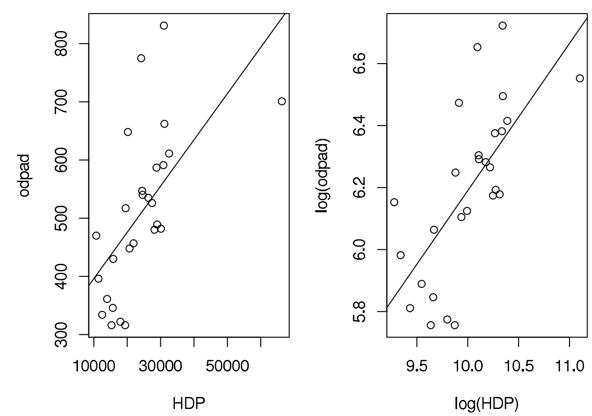


příklad: výška syna a otce

- žádný graf nenaznačuje problém
- test normality: `W = 0.9935, p-value = 0.6445`
- stálost rozptylu: `BP = 1.6422, df = 1, p-value = 0.2`
- ani jeden test nenaznačuje problém, můžeme předpokládat normální rozdělení i konstantní rozptyl
- Durbinův-Watsonův test nemá smysl, protože pořadí napozorovaných hodnot je nahodilé, není v něm systém.

příklad: závislost produkce odpadu na HDP v EÚ

odpad v kg na osobu za rok, HDP jednotkách PPS

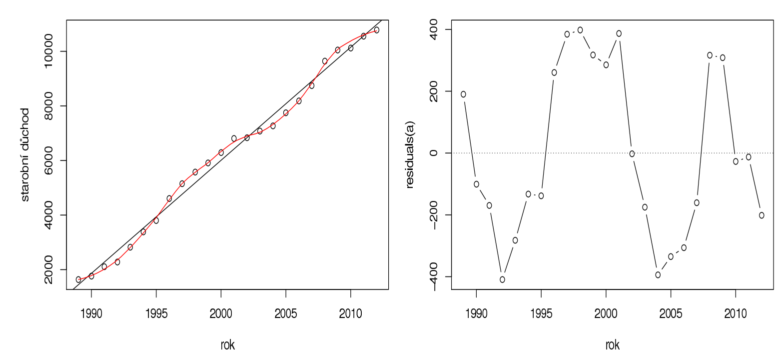


závislost v logaritmech je nepochybně lépe vystižena přímkou

poznámky

- **transformace** mnohdy je užitečné modelovat závislost vhodné funkce závisle proměnné na vhodné (třeba jiné) funkci nezávisle proměnné
- v případě jediné nezávisle proměnné x je koeficient determinace R^2 čtvercem korelačního koeficientu r_{xy}
- **mnohonásobná regrese**: nezávisle proměnných může být více, např. výšku syna lze vysvětlovat výškou otce a výškou matky současně
- pokud data tvoří časovou řadu, předpoklad nezávislosti pozorování mezi sebou nebývá splněn (např. vývoj HDP)

příklad: vývoj starobních důchodů v ČR



na grafu patrné kolísání kolem přímky, graf reziduí to jen zdůrazňuje: je nutno použít složitější model, který přihlédne k závislosti a zřejmé periodicitě

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!

10. přednáška

- multinomické rozdělení
- chí-kvadrát test dobré shody
- kontingenční tabulka
- čtyřpolní tabulka

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 $Y_1, \dots, Y_k$
- parametry $n, \pi_1, \dots, \pi_k$ $(0 < \pi_j < 1, \quad \pi_1 + \dots + \pi_k = 1)$
- 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
- Y_j – počet pokusů, v nichž nastal j-tý možný výsledek, tedy nutně

$$Y_1 + \dots + Y_k = n$$

příklady multinomického rozdělení

- předvolební průzkum
  - n – počet tázaných
  - π_j – skutečný podíl voličů j-té strany v populaci
  - Y_j – počet (četnost) voličů j-té strany ve výběru
- hody hrací kostkou
  - n – počet hodů
  - $\pi_1, \dots, \pi_6$ – pravděpodobnosti jednotlivých stran kostky
  - $Y_1, \dots, Y_6$ – absolutní četnosti jednotlivých stran kostky
- krevní skupiny
  - $k = 4$ (skupiny 0, A, B, AB)
  - $\pi_0, \pi_A, \pi_B, \pi_{AB}$ – pravděpodobnosti skupin 0, A, B, AB
  - Y_0, Y_A, Y_B, Y_{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 Y_j má binomické rozdělení:

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

- střední hodnota: $\mu_{Y_j} = n\pi_j$, rozptyl: $\sigma^2_{Y_j} = n\pi_j(1-\pi_j)$
- (pro zajímavost) kovariance: $\mathrm{cov}(Y_j, Y_t) = -n\pi_j\pi_t \quad j \neq t$
- náhodné veličiny $Y_1, \dots, Y_k$ jsou závislé ($Y_1 + \dots + Y_k = n$)
- asymptotická vlastnost **chí-kvadrát** (velká n, $n\pi_j \geq 5\ \forall j$)

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

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

příklad: hrací kostka A

- **chí-kvadrát test dobré shody**
- $n = 100$ hodů kostkou
- $Y_1 = 12,\ Y_2 = 21,\ Y_3 = 14,\ Y_4 = 15,\ Y_5 = 21,\ Y_6 = 17$
- hypotéza $H_0 : \pi_1 = \dots = \pi_6 = 1/6$ dá očekávané četnosti $n\pi_1 = \dots = n\pi_6 = 100/6 = 16{,}67$ (vždy více než 5))
- $$\chi^2 = \frac{(12 - 16{,}67)^2}{16{,}67} + \dots + \frac{(17 - 16{,}67)^2}{16{,}67} = 4{,}16$$
- $$\chi^2 < \chi^2_5(0{,}95) = 11{,}07, \qquad p = 52{,}7\ \%$$
- neprokázali jsme, že by kostka nebyla symetrická
- neprokázali jsme ani to, že je symetrická
- symetrii můžeme pouze **předpokládat**
- `chisq.test(c(12,21,14,15,21,17),p=rep(1,6)/6)`

příklad: hrací kostka B (1)

- $n = 100$ hodů kostkou
- $Y_1 = 15, Y_2 = 16, Y_3 = 7, Y_4 = 6, Y_5 = 15, Y_6 = 41$
- hypotéza $H_0 : \pi_1 = \dots = \pi_6 = 1/6$ dá očekávané četnosti $n\pi_1 = \dots = n\pi_6 = 100/6 = 16{,}67$

$$\chi^2 = \frac{(15 - 16{,}67)^2}{16{,}67} + \dots + \frac{(41 - 16{,}67)^2}{16{,}67} = 48{,}32$$

- $$\chi^2 > \chi^2_5(0{,}95) = 11{,}07 \qquad p < 0{,}0001$$
- 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
- $Y_1 = 15, Y_2 = 16, Y_3 = 7, Y_4 = 6, Y_5 = 15, Y_6 = 41$
- nulová hypotéza: $\pi_1 = \ldots = \pi_5 = 1/10, \pi_6 = 5/10 = 1/2$
- očekávané četnosti za hypotézy: $n\pi_1 = \ldots = n\pi_5 = 100/10 = 10, n\pi_6 = 100/2 = 50$

$$\chi^2 = \frac{(15-10)^2}{10} + \ldots + \frac{(15-10)^2}{10} + \frac{(41-50)^2}{50} = 12{,}72$$

- $$\chi^2 > \chi^2_5(0{,}95) = 11{,}07 \qquad p = 2{,}6\ \%$$
- 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
- $Y_6 = 41$
- 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% hladině **nelze** zamítnout

```
binom.test(41,100)
```


chí-kvadrát test dobré shody obecně

- $Y_1, \ldots, Y_k$ mají multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$, kde $Y_1 + \cdots + Y_k = n$ a $\pi_1 + \cdots + \pi_k = 1$
- nulová hypotéza tvrdí, že pravděpodobnosti jsou rovny známým hodnotám: $\pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$
- spočítáme očekávané četnosti $o_i = n\pi_i^0$
- ověříme splnění podmínky $o_i \geq 5$
- spočítáme statistiku chí-kvadrát:

$$\chi^2 = \sum_{i=1}^{k} \frac{(Y_i - o_i)^2}{o_i}$$

- nulovou hypotézu zamítáme na hladině významnosti α, je-li $\chi^2 \geq \chi^2_{k-1}(1-\alpha)$

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 testu dobré shody je $\chi^2 = 5{,}34$ ($p = 14{,}9$ %), ale $\chi^2_3(0{,}95) = 7{,}81$
- neprokázali jsme, že by výběr nebyl reprezentativní, můžeme jej za reprezentativní považovat

```
chisq.test(c(180,239,75,4),p=c(34.27,52.03,12.50,1.20)/10
```

příklad: vzdělání snoubenců

tabulka udává četnosti zjištěné u 100 náhodně vybraných snoubenců

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

- zajímá nás, zda vzdělání ženicha a nevěsty spolu souvisí
- lze považovat vzdělání snoubenců za nezávislá?
- tabulka udává sdružené a z nich spočítané marginální četnosti
- vzdělání zde chápeme jen v nominálním měřítku
- četnosti na diagonále převládají, četnosti mimo diagonály jsou spíše menší
- vzhledem k nominálnímu měřítku nelze použít ani Spearmanův korelační koeficient

kontingenční tabulka (závislost vzdělání snoubenců)

- opět použijeme chí-kvadrát test
- očekávané četnosti vycházejí z nulové hypotézy H_0 : vzdělání jsou nezávislá (vzdělání nevěst jsou mají pravděpodobnosti nezávislé na vzdělání ženichů)
- vzdělání nevěst jsou v poměru 34 % : 45 % : 21 %
- stejný poměr by měl být např. pro 27 ženichů vysokoškoláků: 9,18 (tj. 34 % z 27) : 12,15 (tj. 45 %) : 5,67 (tj. 21 %)
- to jsou četnosti očekávané za platnosti nulové hypotézy
- podobně dostaneme očekávané četnosti pro zbývající dvě kategorie ženichů
- statistika chí-kvadrát porovná skutečně zjištěné (empirické) četnosti s četnostmi za nezávislosti očekávanými, spočítá jejich „vzdálenost“

test nezávislosti v kont. tabulce

- u n jedinců (statistických jednotek) vyšetřujeme dva znaky v nominálním měřítku (vzdělání ženicha, vzdělání nevěsty), které mají r a c možných hodnot
- označme n_{ij} počet jedinců s i-tou hodnotou prvního znaku a j-tou hodnotou druhého znaku (např. n_{12} je počet dvojic snoubenců, kdy ženich má základní vzdělání a nevěsta střední)
- spočítáme řádkové marginální četnosti n_{i+} (počet ženichů s i-tým vzděláním) a sloupcové marginální četnosti n_{+j} (počet nevěst s j-tým vzděláním)

test nezávislosti v kont. tabulce

- nulová hypotéza H_0 tvrdí, že dva znaky jsou nezávislé
- ke každé sdružené četnosti spočítáme očekávanou četnost (četnost v průměru očekávanou v případě, že platí H_0)

$$o_{ij} = \frac{n_{i+}n_{+j}}{n}$$

- ověříme podmínku $o_{ij} \geq 5$
- spočítáme statistiku chí-kvadrát

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

- určíme počet stupňů volnosti: $f = (r-1)(c-1)$
- nulovou hypotézu zamítneme (závislost prokážeme) na hladině významnosti α, když bude statistika chí-kvadrát příliš veliká:

$$\chi^2 \geq \chi_f^2(1-\alpha)$$

příklad: vzdělání snoubenců

v závorce jsou očekávané četnosti

	nevěsta			
ženich	základní	střední	VŠ	celkem
základní	24 (13,26)	12 (17,55)	3 (8,19)	39
střední	7 (11,56)	24 (15,30)	3 (7,14)	34
VŠ	3 (9,18)	9 (12,15)	15 (5,67)	27
celkem	34	45	21	100

- $\chi^2 = 43{,}2 > \chi^2_4(0{,}95) = 9{,}5$, $p < 0{,}1$ %
- 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
- POZOR, test nic netvrdí o shodě marginálních pstí (že rozdělení úrovní vzdělání jsou u ženichů a nevěst stejná)

test homogenity

- zjišťujeme četnosti zvoleného znaku, který nabývá c různých hodnot, za r různých podmínek (naše četnosti stran hracích kostek A a B, tedy $c = 6$ $r = 2$)
- četnosti mají za i-té podmínky multinomické rozdělení s pravděpodobnostmi $\pi_{i1}, \ldots, \pi_{ic}$
- rozhodujeme o nulové hypotéze H_0, podle které jsou parametry (pravděpodobnosti) těchto multinomických rozdělení stejné (pravděpodobnosti jedniček jsou u obou kostek stejné, psti dvojek jsou u obou kostek stejné …)
- označme četnost j-té hodnoty za i-té podmínky jako n_{ij}
- očekávané četnosti o_{ij}, počet stupňů volnosti f i statistika chí-kvadrát se určí formálně stejně jako u testu nezávislosti
- stejné je také rozhodování o H_0

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

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

$$\chi^2 > 11{,}07 = \chi^2_5(0{,}95), \qquad p = 0{,}3 \text{ \%}$$

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

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
- například pro

VŠ	strana A	strana B	celkem
ano	11	4	15
ne	6	9	15
celkem	17	13	30

vyjde

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

čtyřpolní tabulka – prokazování závislosti

- chí-kvadrát porovnávající teoretické a očekávané četnosti čtyřpolní tabulky 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 $\chi^2 \geq \chi_1^2(\alpha)$
- 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

$$3{,}39 < 3{,}84 = \chi_1^2(0{,}95), \qquad p = 6{,}5\ \%$$

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 $\chi^2_{\text{Yates}} \geq \chi_1^2(1-\alpha)$
- **Fisherův exaktní test** počítá přímo p-hodnotu