

Statistika

(D360P03Z, D360P03U)
akademický rok 2005/2006

Karel Zvára
karel.zvara@mff.cuni.cz
<http://www.karlin.mff.cuni.cz/~zvára>

naposled upraveno 10. ledna 2006

cvičení, zápočet, zkouška

- cvičení v počítačových učebnách
- MS Excel, volně šiřitelný program R (<http://cran.r-project.org/>)
- (aktivní účast na cvičení, maximálně dvě absence) & (napsání zápočtového testu) ⇒ zápočet
- obsah cvičení více přizpůsoben studovanému oboru
- přednášky formulovány obecněji
- zkouška nejspíš písemná, kombinovaná s ústní, zápočet **musí** zkoušce **předcházet**; přihlašování ke zkoušce přes SIS

literatura

- Z. Pavlík, K. Kühnl: Úvod do kvantitativních metod pro geografii, SPN Praha, 1981
- K. Zvára: Biostatistika, Karolinum Praha, 1998, 2000, 2001, 2003
- T. H. Wonnacot, R. J. Wonnacot: Statistika pro obchod a hospodářství, Victoria Publishing Praha, 1992

přehled témat (1)

- popisná statistika (měřítka, charakteristiky polohy, variability, souvislost znaků)
- souvislost kvalitativních znaků (kontingenční tabulka)
- souvislost kvantitativních znaků (korelační koeficienty)
- pravděpodobnost (základní kombinatorické pojmy, klasická definice, podmíněná pravděpodobnost, nezávislost)
- náhodná veličina (rozdělení, střední hodnota, rozptyl, hustota, distribuční funkce)

přehled témat (2)

- důležitá rozdělení (normální, binomické, Poissonovo, vzájemné vztahy)
- princip statistického usuzování (populace a výběr, parametry a jejich odhady)
- interval spolehlivosti pro parametr, volba rozsahu výběru
- testování hypotéz (chyba 1. druhu, 2. druhu, hladina testu, síla testu, p -hodnota)
- testy (o populačním průměru, populačním podílu či podílech, nezávislosti, regresních koeficientech)
- regrese jako popis závislosti spojitých veličin

5

příklad statistického zjišťování

- zjišťování se týká 200 mužů středního věku
- v souboru je 80 kuřáků a 120 nekuřáků
- 85 mužů má oči modré, 25 hnědé, 90 jiné barvy
- 27 mužů má jen základní vzdělání, 44 neúplné střední, 65 maturitu, 64 vysokoškolské
- 22 se jich narodilo v roce 1942, 19 v roce 1943, 25 v roce 1944, . . . , 18 v roce 1951
- hmotnosti jednotlivých mužů jsou 83, 92, . . . , 63 kg
- Co mají tyto údaje společného? Čím se údaje liší?

6

co měříme (zjišťujeme) a kde

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

7

měřítko (1)

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (země původu, barva očí) jednoznačně dané hodnoty
- **ordinální** (dosažené vzdělání, stupeň bolesti) jednoznačně dané hodnoty, možné hodnoty jsou *uspořádané*
- **intervalové** (teplota v Celsiově stupnici, rok narození) konstantní vzdálenosti mezi sousedními hodnotami, nula jen konvence
- **poměrové** (hmotnost, výška, HDP, počet obyvatel) násobek zvolené jednotky, nula = neexistence měřené vlastnosti

8

měřítka (2)

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

9

příklad: 100 hodů kostkou

počty puntíků 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

11

veličina

- číselně vyjádřený výsledek měření
- hodnoty znaků v intervalovém, 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**)

10

hody kostkou jako hromadný jev

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

12

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
```

17

variační řada, pořadí

- x_1, x_2, \dots, x_n původní (neuspořádaná) data – hodnoty znaku v měřítku aspoň ordinálním uvedené v původním pořadí, bez ohledu na případná opakování
- **variační řada** $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ [\[sort\(x\)\]](#)
data uspořádána tak, aby hodnoty neklesaly (**závorky u indexů**)
- **pořadí** [\[rank\]](#) – umístění pozorování ve variační řadě; shodným hodnotám dáváme průměrné pořadí [\[rank\(x\)\]](#)

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

19

příklad: věk 99 matek – variační řada

uspořádaný soubor hodnot – variační řada

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

18

třídění, třídění četnosti

- spojitá veličina s velkým počtem naměřených hodnot
- obor hodnot rozdělíme na nepřekrývající se třídy (intervaly), nejlépe stejné délky
- všechna pozorování z daného intervalu nahradíme zástupnou hodnotou (zpravidla středem intervalu) x_j^*
- zjistíme (**absolutní**) **četnosti** n_1, \dots, n_k jednotlivých tříd
- **kumulativní četnosti** udávají počet hodnot v dané třídě a třídách předcházejících ($1 \leq j \leq k$) [\[cumsum\(\)\]](#)

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

20

věk matek – třídní četnosti $k = 7$

interval	x_j^*	n_j	$f_j = n_j/n$	N_j	N_j/n
do 20	19	5	0,051	5	0,051
21 až 23	22	27	0,273	32	0,324
24 až 26	25	32	0,322	64	0,646
27 až 29	28	19	0,192	83	0,838
30 až 32	31	8	0,081	91	0,919
33 až 35	34	6	0,061	97	0,980
36 až 38	37	2	0,020	99	1,000

21

grafické znázornění třídních četností

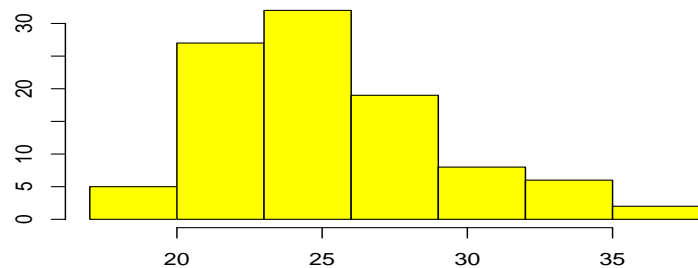
- **histogram** je založen na třídění do intervalů, výjimečně zobrazuje přímo četnosti jednotlivých hodnot (barplot) [\[hist\(\)\]](#)
- každé třídě odpovídá obdélník o **ploše úměrné četnosti** (absolutní nebo relativní)
- při stejných šířkách intervalů h odpovídají četnostem výšky obdélníků
- počet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

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

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

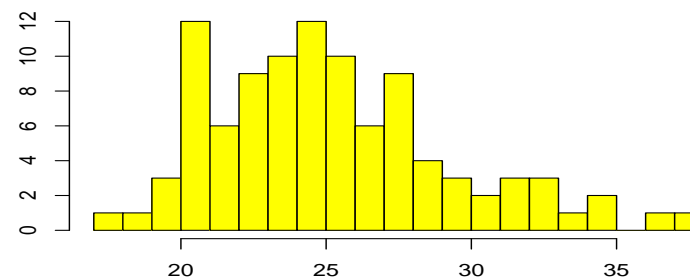
22

histogram, $h = 3$ ($k = 7$) [\[hist\(vek.m,seq\(17,38,by=3\),col="yellow"\)\]](#)



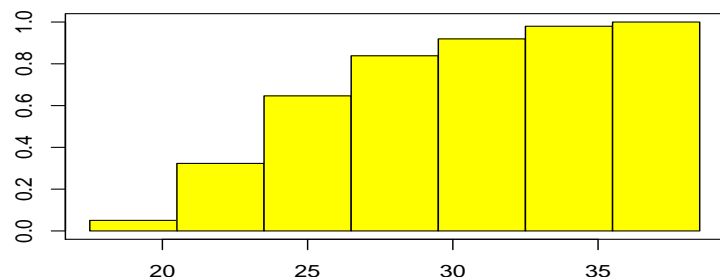
23

histogram, $h = 1$ [\[hist\(vek.m,seq\(17,38,by=1\),col="yellow"\)\]](#)



24

příklad: věk matek (kumulativní relativní četnosti)



25

parametry – odhady, statistiky

- podle toho, jakou roli hraje hodnocený soubor, rozlišujeme charakteristiky
 - **populační**: vztahované k populaci, mnohdy jen ideální, námi představované, jsou to **parametry** modelu
 - **výběrové**: vztahované k výběru z nějaké populace, takže jde o **odhady** nějakých populačních parametrů, jsou to **statistiky** spočítané z výběru
- příkladem dvojice odhad – parametr je dvojice relativní četnost – pravděpodobnost (17/100 vers. 1/6)
- statistiky se používají při **statistické indukci** (statistickém rozhodování) [statistical inference (decisions)]

27

populace

- velká populace, spojitá veličina – intervaly mohou být krátké, obálce histogramu relativních četností odpovídá **hustota** $f_X(x)$ [density]
- podobně kumulativním relativním četnostem odpovídá **distribuční funkce** [distribution function]
- hodnota distribuční funkce $F_X(x)$ je pravděpodobnost, že náhodná veličina X nepřekročí x :

$$F_X(x) = P(X \leq x)$$

- souvislost: hustota je derivace distribuční funkce:

$$f_X(x) = F'_X(x)$$

26

charakteristiky polohy (1)

- **medián** (prostřední hodnota) \tilde{x} [median] [median(x)]

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

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

medián je číslo, které dělí data na dvě stejně velké části (velkých hodnot a malých hodnot)

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

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

populační medián číslo, které dělí možné hodnoty náhodné veličiny na dva stejně pravděpodobné intervaly

28

charakteristiky polohy (2)

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

$$F_X(\mu_p) = P(X \leq \mu_p) = p$$

29

příklad: věk 99 matek – variační řada

variační řada, medián $\tilde{x} = 25$, kvartily $Q_1 = 23$, $Q_3 = 28$

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

31

pro zajímavost algoritmus výpočtu percentilu v R (jedna z možných definic – Gumbel(1939))

- najde se celé číslo k splňující $([x])$ znamená celou část z x

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

tedy

$$k = [1 + (n-1) \cdot p]$$

- provede se lineární interpolace mezi $x_{(k)}$ a $x_{(k+1)}$ ($\{x\}$ znamená zlomkovou část x , o kolik přesahuje celé číslo)

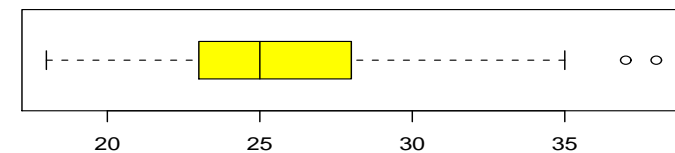
$$q = \{1 + (n-1) \cdot p\} = (1 + (n-1) \cdot p) - k$$

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

30

krabicový diagram

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



32

charakteristiky polohy (3)

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

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

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

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

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

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

- **populační průměr**: značíme μ

33

charakteristiky polohy (4)

- u nula-jedničkového měřítka: průměr = relativní četnost jedniček, populační průměr = pravděpodobnost jedničky
- **modus** \hat{x} [mode] nejčastější hodnota (lze počítat také pro nominální či ordinální měřítka)
- modus nemusí být určen jednoznačně
- **populační modus**
 - pro spojitou veličinu – hodnota, kde je hustota maximální
 - pro diskrétní veličinu (četnosti) – nejpravděpodobnější hodnota

34

příklad – věk matek

- průměr

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

- vážený průměr založený na třídění

$$\begin{aligned} \bar{x} &= \frac{5 \cdot 19 + 27 \cdot 22 + 32 \cdot 25 + 19 \cdot 28 + 8 \cdot 31 + 6 \cdot 34 + 2 \cdot 37}{5 + 27 + 32 + 19 + 8 + 6 + 2} \\ &= \frac{2547}{99} \doteq 25,7 \end{aligned}$$

- modus není určen jednoznačně: $\hat{x} = 21$, $\hat{x} = 25$

35

charakteristiky polohy (5)

- **alfa-useknutý průměr** [trimmed mean]: nejprve se oddělí (usekne) 100α % nejmenších a 100α % největších hodnot, ze zbytku se spočítá průměr
- je robustní vůči odlehlým hodnotám
- volí se zpravidla $\alpha = 0,1$ (0,15)
- věk matek [mean(vek.m,trim=0.1)]

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

36

příklad: věk 99 matek

vyloučí se $[0, 1.99] = [9, 9] = 9$ ($[x]$ znamená celou část z x)
nejmenších a 9 největších hodnot

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

37

charakteristiky polohy v geografii/demografii

- často známe jen průměry v dílčích souborech a četnosti: průměry se použijí jako x_j^* , četnosti standardně
- příklad: věk nových profesorů a docentů UK 2002:
41 profesorů, průměrný věk 51,1
77 docentů, průměrný věk 47,8
celkový průměr nově habilitovaných (**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))]

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

39

vlastnosti charakteristik polohy

- změníme-li všechny hodnoty x_i tak, že přidáme ke každé stejnou konstantu a , změní se o tutéž konstantu také charakteristika polohy (posunutí)
- změníme-li všechny hodnoty x_i tak, že je vynásobíme kladnou konstantou b , toutéž konstantou musíme vynásobit původní charakteristiku polohy, abychom dostali charakteristiku polohy pro upravená data (změna měřítka)
- obecně pro míru polohy $m(x)$

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

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

38

charakteristiky polohy v geografii/demografii (2)

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

40

charakteristiky variability (1)

- měří nestejnost (**variabilitu**) hodnot spojité veličiny
- obecně pro míru variability $s(x)$

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

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

41

charakteristiky variability (2)

- (výběrový) **rozptyl** (variance) [variance] [var(x)]
(nevyhovuje obecnému požadavku přesně: $s_{a+b \cdot x}^2 = b^2 \cdot s_x^2$)

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

- **populační rozptyl** σ^2 (případ, kdy populace má konečný počet prvků): ve jmenovateli je n

42

charakteristiky variability (3)

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

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

- vyhovuje obecnému požadavku
- výhoda směrodatné odchylky:
stejný fyzikální rozměr jako původní data
- výběrový rozptyl z *třídních* četností Sheppardova korekce:

$$\text{odečti } \frac{h^2}{12}$$

43

příklad – věk matek

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

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

- směrodatná odchylka je 4,12

44

příklad – věk matek

- pomocí třídnicích četností

$$s^2 = \frac{1}{98} \left((5 \cdot 19^2 + 27 \cdot 22^2 + \dots + 6 \cdot 34^2 + 2 \cdot 37^2) - 99 \cdot \left(\frac{2547}{99} \right)^2 \right) \\ = 16,36 = (4,05)^2$$

- navíc Sheppardova korekce

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

45

normované charakteristiky rozptýlenosti

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

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

- **(Giniho) koeficient koncentrace**

$$G = \frac{\Delta}{2\bar{x}}$$

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

47

charakteristiky variability (4)

- **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**: průměr vzájemných vzdáleností všech dvojic (je jich n^2)

$$\Delta = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|$$

46

Lorenzova křivka

- variační řada: $0 < x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ [sort(x)]
- kumulativní součty pro $j = 0, 1, \dots, n$ [cumsum(sort(x))]

$$t_0 = 0 \quad t_j = x_{(1)} + x_{(2)} + \dots + x_{(j)} = \sum_{i=1}^j x_{(i)}$$

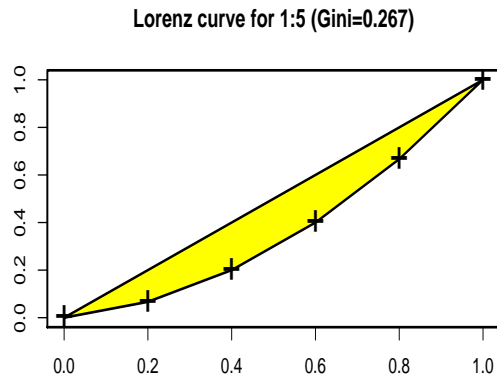
- úsečkami spojit body $[j/n; t_j/t_n]$, $0 \leq j \leq n$
- zajímá nás plocha nad touto lomenou čarou a pod úhlopříčkou jednotkového čtverce
- plocha měří nerovnoměrnost rozdělení nějakého zdroje
- kdyby dostal každý stejně, bude velikost plochy nulová
- Giniho koeficient koncentrace je dvojnásobkem této plochy

48

příklad

x_1, \dots, x_5 : 1, 2, 3, 4, 5

$x_{(j)}$	t_j	t_j/t_n
1	1	0,067
2	3	0,200
3	6	0,400
4	10	0,667
5	15	1,000



49

výpočet Giniho koeficientu ($n = 5$):

$$\begin{aligned}
 5^2 \cdot \Delta &= |1-1| + |1-2| + |1-3| + |1-4| + |1-5| \\
 &\quad + |2-1| + |2-2| + |2-3| + |2-4| + |2-5| \\
 &\quad + |3-1| + |3-2| + |3-3| + |3-4| + |3-5| \\
 &\quad + |4-1| + |4-2| + |4-3| + |4-4| + |4-5| \\
 &\quad + |5-1| + |5-2| + |5-3| + |5-4| + |5-5| \\
 &= 10 + 7 + 6 + 7 + 10
 \end{aligned}$$

$$\Delta = 40/25 = 1,6$$

$$\bar{x} = 3$$

$$G = \frac{1,6}{2 \cdot 3} = \frac{1,6}{6} = 0,267$$

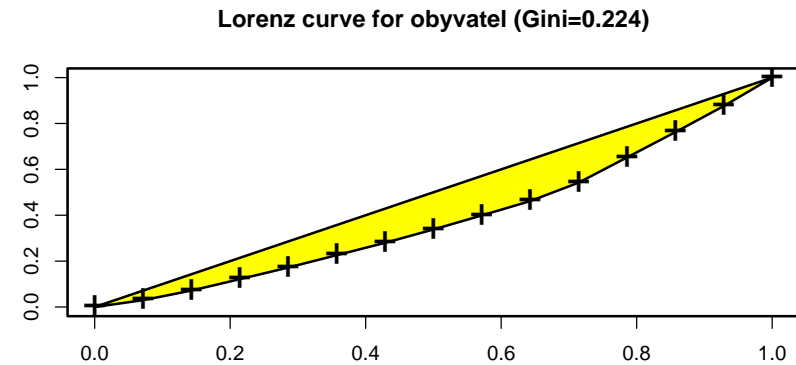
50

příklad: obyvatelé krajů (počet hejtmanů)

j	$x_{(j)}$	t_j	j/n	t_j/t_n	$x_{(j)}$	t_j	t_j/t_n
0	—	0	0,000	0,000	—	0	0,000
1	303761	303761	0,071	0,030	1	1	0,071
2	427418	731179	0,143	0,072	1	2	0,143
3	506849	1238028	0,214	0,121	1	3	0,214
4	517959	1755987	0,286	0,172	1	4	0,286
5	548698	2304685	0,357	0,226	1	5	0,357
6	549369	2854054	0,429	0,280	1	6	0,429
...
13	1158800	8936427	0,929	0,876	1	13	0,929
14	1264347	10200774	1,000	1,000	1	14	1,000

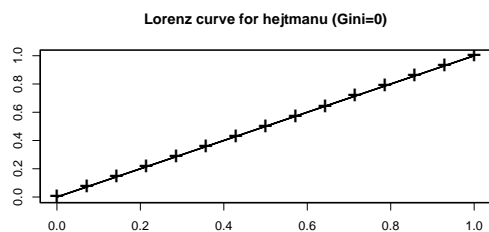
51

Lorenzova křivka (obyvatelé – kraje)



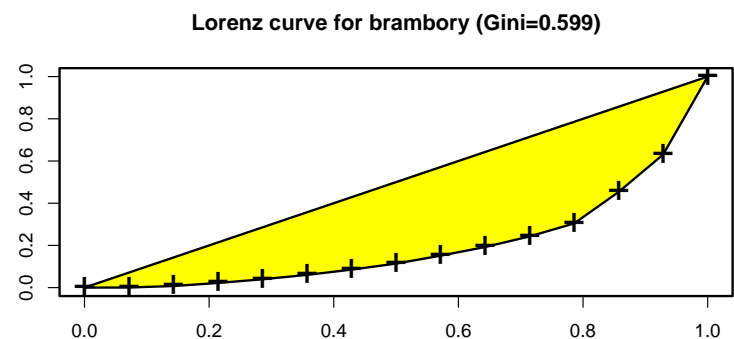
52

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



53

roční úroda brambor je mezi kraji rozdělena mnohem nerovnoměrněji: např. 70 % brambor se pěstuje ve třech krajích (Vys, Stč, Jč)



54

případ s vahami

- někdy nutno přihlédnout k velikosti jednotek, pak měříme nerovnoměrnost **hustoty** rozdělení zdroje (x_i – velikost jednotky, y_i – velikost produktu, y_i/x_i – hustota, \bar{y}_x – vážený průměr hustot y_i/x_i s vahami x_i)

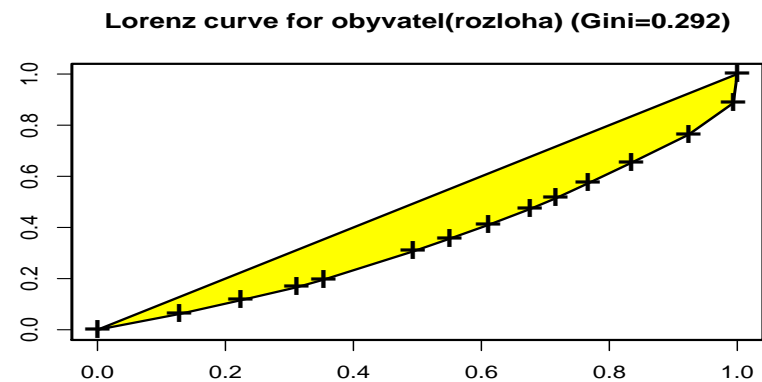
$$\Delta = \frac{1}{(\sum x_i)^2} \sum_{i=1}^n \sum_{j=1}^n x_i x_j \left| \frac{y_i}{x_i} - \frac{y_j}{x_j} \right|$$

$$G = \frac{\Delta}{2\bar{y}_x}$$

$$\bar{y}_x = \frac{\sum_i x_i (y_i/x_i)}{\sum_i x_i} = \frac{\bar{y}}{\bar{x}}$$

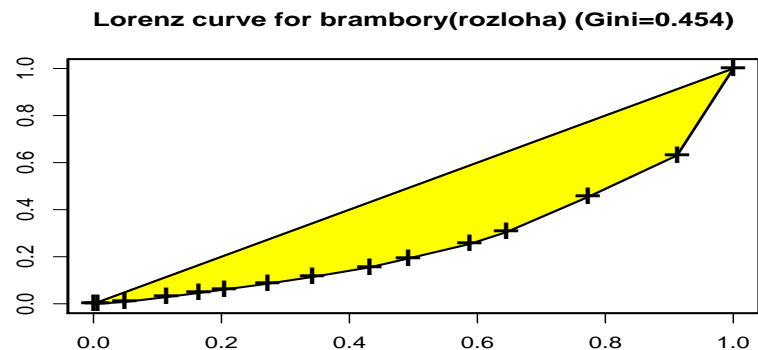
55

Lorenzova křivka (obyvatelé a rozloha krajů)



56

roční úroda brambor s přihlédnutím k velikosti krajů



57

z-skór, standardizace

- variační koeficient, Giniho koeficient – příklady bezrozměrných veličin

- z-skóry $[(x - \text{mean}(x))/\text{sd}(x)]$ nebo $[z <- c(\text{scale}(x))]$

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

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

58

šikmost, špičatost

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

- **špičatost** b_2 – průměr ze 4. mocnin z-skórů (někdy se odečítá 3)

$$\sqrt{b_1} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right)^3 \quad b_2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right)^4$$

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

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

59

dvojice znaků (veličin)

- na jedné statistické jednotce se měří dva znaky
- lze vyšetřovat závislost
- postupy (i grafické) závisí na měřících obou znaků
 - kvalitativní – kvalitativní
 - kvalitativní – kvantitativní
 - kvantitativní – kvantitativní
- zatím popisné charakteristiky, prokazování závislosti později

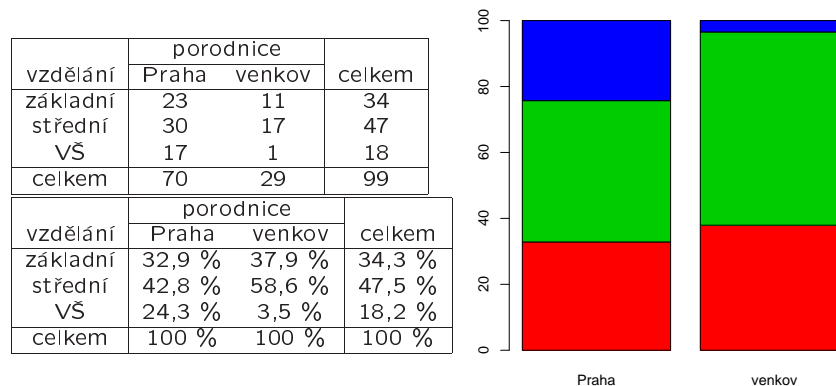
60

kvalitativní – kvalitativní

- kvalitativní data – nominální (ordinální) měřítko, vyjadřujeme pomocí četností
- dva znaky – četnosti možných **dvojic hodnot** n_{ij}
- zapisujeme do **kontingenční tabulky** [contingency table]
[table(x,y)]
- doplňujeme **marginální četnosti** [marginal frequencies] – součty po řádcích a po sloupcích - četnosti jednotlivých znaků zvlášť
- oba znaky nula-jedničkové – kontingenční tabulka 2x2, **čtyřpolní tabulka** [fourfold table]

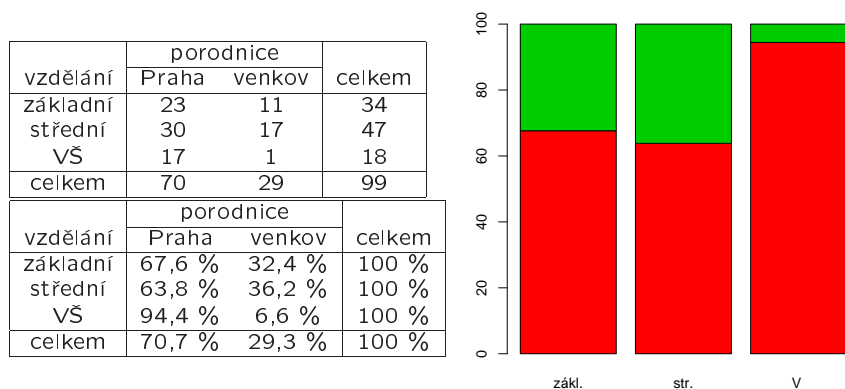
61

příklad – vzdělání matek (pozor na orientaci)



62

příklad – vzdělání matek (pozor na orientaci)



63

příklad – vzdělání matek (očekávané četnosti)

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

vzdělání	porodnice		celkem	vzdělání	porodnice		celkem
	Praha	venkov			Praha	venkov	
základní	23	11	34	základní	24,0	10,0	34
střední	30	17	47	střední	33,2	13,8	47
VŠ	17	1	18	VŠ	12,7	5,3	18
celkem	70	29	99	celkem	70	29	99

64

příklad – vzdělání matek (očekávané četnosti)

vzdělání	porodnice		celkem	vzdělání	porodnice		celkem
	Praha	venkov			Praha	venkov	
základní	23	11	34	základní	24,0	9,9	34
střední	30	17	47	střední	33,3	13,8	47
VŠ	17	1	18	VŠ	12,7	5,3	18
celkem	70	29	99	celkem	70	29	99

empirické a **očekávané** [expected] četnosti porovnáme pomocí statistiky **chi-kvadrát** [chi-squared]: `[chisq.test(table(Vzdelani, Porodnice))]`

$$\chi^2 = \frac{(23 - 24)^2}{24} + \frac{(11 - 9,9)^2}{9,9} + \frac{(30 - 33,3)^2}{33,3} + \dots + \frac{(1 - 5,3)^2}{5,3} = 6,12$$

velká hodnota χ^2 svědčí o velké neshodě

65

příklad: plánovaná těhotenství

- je souvislost mezi odpověďmi o plánovaném těhotenství a vzděláním matek?

vzdělání	plánované		celkem	vzdělání	plánované		celkem
	ne	ano			ne	ano	
základní	20	14	34	základní	58,8 %	42,1 %	100 %
střední	16	31	47	střední	34,0 %	66,0 %	100 %
VŠ	5	13	18	VŠ	27,8 %	72,2 %	100 %
celkem	41	58	99	celkem	41,4 %	58,6 %	100 %

66

příklad: plánovaná těhotenství (očekávané četnosti)

vzdělání	plánované		celkem
	ne	ano	
základní	14,08	19,92	34
střední	19,46	27,54	47
VŠ	7,46	10,54	18
celkem	41	58	99

$$99 \cdot \frac{41}{99} \cdot \frac{34}{99} = \frac{41 \cdot 34}{99} = 14,08$$

$$99 \cdot \frac{58}{99} \cdot \frac{34}{99} = \frac{58 \cdot 34}{99} = 19,92$$

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

67

příklad: předvolební výzkum

30 voličů bylo dotázáno, které ze dvou stran dají přednost; souvisí odpovědi s pohlavím?

	strana		celkem
	A	B	
muž	11	4	15
žena	6	9	15
celkem	17	13	30

	strana		celkem		strana		celkem
	A	B			A	B	
muž	73 %	27 %	100 %	muž	65 %	31 %	50 %
žena	40 %	60 %	100 %	žena	35 %	69 %	50 %
celkem	57 %	43 %	100 %	celkem	100 %	100 %	100 %

68

čtyřpolní tabulka

- obecné označení četností v čtyřpolní tabulce

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

- **sílu závislosti** lze měřit ϕ -koeficientem [phi coefficient] (čtyřpolním korelačním koeficientem)

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

- ϕ je mezi -1 a 1

- např. pro

11	4	15
6	9	15
17	13	30

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

69

příklad: předvolební průzkum

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

	strana		celkem
	A	B	
muž	11	4	15
žena	6	9	15
celkem	17	13	30

- v našem příkladu vychází $\phi = 0,34 > 0$

(tedy kladné), protože je $11 \cdot 9 > 6 \cdot 4$

70

čtyřpolní tabulka – prokazování závislosti

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

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

- příklad (předvolební průzkum)

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

71

Simpsonův paradox dílčí tabulky mají závislost jiného směru, než jejich součet (zde bez ohledu na to, kde žijí)

venkov	A	B	celkem
muž	6	7	13
žena	2	3	5
celkem	8	10	18

$\phi = 0,055$

celkem	A	B	celkem
muž	11	9	20
žena	13	8	21
celkem	24	17	41

$\phi = -0,07$

město	A	B	celkem
muž	5	2	7
žena	11	5	16
celkem	16	7	23

$\phi = 0,027$

kdyby byl stejný poměr mezi počtem mužů a počtem žen oslovených ve městě a na venkově, problém by nevznikl

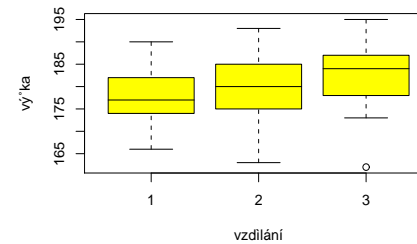
72

dvojice kvalitativní – kvantitativní

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

73

příklad: výška otce ~ vzdělání matky [boxplot(vyska.o~Vzdelani)]



vzdělání	rozsah	průměr	sm. odch.
základní	34	177,1	6,0
střední	47	179,5	6,4
VŠ	18	182,8	7,8
celkem	99	179,3	6,8

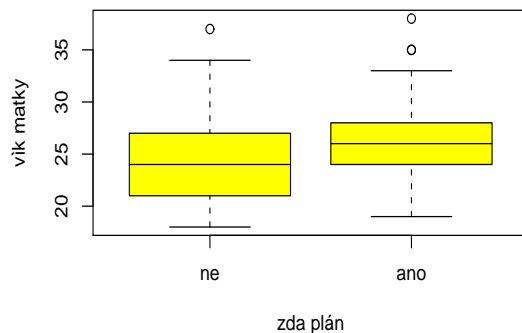
$$\bar{x} = \frac{34 \cdot 177,1 + 47 \cdot 179,5 + 18 \cdot 182,8}{34 + 47 + 18} = 179,3$$

$$s^2 = 6,8^2 > \frac{34 \cdot 6,0^2 + 47 \cdot 6,4^2 + 18 \cdot 7,8^2}{34 + 47 + 18} = 6,6^2$$

74

příklad: věk matek

[boxplot(vek.m~Plan)]



plán	ne	ano
n	41	58
\bar{x}	24,7	26,4
\tilde{x}	24,0	26,0
Q_1	21,0	24,0
Q_2	27,0	28,0
sd	4,24	3,93
R_Q	6,00	4,00

75

závislost nula-jedničkové – kvantitativní

[cor(vek.m,Plan)]

- pro nula-jedničkové x sílu závislosti x, y vyjadřuje **bodově biseriální korelační koeficient** [point-biserial]

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

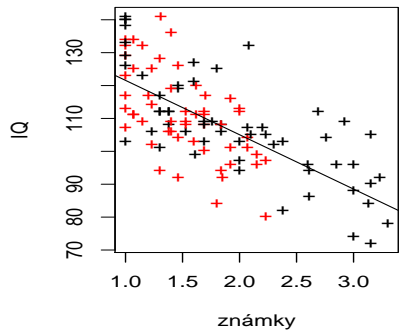
- kde \bar{y}_1 je průměr těch y , u nichž je $x = 1$
- kde \bar{y}_0 je průměr těch y , u nichž je $x = 0$
- kde s je směrodatná odchylka *všech* y ($n - 1$ ve jmenovateli)
- kde n_0 je počet nul a n_1 počet jedniček mezi x

- platí $-1 \leq r_{\text{bis}} \leq 1$

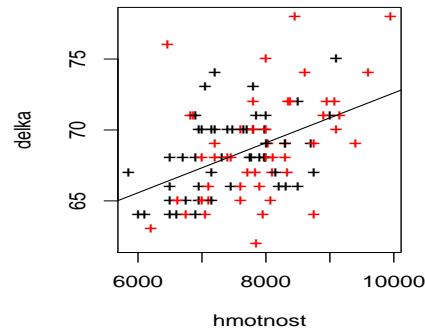
- příklad: $r_{\text{bis}} = \frac{26,4 - 24,7}{4,12} \sqrt{\frac{41 \cdot 58}{99 \cdot 98}} = 0,20$

76

dvojice kvantitativních veličin [plot(iq zn7,data=Iq,col=1+(divka))]



$r = -0,69$



$r = 0,45$

77

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

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

79

závislost spojité veličin

[cor(vek.o,vek.m)]

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

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

- (Pearsonův, momentový) korelační koeficient [(Pearson, product-moment) correlation coefficient] lze zapsat pomocí z-skórů

$$r = \frac{s_{xy}}{s_x s_y} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y} \right) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- bodově biseriální korelační koeficient i ϕ -koeficient (čtyřpolní korelační koeficient) jsou speciální případy Pearsonova korelačního koeficientu, když za nula-jedničkovou veličinu použijeme opravdu nuly a jedničky

78

vlastnosti Pearsonova korelačního koeficientu

- vypovídá o směru závislosti
- při $r < 0$ s rostoucím x v průměru y klesá
- platí $-1 \leq r \leq 1$
- $|r| = 1$ jedině, když body $[x; y]$ leží na přímce
- vzájemné nezávislosti x, y odpovídají r blízká nule
- hranice statistické průkaznosti závisí na n , čím větší n , tím menší $|r|$ stačí k prokázání závislosti (tabulky)
- takto lze závislost prokazovat jen někdy (normální rozdělení)
- špatně zachytí křivočarou (nelineární) závislost

80

Spearmanův korelační koeficient `[cor(x,y,method="spearman")]`

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

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

- vhodný pro nelineární monotonní **závislost**, nevadí odlehle hodnoty
- při testování nemusí být normální rozdělení

81

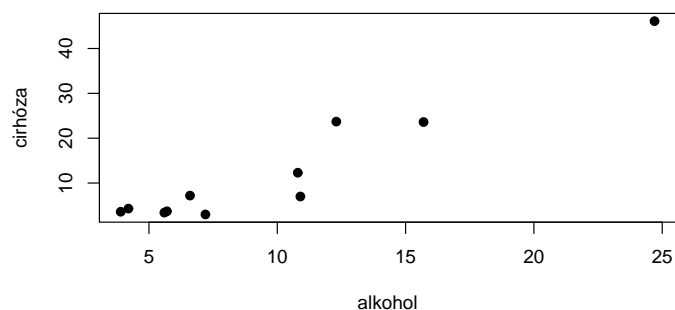
příklad: alkohol a úmrtnost na cirhózu

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

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

82

příklad: spotřeba alkoholu a cirhóza jater



83

pravděpodobnost

- **pokus** – dobře definovaná situace (postup), která končí jedním z řady možných výsledků
- **náhodný pokus** – pokus, u něhož předem nevíme, který výsledek nastane; 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
- při velkém počtu opakování pokusu se relativní četnost jevu blíží k pravděpodobnosti tohoto jevu

84

klasická pravděpodobnost

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

$$P(A) = \frac{M_A}{M}$$

85

příklad: hrací kostka

- idealizovaná symetrická homogenní kostka
- každá strana má stejnou pravděpodobnost
- A – padne šestka, B – padne sudé číslo
- $M = 6$
- $M_A = 1$, tedy $P(A) = 1/6$
- $M_B = 3$, tedy $P(B) = 3/6 = 1/2$

86

faktoriál

[factorial(n)]

- **faktoriál** $n! = n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1$ $0! = 1$
- 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:
 $14! = 14 \cdot 13 \cdot 12 \cdot \dots \cdot 2 \cdot 1 = 87\,178\,291\,200 = 8,7 \cdot 10^{10}$

87

počet kombinací

[choose(n, k)]

- **kombinační číslo** $\binom{n}{k}$ (čti „ n nad k “)
- počet k -prvkových podmnožin množiny o n prvcích nezávisle na jejich pořadí

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

- kolika způsoby si mohou 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 mohou vybrat tři knihy? (10)

88

příklad: losování otázek (1)

- student *neumí* 5 otázek, *umí* 10 otázek
- losuje se dvojice otázek z oněch 15 otázek
- pravděpodobnost, že student nezná ani jednu z vylosovaných:
- elementární jevy: první losovaná otázka – 15 možností, druhá jen 14 možností, nezáleží na pořadí, tedy dělit 2

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

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

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

89

příklad: losování otázek (2)

- pravděpodobnost, že zná *právě* jednu otázku

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

- pravděpodobnost, že zná *právě* dvě otázky

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

- pravděpodobnost, že zná *aspoň jednu* otázku

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

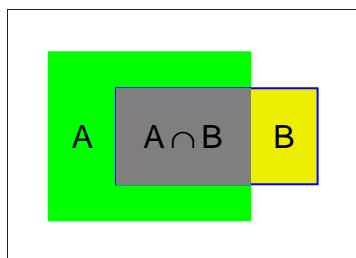
- kontrola: $M_D + M_A = M$

90

pravidla pro pravděpodobnost (1)

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

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



$P(A \cup B)$ = 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 + žlutá + 2 · šedivá plocha
 $P(A \cup B) = 0,42 + 0,24 - 0,16 = 0,50$

91

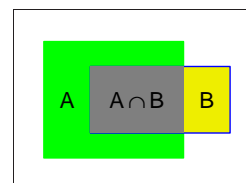
pravidla pro pravděpodobnost (2)

- **neslučitelné jevy**: nemohou nastat nikdy současně, navzájem se vylučují; platí pro ně

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

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

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

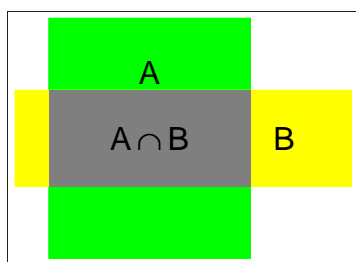


$P(B) = 0,24$ = žlutá + šedivá plocha
 $P(A \cap B) = 0,16$ = šedivá plocha
 $P(A|B)$ = šedivá vzhledem k (žlutá + šedivá)
 $P(A|B) = 0,16/0,24 = 0,67$, ale $P(A) = 0,42$

92

- **nezávislé jevy**: výskyt jednoho jevu **neovlivní** pravděpodobnost výskytu druhého (definice **nezávislosti** náhodných jevů):

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



$P(A) = 0,60 = \text{zelená} + \text{šedivá}$
 $P(B) = 0,40 = \text{žlutá} + \text{šedivá plocha}$
 $P(A \cap B) = 0,24 = \text{šedivá plocha}$
 $P(A|B) = \text{šedivá vzhledem k (žlutá + šedivá)}$
 $P(A|B) = 0,24/0,40 = 0,60$

93

idealizovaný příklad

- A – jednička ze statistiky, $P(A) = 0,3$
- B – jednička z matematiky, $P(B) = 0,2$
- $A \cap B$ – jednička z obou předmětů, $P(A \cap B) = 0,1$
- jsou jevy A, B nezávislé (je výskyt jedniček ze dvou předmětů nezávislý)? NE, protože $0,3 \cdot 0,2 \neq 0,1$
- jaká je pst jedničky ze statistiky, když už je z matematiky?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0,1}{0,2} = 0,5$$

- pravděpodobnost, že aspoň jedna jednička:

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

94

rozdělení náhodné veličiny

- **náhodná veličina** – číselně vyjádřený výsledek náhodného pokusu
- **diskrétní rozdělení** (pro četnosti) určeno seznamem možných hodnot a jejich pravděpodobnostmi:

$$x_1, x_2, \dots$$

$$P(X = x_1), P(X = x_2), \dots$$

- **spojité rozdělení** (pro spojité měřítko) určeno **distribuční funkcí**

$$F_X(x) = P(X \leq x)$$

nebo **hustotou**

$$f_X(x) = \frac{d}{dx} F_X(x), \quad F_X(x) = \int_{-\infty}^x f_X(t) dt$$

95

příklad diskrétního rozdělení: známka u zkoušky

známka k	1	2	3	4
$P(X = k)$	0,3	0,4	0,2	0,1
$P(Y = k)$	0,3	0,3	0,2	0,2

(Z této tabulky *nic* nepoznáme o případné závislosti!)

Jak jedním číslem charakterizovat úroveň známek?

Obyčejný průměr by X, Y nerozlišil \Rightarrow **vážený průměr** vahami známek budou jejich pravděpodobnosti

dostaneme tak **populační průměry**

$$\mu_X = 1 \cdot 0,3 + 2 \cdot 0,4 + 3 \cdot 0,2 + 4 \cdot 0,1 = 2,1$$

$$\mu_Y = 1 \cdot 0,3 + 2 \cdot 0,3 + 3 \cdot 0,2 + 4 \cdot 0,2 = 2,3$$

96

charakteristiky rozdělení náhodné veličiny (1)

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

$$\mu_X = E X = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \dots = \sum_k x_k \cdot P(X = x_k)$$

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

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

97

charakteristiky rozdělení náhodné veličiny (2)

- **(populační) rozptyl** náhodné veličiny X – vážený průměr čtverců vzdáleností možných hodnot od střední hodnoty

$$\begin{aligned} \sigma_X^2 &= E (X - \mu_X)^2 = (x_1 - \mu_X)^2 P(X = x_1) + (x_2 - \mu_X)^2 P(X = x_2) + \dots \\ &= \sum_k (x_k - \mu_X)^2 P(X = k) \end{aligned}$$

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

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

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

99

příklad diskrétního rozdělení: známka u zkoušky

známka k	1	2	3	4	μ	σ^2	σ
$P(X = k)$	0,3	0,4	0,2	0,1	2,1	0,89	0,943
$P(Y = k)$	0,3	0,3	0,2	0,2	2,3	1,21	1,100

Jak jedním číslem charakterizovat kolísání známek (jejich **variabilitu**)?
vážený průměr čtverců vzdáleností od střední hodnoty, vahami jsou známky = **(populační) rozptyl**

$$\begin{aligned} \sigma_X^2 &= (1 - 2,1)^2 \cdot 0,3 + (2 - 2,1)^2 \cdot 0,4 \\ &\quad + (3 - 2,1)^2 \cdot 0,2 + (4 - 2,1)^2 \cdot 0,1 = 0,89 = 0,943^2 \\ \sigma_Y^2 &= (1 - 2,3)^2 \cdot 0,3 + (2 - 2,3)^2 \cdot 0,3 \\ &\quad + (3 - 2,3)^2 \cdot 0,2 + (4 - 2,3)^2 \cdot 0,2 = 1,21 = 1,1^2 \end{aligned}$$

98

vlastnosti střední hodnoty a rozptylu

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

$$\mu_{a+X} = E (a + X) = a + E X = a + \mu_X$$

$$\mu_{b \cdot X} = E (b \cdot X) = b \cdot E X = b \cdot \mu_X$$

$$\mu_{X+Y} = E (X + Y) = E X + E Y = \mu_X + \mu_Y$$

$$\sigma_{a+X}^2 = \sigma_X^2$$

$$\sigma_{bX}^2 = b^2 \sigma_X^2$$

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{X,Y}$$

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

$$= (x_1 - \mu_X)(y_1 - \mu_Y) P(X = x_1, Y = y_1)$$

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

(sčítá se přes všechny možné dvojice)

100

nezávislé náhodné veličiny

náhodné veličiny X, Y jsou **nezávislé**, když pro všechny dvojice možných hodnot (x_i, y_j) platí

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$$

jsou-li X, Y **nezávislé**, pak $\sigma_{X,Y} = 0$, tedy $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$

rozptyl součtu **nezávislých** náhodných veličin = součet rozptylů

101

příklad: známky u zkoušky

X	Y				P(X = k)
	1	2	3	4	
1	0,15	0,10	0,05	0,00	0,3
2	0,10	0,15	0,10	0,05	0,4
3	0,05	0,05	0,05	0,05	0,2
4	0,00	0,00	0,00	0,10	0,1
	0,3	0,3	0,2	0,2	1,0

$$\sigma_{X,Y} = (1-2,1)(1-2,3) \cdot 0,15 + (1-2,1)(2-2,3) \cdot 0,10 + \dots + (4-2,1)(3-2,3) \cdot 0,00 + (4-2,1)(4-2,3) \cdot 0,10 = 0,57$$

$$\rho_{X,Y} = \frac{0,57}{0,943 \cdot 1,1} = 0,55$$

103

(populační) korelační koeficient

Pearsonův korelační koeficient:

$$r_{x,y} = \frac{s_{xy}}{s_x s_y}$$

jeho **populační protějšek**

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

ρ_{XY} má stejné vlastnosti jako r_{xy} , zejména platí $|\rho_{XY}| \leq 1$ pro **nezávislé** náhodné veličiny X, Y je $\rho_{XY} = 0$

102

alternativní rozdělení

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

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

- (populační) **rozptyl**

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

104

binomické rozdělení $bi(n, \pi)$ (1)

- diskrétní rozdělení s parametry n, π ($0 < \pi < 1$)
- n **nezávislých** pokusů
- v každém zdar s pravděpodobností π , nezdar s $1 - \pi$
- X (**celk. počet zdarů**) má binomické rozdělení s parametry n, π
- X je součet n nezávislých náhodných veličin X_i , každá z nich má alternativní rozdělení s parametrem π

$\mu_X = n\pi$ z vlastnosti střední hodnoty součtu

$\sigma_X^2 = n\pi(1 - \pi)$ z vlastnosti součtu nezávislých náhodných veličin

105

binomické rozdělení $bi(n, \pi)$ (2)

- pravděpodobnosti možných hodnot

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

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

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

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

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

106

příklad: zkoušky

- C – zdar = udělat zkoušku, $P(C) = 0,8$
- zkoušku dělá $n = 10$ studentů stejně připravených (u všech stejná pravděpodobnost π), studenti neopisují (nezávislost)

- pravděpodobnost, že zkoušku udělá nějakých 9 studentů

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

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

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

- pravděpodobnost, že zkoušku udělá **daných** 9 studentů: 0,0268

107

příklad: kouření

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

•

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

- ukázky pravděpodobností možných hodnot [\[dbinom\(15,60,0.35\)\]](#)

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

108

Poissonovo rozdělení $Po(\lambda)$ (1)

- diskrétní rozdělení (zákon vzácných jevů)
- Y – počet výskytů jevu ve zvolené časové (prostorové, plošné ...) jednotce
- $\lambda > 0$ – jediný parametr, intenzita výskytu jevu (jak často se v průměru vyskytuje ve zvolené jednotce)

$$P(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad k = 0, 1, \dots$$
$$\mu_Y = \lambda \quad \sigma_Y^2 = \lambda$$

109

Poissonovo rozdělení $Po(\lambda)$ (2)

- je-li λ parametr (populační průměr počtu případů na jednotku), pak při počítání pravděpodobností toho, kolikrát najdeme případ na trojnásobku jednotky (trojnásobné ploše, ve trojnásobném čase ...), pak parametrem bude 3λ
- analogicky pro jiné kladné násobky
- $X \sim bi(n, \pi)$, n velké, π malé, pak pravděpodobnosti hodnot X lze aproximovat (přibližně vyjádřit) pomocí pravděpodobností hodnot $Y \sim Po(n\pi)$

110

příklady Poissonova rozdělení

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

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

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

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

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

111

příklady

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

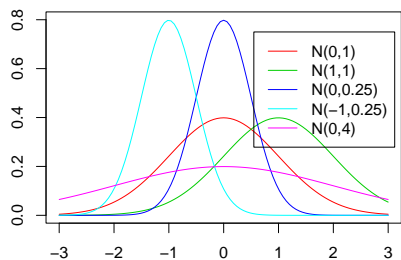
$$P(X = 12) = \binom{50}{12} \cdot 0,2^{12} \cdot 0,8^{38} = 0,103$$

- Poissonovo rozdělení $Po(50 \cdot 0,2) = Po(10)$ [\[dpois\(12,10\)\]](#)

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

112

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



- spojité rozdělení, symetrické okolo střední hodnoty μ
- maximální hodnota hustoty úměrná $1/\sigma$
- model vzniku: součet velkého počtu nepatrných příspěvků

113

- pro $X \sim N(\mu, \sigma^2)$ platí

$$\mu_X = E X = \mu \quad \sigma_X^2 = E (X - \mu_X)^2 = \sigma^2$$

$$P(|X - \mu| < 1,00\sigma) = 0,68, \text{ tj. } 68 \%$$

$$P(|X - \mu| < 1,96\sigma) = 0,95, \text{ tj. } 95 \%$$

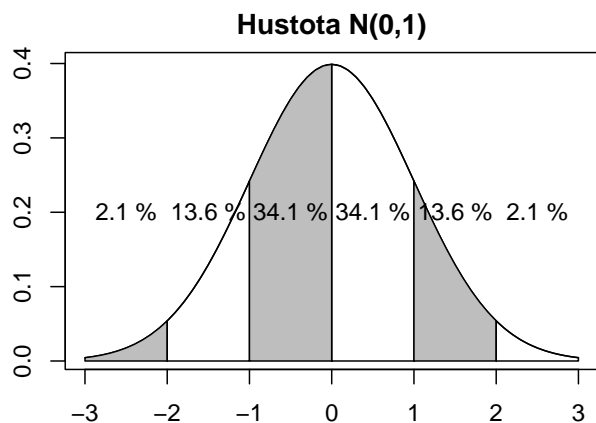
$$P(|X - \mu| < 2,00\sigma) = 0,9545, \text{ tj. } 95,45 \%$$

$$P(|X - \mu| < 3,00\sigma) = 0,9973, \text{ tj. } 99,73 \%$$

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

114

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



115

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

- tabelováno
 - hustota $\varphi(z)$
 - distribuční funkce $\Phi(z) = P(Z \leq z)$
 - **kritické hodnoty** $z(\alpha)$: $P(Z \leq z(\alpha)) = \Phi(z(\alpha)) = 1 - \alpha$

[dnorm(z)]

[pnorm(z)]

$$z(0,025) = 1,96 \text{ tj. } P(|Z| > 1,96) = 5 \%$$

$$z(0,025) = 1,96 \text{ tj. } P(Z > 1,96) = 2,5 \%$$

$$z(0,025) = 1,96 \text{ tj. } P(Z < -1,96) = 2,5 \%$$

$$z(0,005) = 2,58 \text{ tj. } P(|Z| > 2,58) = 1 \%$$

$$z(0,005) = 2,58 \text{ tj. } P(Z > 2,58) = 0,5 \%$$

$$z(0,050) = 1,64 \text{ tj. } P(|Z| > 1,64) = 10 \%$$

116

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

- u každého spojitého rozdělení je $P(X < x) = P(X \leq x)$, tedy i u Z
- $Z \sim N(0, 1)$, $a < b$, pak

$$P(a < Z < b) = \Phi(b) - \Phi(a)$$

- odvození: jevy $(Z \leq a)$ a $(a \leq Z \leq b)$ jsou neslučitelné (tvrzení nemohou platit současně), jejich sjednocením je jev $(Z \leq b)$, proto

$$P(Z \leq b) = P(Z \leq a) + P(a < Z \leq b)$$

$$\Phi(b) = \Phi(a) + P(a < Z \leq b)$$

- příklad: $P(1 < Z < 2) = \Phi(2) - \Phi(1) = 0,977 - 0,841 = 0,136$, jak bylo na obrázku

[pnorm(2)-pnorm(1)]

117

v R je výpočet snazší, protože máme k dispozici distribuční funkci se zvolenou střední hodnotou μ a zvolenou směrodatnou odchylkou σ

[pnorm(140.5,mean=136.1,sd=6.4)]

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

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

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

119

výpočet pro $X \sim N(\mu, \sigma^2)$

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

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right)$$

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

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

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

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

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

118

chování výběrového průměru

- nechť X_1, X_2, \dots, X_n jsou nezávislé náhodné veličiny s **libovolným rozdělením** se střední hodnotou μ a rozptylem σ^2 , tj. náhodný výběr z onoho rozdělení
- pro průměr z těchto veličin platí (víme, že $\mu_{X+Y} = \mu_X + \mu_Y$, $E bX = bE X$, $\sigma_{bX}^2 = b^2\sigma_X^2$, pro nezávislé X, Y také $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$)

$$\mu_{\bar{X}} = \mu \quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

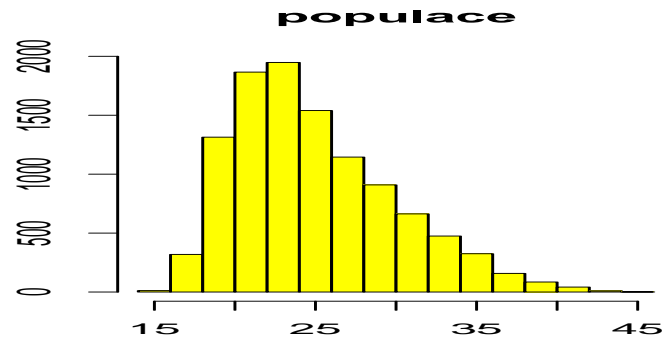
- průměr \bar{X} má tedy rozptyl n -krát menší, než jednotlivá pozorování
- **střední chyba** průměru = směrodatná odchylka průměru

$$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

120

příklad: věk matek

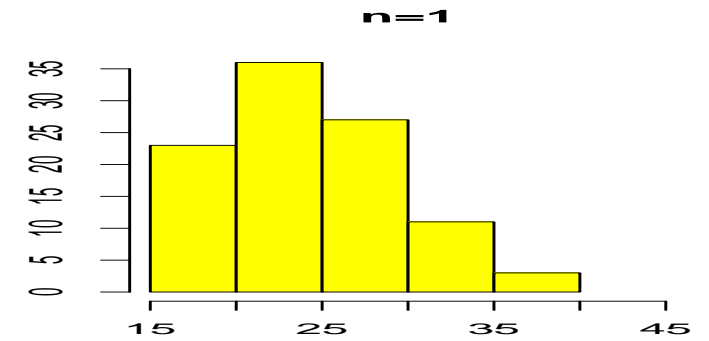
- velká populace rodičů (11 tisíc), zřejmá je nesymetrie rozdělení



121

příklad: věk matek

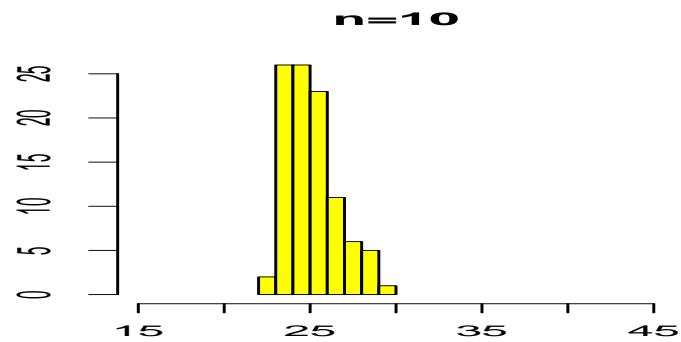
- náhodně vybráno 100 matek (průměry rozsahu $n = 1$)



122

příklad: věk matek

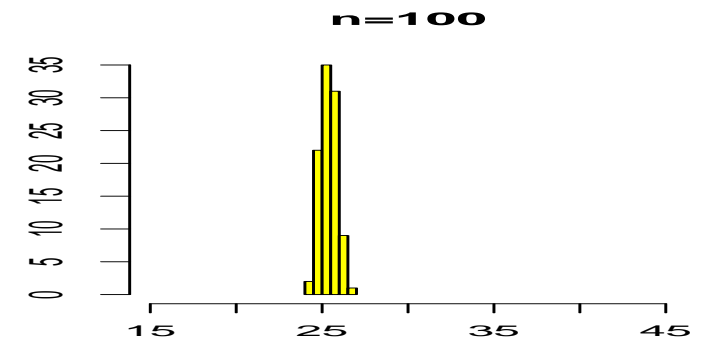
- náhodně vybráno 100 krát po $n = 10$ matkách, průměry:



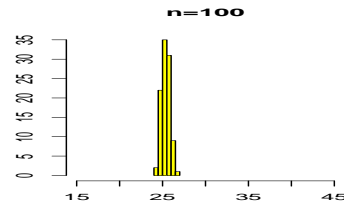
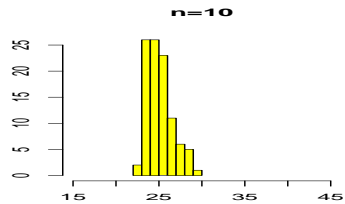
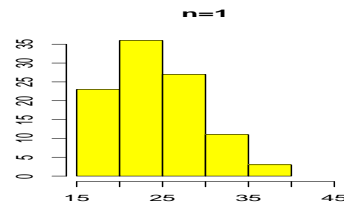
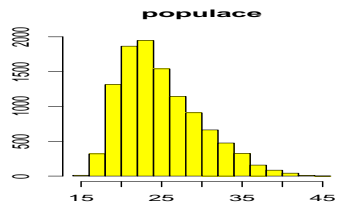
123

příklad: věk matek

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



124



125

shrnutí

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

126

výběrový průměr z normálního rozdělení

- nechť X_1, X_2, \dots, X_n jsou nezávislé náhodné veličiny s rozdělením $N(\mu, \sigma^2)$ – **náhodný výběr** z $N(\mu, \sigma^2)$
- pro průměr z nich platí

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- opět je střední chyba \bar{X} rovna $\frac{\sigma}{\sqrt{n}}$
- proto je

$$Z = \frac{\bar{X} - E\bar{X}}{\text{S.E.}(\bar{X})} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

- chování Z lze tedy popsat pomocí distribuční funkce $\Phi(z)$

127

interval spolehlivosti pro normální rozdělení (1)

- protože je $X \sim N(\mu, \sigma^2)$, platí $\bar{X} \sim N(\mu, \sigma^2/n)$ a tedy

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

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

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



128

interval spolehlivosti pro normální rozdělení (2)

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

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

129

příklad výška postavy

- studenti odhadovali výšku přednášejícího; předpokládejme, že nezávisle a nezávisle na sobě
- $n = 22$, $\bar{x} = 170,4$, $s_x = 4,032$
- $t_{21}(0,05) = 2,080$ z tabulek

$$\left(170,4 - \frac{4,032}{\sqrt{22}} \cdot 2,080; 170,4 + \frac{4,032}{\sqrt{22}} \cdot 2,080\right) \\ (170,7; 174,2)$$

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

131

interval spolehlivosti pro normální rozdělení (3)

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

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

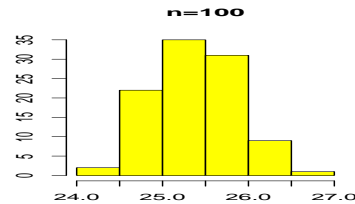
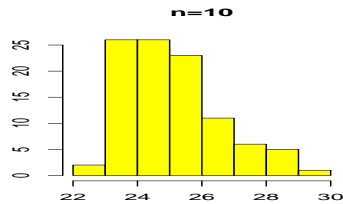
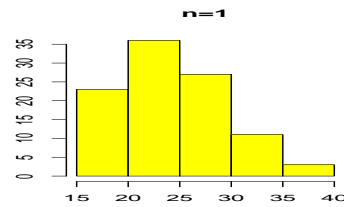
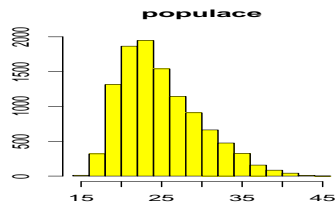
- interval spolehlivosti lze počítat i pro jiné parametry
- obecně je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

130

centrální limitní věta (CLV)

- Nechť X_1, X_2, \dots, X_n jsou nezávislé náhodné veličiny se stejným rozdělením (**nemusí** mít normální rozdělení!), se střední hodnotou μ a rozptylem $\sigma^2 > 0$. Potom pro velké n má průměr z nich rozdělení $N\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $N(n\mu, n\sigma^2)$
- prakticky: pro dost velká n má průměr normální rozdělení
- příklad: průměrný věk matek z velkých výběrů už (téměř) normální rozdělení (na následujících histogramech nejsou stejná měřítka!)
- následují stejné histogramy, ale s **nestejným měřítkem**, zajímá nás **tvar** rozdělení

132



133

příklad: věk matek

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

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

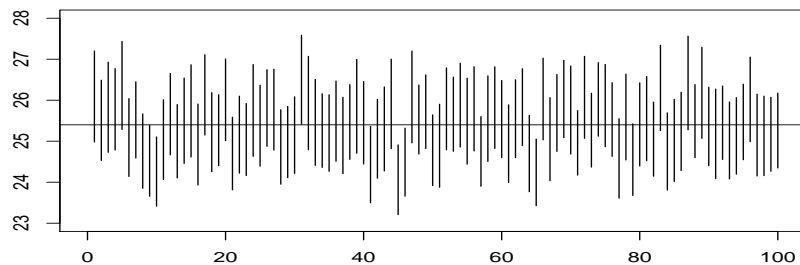
- 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)

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

- větší jistota \Leftrightarrow větší šířka

134

příklad: simulované výběry pro $n = 100$



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

135

centrální limitní věta pro četnosti

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

136

relativní četnost ve výběru

- π je **podíl** prvků s danou vlastností v populaci (např. $\pi = 45\%$)
- π – **pravděpodobnost**, že vlastnost má náhodně vybraný prvek
- Y **četnost** prvků s vlastností ve výběru rozsahu n , $Y \sim \text{bi}(n, \pi)$
- $f = \frac{Y}{n}$ **relativní četnost** prvků s danou vlastností ve výběru
- relativní četnost je průměr nula-jedničkové veličiny – pro velké n má přibližně normální rozdělení
- nula-jedničková veličina má rozptyl $\pi(1 - \pi)$, tedy relativní četnost (je to průměr) má rozptyl $\frac{\pi(1 - \pi)}{n}$
- CLV $\Rightarrow \boxed{f \sim N(\pi, \pi(1 - \pi)/n)}$, $\boxed{Y \sim N(n\pi, n\pi(1 - \pi))}$

137

interval spolehlivosti pro podíl π

- střední chyba relativní četnosti = směrodatná odchylka relativní četnosti = odmocnina z rozptylu je tedy $\sqrt{\frac{\pi(1 - \pi)}{n}}$
- pravděpodobnost π neznáme, odhadneme ji pomocí relativní četnosti f
- odtud je 95% interval spolehlivosti pro π
$$\left(f - 1,96 \cdot \sqrt{\frac{f(1 - f)}{n}}; f + 1,96 \cdot \sqrt{\frac{f(1 - f)}{n}} \right)$$
- existuje přesnější (pracnější) postup

139

příklad na aproximaci binomického rozdělení normálním

- za zkušenosti je známo, že mezi uchazeči o studium matematiky na MFF bývá 45 % dívek
- s jakou pravděpodobností bude při 500 přihláškách počet dívek mezi 200 a 220 (včetně)?
- $X \sim \text{bi}(500, 0,45)$ má $\mu_X = 500 \cdot 0,45 = 225$, $\sigma_X^2 = 500 \cdot 0,45 \cdot 0,55 = 123,75$, tedy $\sigma_X = 11,1$

$$P(200 \leq X \leq 220) = \Phi\left(\frac{220,5 - 225}{11,1}\right) - \Phi\left(\frac{199,5 - 225}{11,1}\right) = 0,343 - 0,011 = 0,332$$

- hledaná pravděpodobnost je přibližně 33,2 % (přesně 33,3 %)

138

příklad: hody s hrací kostkou

- odhadujeme pravděpodobnost šestky

- kostka A: $n = 100, n_A = 17, f_A = 0,17$

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

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

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

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

140

proč testování hypotéz

- nelze bezpečně poznat, že kostka B je falešná nebo že kostka A není falešná
- intervaly spolehlivosti určily rozmezí, kde by skutečná pravděpodobnost šestky měla být, jejich spolehlivost je velká, ale omezená
- znamená něco, když $1/6$ neleží v 95% intervalu spolehlivosti?
- musíme připustit, že jsme mohli mít smůlu, že se v našich pokusech náhodou realizovaly málo pravděpodobné možnosti, přestože k takové smůle dochází jen zřídka

141

testování hypotéz (2)

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

143

testování hypotéz (1)

- **(nulová) hypotéza** H_0 : – zjednodušuje situaci, zpravidla se jí snažíme vyvrátit, abychom věcně něco prokázali
- **alternativa** H_1 : (**alternativní hypotéza**) – opak nulové hypotézy, zpravidla to, co chceme věcně dokázat
- možná rozhodnutí
 - **zamítnout** H_0 pokud naše data svědčí proti H_0
 - **nezamítnout** H_0 (přijmout H_0) pokud *není dost důvodů* H_0 zamítnout
- nelze zaručit bezchybnost rozhodnutí

142

schéma testování hypotéz

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 ($pst \beta$)

144

postup při rozhodování

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

145

příklad přesné volby kritického oboru

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

- podmínku $P(Y \geq y_0) \leq 0,05$ splňuje $y_0 = 24$
- padne-li ve 100 nezávislých hodech kostkou aspoň 24 šestek, budeme na **5% hladině zamítnout 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**, což neznamená, že bychom hypotézu prokázali
- na kostce B nám padlo 41 šestek, hypotézu **zamítáme**
- pro $\alpha = 10\%$ bychom zvolili $y_0 = 22$

147

příklad: padá na kostce šestka příliš často?

- chceme na 5% hladině prokázat, že pravděpodobnost šestky na dané kostce je větší, než by měla být (tj. větší než 1/6)
- $H_0 : P(\text{padne šestka}) = 1/6 \quad (\pi = \pi_0)$
- $H_1 : P(\text{padne šestka}) > 1/6 \quad (\pi > \pi_0)$
- provedeme $n = 100$ pokusů, Y počet šestek
- co svědčí pro neplatnost hypotézy?
„šestka padá mnohem častěji, než by měla za H_0 “
- hypotézu budeme zamítnat, když $Y \geq y_0$ (tvar krit. oboru)
- za platnosti H_0 má počet šestek Y rozdělení $bi(n, 1/6)$
- y_0 zvolit tak, aby za hypotézy bylo $P(Y \geq y_0) \leq \alpha$

146

příklad: volba kritického oboru (přibližně)

- použijme přibližné tvrzení: za $H_0 \quad Y \sim N(n\pi_0, n\pi_0(1-\pi_0))$, potom

$$\begin{aligned}
 P(Y \geq y_0) &= 1 - P(Y < y_0) = 1 - P(Y \leq y_0 - 0,5) \\
 &= 1 - P\left(\frac{Y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} < \frac{y_0 - 0,5 - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}\right) \\
 &\doteq 1 - \Phi\left(\frac{y_0 - 0,5 - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}\right) = \alpha (= 0,05)
 \end{aligned}$$

- tabulka kritických hodnot dá $z(\alpha)$, musí platit $z(\alpha) = \frac{y_0 - 0,5 - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$
tedy
 $y_0 = n\pi_0 + 0,5 + z(\alpha)\sqrt{n\pi_0(1-\pi_0)}$, v našem příkladu
 $y_0 = 100/6 + 1/2 + 1,645 \cdot \sqrt{500/36} = 23,3 \doteq 23$

148

p -hodnota

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

149

příklad rozhodování pomocí p -hodnoty

- snažíme se prokázat, že šestka padá příliš často
- padlo nám $Y = 17$, proto (vzorec pro psti binomického rozdělení)

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

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

150

příklad: kostka a oboustranná alternativa

- chceme ověřit, zda je kostka v pořádku
- pokusíme se prokázat, že šestka padla příliš často nebo příliš zřídka
- $H_0 : P(\text{padne šestka}) = 1/6 \quad (\pi = \pi_0)$
- $H_1 : P(\text{padne šestka}) \neq 1/6 \quad (\pi \neq \pi_0)$
- je to **oboustranná alternativa** (na rozdíl od jednostranné)
- *proti* hypotéze svědčí malé *nebo* velké hodnoty Y
- pst chyby 1. druhu α rozdělíme na dvě poloviny: pro příliš malé a příliš velké Y

151

příklad: kostka, oboustranná alternativa

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

- H_0 zamítneme, když bude $Y \leq 9$ *nebo* když bude $Y \geq 25$
- skutečná pst chyby 1. druhu bude $0,021 + 0,022 = 0,043$
- `[pbinom(9,100,1/6)+(1-pbinom(24,100,1/6))]`
(nezapomeňte, že hodnota distribuční funkce je $P(X \leq x)$)
- hodnoty v rozmezí 10 až 24 (včetně obou mezí) nnesvědčí proti H_0

152

oboustranná alternativa přibližně

- $H_0 : P(\text{padne šestka}) = 1/6 \quad (\pi = \pi_0)$
 $H_1 : P(\text{padne šestka}) \neq 1/6 \quad (\pi \neq \pi_0)$
- proti alternativě svědčí Y hodně daleko od $\mu_Y = n\pi_0$ (počítáme za platnosti hypotézy), tj. rel. četnost $f = Y/n$ daleko od π_0 :

$$P\left(\left|\frac{Y - n\pi_0}{\sqrt{n\pi_0(1 - \pi_0)}}\right| \geq z(\alpha/2)\right) = \alpha$$

- zamítáme tedy, je-li [prop.test(9,100,1/6)]

$$Y \leq n\pi_0 - z(\alpha/2)\sqrt{n\pi_0(1 - \pi_0)} \doteq 9,36$$

nebo

$$Y \geq n\pi_0 + z(\alpha/2)\sqrt{n\pi_0(1 - \pi_0)} \doteq 23,97$$

153

výšky desetiletých hochů

- alternativě nasvědčují průměry \bar{X} o hodně větší než $\mu_0 = 136,1$
- kritický obor: $\bar{X} \geq x_0$, kde x_0 je zvoleno tak, aby za platnosti hypotézy tato nerovnost nastala s pstí nejvýš 5 %
- platí (za platnosti hypotézy)

$$\bar{X} \sim N(136,1, 6,4^2/15) \Rightarrow Z = \frac{\bar{X} - 136,1}{\text{S.E.}(\bar{X})} = \frac{\bar{X} - 136,1}{6,4} \sqrt{15} \sim N(0, 1)$$

- proto hypotézu zamítáme, je-li $Z \geq z(0,05) = 1,645$
- v našem příkladu je $Z = \frac{139,13 - 136,1}{6,4} \sqrt{15} = 1,82 > 1,645$, takže na 5% hladině hypotézu **zamítáme ve prospěch jednostranné alternativy, že populační průměr za deset roků vzrostl**

155

příklad: výšky desetiletých hochů

- velký výběr v roce 1951 dal průměr 136,1 cm, rozptyl $6,4^2 \text{ cm}^2$ (dál interpretujeme jako známé konstanty)
- v roce 1961 naměřeno v náhodném výběru $n = 15$ hodnot s průměrem $\bar{X} = 139,13 \text{ cm}$ (lze předpokládat nezměněný rozptyl)
- prokázali jsme na 5% hladině představu, že desetiletí hoši jsou (co do populačního průměru) v roce 1961 *větší* než desetiletí hoši v roce 1951?
- hypotéza $H_0 : \mu = \mu_0 = 136,1$ (stejný postup by byl pro $\mu \leq 136,1$,)
- alternativa $H_1 : \mu > \mu_0 = 136,1$

154

obecně (jednostranná alternativa)

- $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$, nezávislé
- $H_0 : \mu = \mu_0 \quad H_1 : \mu > \mu_0$
- kritický obor: $\bar{X} \geq x_0$, kde x_0 je zvoleno tak, aby za platnosti hypotézy bylo překročeno s pstí nejvýš 5 %
- platí (za platnosti hypotézy)

$$\bar{X} \sim N(\mu_0, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \sim N(0, 1)$$

- proto hypotézu zamítáme na hladině α , je-li $Z \geq z(\alpha)$
- je-li zvoleno $H_1 : \mu < \mu_0$, podobně jako výše hypotézu H_0 zamítáme na hladině α , je-li $Z \leq -z(\alpha)$

156

výpočet p -hodnoty pro výšky desetiletých hochů

- uvažujeme jednostrannou alternativu $H_1 : \mu > \mu_0 = 136,1$
- skutečně platnou střední hodnotu uvedeme jako dolní index u P
- za H_0 je $Z = \frac{\bar{X} - 136,1}{6,4} \sqrt{15} \sim N(0, 1)$
[1-pnorm(139.1, 136.1, 6.4/sqrt(15))]

$$\begin{aligned} p &= P_{136,1}(\bar{X} \geq 139,1) \quad (\text{protože nám vyšlo } \bar{X} = 139,1) \\ &= P_{136,1}\left(\frac{\bar{X} - 136,1}{6,4} \sqrt{15} \geq \frac{139,13 - 136,1}{6,4} \sqrt{15}\right) \\ &= P(Z \geq 1,82) = 1 - \Phi(1,82) = 1 - 0,965 = 0,035 < 0,05 \end{aligned}$$

- na 5% hladině jsme zamítli hypotézu ve prospěch jednostranné alternativy (kterou jsme zvolili předem, bez znalosti dat!)
- prokázali jsme na 5% hladině vzrůst populačního průměru

157

síla testu pro $\mu = 140$

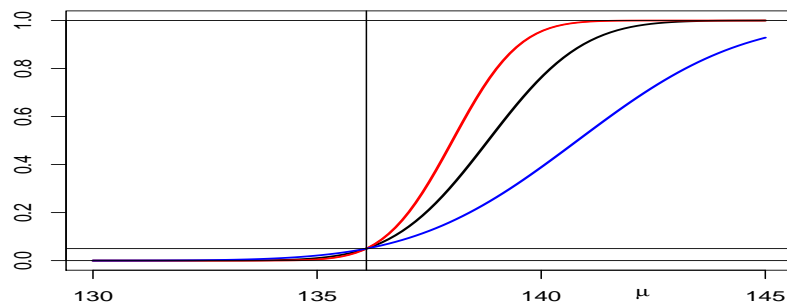
- síla testu = $pst(\text{zamítnout hypotézu, když tato neplatí})$
- zkusme předpokládat, že ve skutečnosti platí $\mu = 140$, připomeňme také kritickou hodnotu $z(0,05) = 1,645$

$$\begin{aligned} 1 - \beta(140) &= P_{140}\left(\frac{\bar{X} - 136,1}{6,4} \sqrt{15} \geq 1,645\right) \\ &= P_{140}\left(\frac{\bar{X} - 140}{6,4} \sqrt{15} + \frac{140 - 136,1}{6,4} \sqrt{15} \geq 1,645\right) \\ &= P\left(Z \geq 1,645 - \frac{140 - 136,1}{6,4} \sqrt{15}\right) = P(Z \geq -0,715) \\ &= 1 - \Phi(-0,715) = 1 - 0,237 = 0,763 \end{aligned}$$

- je-li opravdu $\mu = 140$, pak máme 76% naději, že to odhalíme

158

síla testu v závislosti na μ , $n = 15$ (30, 5)



159

obecně pro oboustrannou alternativu

- $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$
- $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$
- kritický obor: \bar{X} je příliš daleko od μ_0 ,
- platí (za platnosti hypotézy)

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right) \Rightarrow Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \sim N(0, 1)$$

- protože hladinu musíme rozdělit na dvě části (pro $\bar{X} \ll \mu_0$, tj. $Z \ll 0$ nebo pro $\bar{X} \gg \mu_0$, tj. $Z \gg 0$) hypotézu zamítáme na hladině α , je-li $|Z| \geq z(\alpha/2)$

160

shrnutí: $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, nezávislé

- předpokládáme, že $\sigma > 0$ známe
- $H_0 : \mu = \mu_0$ (μ_0 známá konstanta)

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

- kdy hypotézu H_0 zamítáme (kritický obor):
 - $H_1 : \mu \neq \mu_0$ (oboustranná alternativa) $|Z| \geq z(\alpha/2)$
 - $H_1 : \mu > \mu_0$ (jednostranná alternativa) $Z \geq z(\alpha)$
 - $H_1 : \mu < \mu_0$ (jednostranná alternativa) $Z \leq -z(\alpha)$

161

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$, nezávislé, σ^2 neznámé

- neznámé $\sigma > 0$ odhadneme pomocí $s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$
- $H_0 : \mu = \mu_0$ (μ_0 známá konstanta)

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

- kdy hypotézu H_0 zamítáme (kritický obor):
 - $H_1 : \mu \neq \mu_0$ (oboustranná alternativa) $|T| \geq t_{n-1}(\alpha)$
 - $H_1 : \mu > \mu_0$ (jednostranná alternativa) $T \geq t_{n-1}(2\alpha)$
 - $H_1 : \mu < \mu_0$ (jednostranná alternativa) $T \leq -t_{n-1}(2\alpha)$

162

souvislost s intervalem spolehlivosti

- připomeňme interval spolehlivosti pro μ

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

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

což lze přepsat jako

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

- $H_0 : \mu = \mu_0$ tedy **nezamítáme** na hladině α při oboustranné alternativě, právě když μ_0 leží v $100(1 - \alpha)\%$ intervalu spolehlivosti
- **interval spolehlivosti tedy obsahuje takové hodnoty μ_0 , které bychom jako hypotézu nezamítli**

163

výšky desetiletých hochů (σ^2 neznámé)

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

$$s_x = \sqrt{\frac{1}{15-1} ((130-139,13)^2 + \dots + (141-139,13)^2)} = \sqrt{42,98} = 6,56$$

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

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

164

nová úloha: porovnání dvou populací

- liší se desetileté dívky výškou postavy od desetiletých hochů?
- lze předpokládat, že výšky hochů

$$X_i \sim N(\mu_1, \sigma^2), \quad i = 1, \dots, n_1$$

- lze předpokládat, že výšky dívek

$$Y_i \sim N(\mu_2, \sigma^2), \quad i = 1, \dots, n_2$$

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

165

porovnání středních hodnot nezávislých výběrů

- zřejmě $H_0: \mu_1 = \mu_2$ (není rozdíl: $\mu_1 - \mu_2 = 0$ **nulová** hypotéza)
- možné alternativy
 - $H_1: \mu_1 \neq \mu_2$ (není-li důvod k jednostranné alternativě)
 - $H_1: \mu_1 > \mu_2$ (bylo cílem dokázat, že hoši jsou větší dívek)
 - $H_1: \mu_1 < \mu_2$ (bylo cílem dokázat, že hoši jsou menší dívek)
- 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 $\mu_1 - \mu_2$

166

porovnání středních hodnot nezáv. výběrů (2)

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

$$\begin{aligned} s^2 &= \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 \right) \\ &= \frac{n_1 - 1}{n_1 + n_2 - 2} s_x^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} s_y^2 \end{aligned}$$

(vážený průměr odhadů rozptylu v obou výběrech)

- výška desetiletých dětí: $n_1 = 15$, $n_2 = 12$, $\bar{X} = 139,13$, $\bar{Y} = 140,83$, $s_x^2 = 42,98$, $s_y^2 = 33,79$, tudíž

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

167

kritické obory

- o hypotéze $H_0: \mu_1 = \mu_2$ se rozhoduje pomocí
$$T = \frac{\bar{X} - \bar{Y}}{\widehat{\text{S.Ě.}}(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$
- $H_1: \mu_1 \neq \mu_2$ zamítáme pokud $|T| \geq t_{n_1+n_2-2}(\alpha)$
- $H_1: \mu_1 > \mu_2$ zamítáme pokud $T \geq t_{n_1+n_2-2}(2\alpha)$
- $H_1: \mu_1 < \mu_2$ zamítáme pokud $T \leq -t_{n_1+n_2-2}(2\alpha)$
- výšky desetiletých: $T = -0,70 \Rightarrow |-0,70| < 2,06 = t_{15+12-2}(0,05)$
- na 5% hladině jsme **neprokázali** rozdíl mezi výškami desetiletých hochů a dívek ($p = 48,8\%$) [[t.test\(vyska~Divka,var.equal=TRUE\)](#)]

168

souvislost s intervalem spolehlivosti

- $\mu_1 - \mu_2 = \delta$ o kolik se liší populační průměrné výšky
- odhadem pro δ je $d = \bar{X} - \bar{Y} = -1,7$
- interval spolehlivosti pro rozdíl δ je

$$(\bar{X} - \bar{Y}) - \widehat{S.E.} \cdot (\bar{X} - \bar{Y}) \cdot t_{n_1+n_2-2}(\alpha) < \delta < (\bar{X} - \bar{Y}) + \widehat{S.E.} \cdot (\bar{X} - \bar{Y}) \cdot t_{n_1+n_2-2}(\alpha)$$
 H_0 zamítáme právě tehdy, když nula **není** v int. spol. pro δ
- při porovnání výšek hochů a dívek je 95% interval pro δ

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

(-6,7; 3,3)

169

provedení v MS Excelu (stejně rozptyly)

přednáška	Excel	Soubor 1	Soubor 2
průměr	Stř. hodnota	139.133	140.833
rozptyl	Rozptyl	42.981	33.788
rozsah výběru	Pozorování	15	12
spol. odhad rozpt.	Společný rozptyl	38.936	
$H_0 : \mu_1 - \mu_2 =$	Hyp. rozdíl stř. hodnot	0	
stupně vol.	Rozdíl	25	
T	t stat	-0.733	
p jednostr. testu	P(T<=t) (1)	0.244	jen někdy!
$t_{n_1+n_2-2}(2\alpha)$	t krit (1)	1.708	
p oboustr. testu	P(T<=t) (2)	0.488	
$t_{n_1+n_2-2}(\alpha)$	t krit (2)	2.060	

při oboustranné alternativě nelze nulovou hypotézu zamítnout

170

problém nestejných rozptylů

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

$$\text{buď } F = \frac{s_x^2}{s_y^2} \geq F_{n_1-1, n_2-1}(\alpha/2) \text{ nebo } \frac{1}{F} = \frac{s_y^2}{s_x^2} \geq F_{n_2-1, n_1-1}(\alpha/2)$$

- vlastně se větší odhad rozptylu dělí menším odhadem, k tomu se musí zvolit správné pořadí stupňů volnosti a hladina
- příklad výšky desetiletých dětí: $F = \frac{42,98}{38,94} = 1,27 < F_{14,11}(0,025) = 3,36$
- [var.test(vyska~Divka)]

171

MS Excel: Dvouvýběrový F-test pro rozptyl

přednáška	Excel	Soubor 1	Soubor 2
průměr	Stř. hodnota	139.13	140.83
rozptyl	Rozptyl	42.98	33.79
rozsah	Pozorování	15	12
stupně vol.	Rozdíl	14	11
F	F	1.27	
p	P(F <= f) (1)	0.349	
	F krit (1)	2.739	

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

172

dvouvýběrový t -test při nesterjných rozptylech [t.test(vyska~Divka)]

- není-li udržitelny předpoklad o stejnyh rozptylech, lze použit přibližný t -test (Welchův, s jiným odhadem S.E. $(\bar{X} - \bar{Y})$)

$$T = \frac{\bar{X} - \bar{Y}}{\widehat{\text{S.E.}}(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}}$$

- kde $s_{\bar{X}-\bar{Y}}$ je střední chyba $\bar{X} - \bar{Y}$

$$s_{\bar{X}-\bar{Y}} = \sqrt{v_x + v_y} \quad v_x = s_x^2/n_1 \quad v_y = s_y^2/n_2$$

- H_0 se zamítá, je-li $|T| \geq t_f(\alpha)$, kde $f = \frac{s_{\bar{X}-\bar{Y}}^4}{\frac{v_1^2}{n_1-1} + \frac{v_2^2}{n_2-1}}$

- náš příklad $T = -0,713$, $f = 24,69$, $t_f(0,05) = 2,061$, $p = 0,482$

173

souvislost s bodově biseriálním korel. koef.

- bodově biseriální korelační koeficient (str. 76) vypovídá o síle závislosti mezi spojitou a nula-jedničkovou veličinou (v současném označení)

$$r_{\text{bis}} = \frac{\bar{X} - \bar{Y}}{s_{\text{all}}} \sqrt{\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)}}$$

- pomocí r_{bis} lze vyjádřit testovou statistiku dvouvýběrového t -testu

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

- náš příklad $r_{\text{bis}} = -0,139$

175

provedení v MS Excelu (nesterjné rozptyly)

		Soubor 1	Soubor 2
průměr	Stř. hodnota	139.133	140.833
rozptyl	Rozptyl	42.981	33.788
rozsah	Pozorování	15	12
$H_0 : \mu_1 - \mu_2 =$	Hyp. rozdíl stř. hodnot	0	
stupně vol. f	Rozdíl	25	
T	t stat	-0.713	
p jednostr. testu	P(T<=t) (1)	0.241	
$t_f(2\alpha)$	t krit (1)	1.708	
p oboustr. testu	P(T<=t) (2)	0.482	
$t_f(\alpha)$	t krit (2)	2.060	

při oboustranné alternativě nelze nulovou hypotézu zamítnout

174

Mannův-Whitneyův (Wilcoxonův) test

- co když nelze předpokládat normální rozdělení?
- necht X_1, \dots, X_{n_1} a Y_1, \dots, Y_{n_2} jsou **nezávislé** výběry ze spojitého rozdělení (například věk matek, střední délka života mužů při narození ve dvou skupinách zemí, potratovost ...)
- postup založen na pořadí bez ohledu na výběr
- idea: kdyby nebyl mezi populacemi rozdíl, byla by takto zjištěná průměrná pořadí v obou výběrech podobná

176

příklad: potratovost (Čechy vers. Morava)

kraj	Pha	Stč	Jč	Pl	KV	Ús	Lb
potratovost	4.03	4.02	4.11	4.70	5.65	5.80	4.98
pořadí	7	6	8	10	12	13	11
kraj	HK	Par	Vys	JM	Ol	Zl	MS
potratovost	4.33	3.38	3.57	3.70	3.65	3.42	3.87
pořadí	9	1		4	3	2	5

- H_0 : shoda populací (zejm. mediánů), H_1 : neshoda
- nejasné, kam patří kraj Vysočina; vynecháme jej
- průměrné pořadí českých krajů: $77/9=8,56$
 $W_1=7+6+8+10+12+13+11+9+1=77$
- průměrné pořadí moravských krajů: $14/4=3,5$
 $W_2=4+3+2+5=14$

177

přibližné rozhodování (n_1, n_2 desítky)

- W_1, W_2 součty pořadí, použitím centrální limitní věty

$$Z = \frac{W_1 - n_1(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

- za hypotézy (není rozdíl mezi populacemi) je $Z \sim N(0, 1)$
- hypotézu zamítáme, je-li $|Z| \geq z(\alpha/2)$

- náš příklad:

[wilcox.test(potr~Cechy)]

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

- na 5% hladině jsme prokázali rozdíl

178

přesný výpočet p -hodnoty

- zajímá nás, nakolik je náš výsledek ($W_1 = 77, W_2 = 14$) výjimečný
- máme celkem $n_1 + n_2 = 13$ pozorování, čtyři z nich (Morava) lze vybrat celkem $\binom{13}{4} = 715$ způsoby
- kolik z nich vede k tak extrémně nestejným průměrným pořadím?
- budeme hledat, kolik čtveřic označených za moravské by dalo v součtu nejvýš 14, jak nám doopravdy vyšlo
- vždy platí $W_1 + W_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2 = 91$
(součet čísel $1 + 2 + \dots + n_1 + n_2$)
- stačí zabývat se jedinou ze statistik W_1, W_2 , zpravidla tou pro menší výběr

179

přehled možných čtveřic,

v nichž je součet pořadí nejvýš 14

1	1	1	1	1	1	1	1	1	1	1	2	1	1
2	2	2	2	2	2	3	2	2	2	3	3	2	2
3	3	3	4	3	4	4	3	4	5	4	4	3	4
4	5	6	5	7	6	5	8	7	6	6	5	9	8
10	11	12	12	13	13	13	14	14	14	14	14	15	15

- nejvýš 14 mohl být součet pořadí za platnosti hypotézy s pravděpodobností $p_1 = 12/715 = 0,01678$
- musíme vzít v úvahu také situaci, kdy by byla na Moravě velká pořadí, p -hodnotu nutno zdvojnásobit, tedy $p = 24/715 = 3,4 \%$

180

párové testy

- předpoklad **nezávislosti** porovnávaných výběrů musí opravdu být splněn, jinak dostaneme nesmysl
- typické porušení předpokladu nezávislosti je u párových dat
 - měření na stejných objektech ve dvou různých časech
 - měření na stejných objektech před zásahem a po něm (ošetření)
 - měření na rodičích
- postup
 - spočítají se a hodnotí rozdíly (změny)
 - přejde se k úloze s jediným výběrem
 - mají-li rozdíly normální rozdělení, pak párový t -test

181

příklad: výška rodičů

- rozhodnout o tvrzení, že populační průměr výšek otců je o 10 cm větší než populační průměr výšek matek
- otcové: $\bar{Y} = 179,26$, $s_Y = 6,78$, $n_1 = 99$
matky: $\bar{Z} = 166,97$, $s_Z = 6,11$, $n_2 = 99$
- otcové jsou (ve výběru) v průměru o $\bar{Y} - \bar{Z} = 12,29$ cm vyšší směrodatná odchylka **rozdílů** je 8,14 (méně, než kdyby byly výšky rodičů nezávislé . . . $6,78^2 + 6,11^2 = 9,13^2$)
střední chyba rozdílu průměrů je $8,14/\sqrt{99} = 0,819$
- rozhodneme podle statistiky `[t.test(vyska.o-vyska.m,mu=10)]`

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

182

párový t -test:

- nechť $(Y_1, Z_1) \dots, (Y_n, Z_n)$ nezávislé dvojice, $X_i = Y_i - Z_i$
- nechť $X_i \sim N(\mu, \sigma^2)$
- neznámé $\sigma > 0$ odhadneme pomocí $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$
- $H_0 : \mu = \mu_0$ (μ_0 známá konstanta, zpravidla 0)

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

- hypotézu H_0 zamítáme (kritický obor):
 - $H_1 : \mu \neq \mu_0$ (oboustranná alternativa) $|T| \geq t_{n-1}(\alpha)$
 - $H_1 : \mu > \mu_0$ (jednostranná alternativa) $T \geq t_{n-1}(2\alpha)$
 - $H_1 : \mu < \mu_0$ (jednostranná alternativa) $T \leq -t_{n-1}(2\alpha)$

183

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

Y_i	24.7	25.7	31.6	24.3	26.8	30.6	21.1	23.5	26.9	22.5	23.1	24.9
Z_i	23.1	23.6	27.9	22.2	23.4	27.9	21.5	26.0	24.3	23.9	21.2	25.7
X_i	1.6	2.1	3.7	2.1	3.4	2.7	-0.4	-2.5	2.6	-1.4	1.9	-0.8
R_i^+	4	6	12	7	11	10	1	8	9	3	5	2

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

184

párový Wilcoxonův (Wilcoxon signed rank) test

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

185

párový znaménkový (sign) test

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

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

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

187

rozhodování `[wilcox.test(potr00-potr01, alternative="greater")]`

- na základě centrální limitní věty lze použít

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

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

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

186

poznámky `[binom.test(sum(potr00>potr01), 12, alt="gr")]`

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

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

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

$$Z_{\text{Yates}} = \frac{|Y - n/2| - 1/2}{\sqrt{n/4}} \text{sign}(Y - n/2) = \frac{|2Y - n| - 1}{\sqrt{n}} \text{sign}(2Y - n)$$

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

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

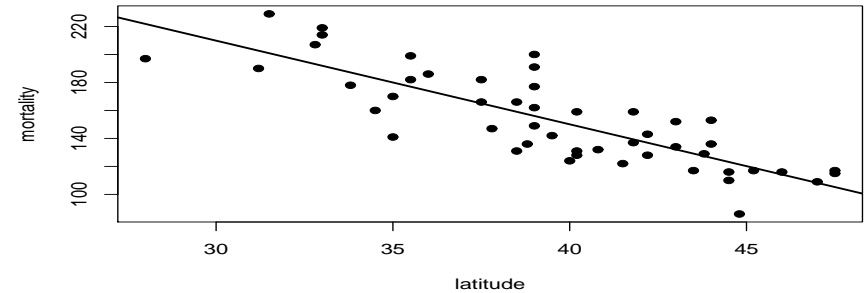
188

Regrese

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

189

příklad: souvisí úmrtnost se zeměpisnou šířkou?



- úmrtnost na melanom na 10 000 000 obyvatel v státech USA

190

regresní přímka

- chování Y (úmrtnost, mortality) co nejlépe (nejvíce) vysvětlit lineární závislostí na x (zeměpisná šířka, latitude)
- (naše představa, předpoklad:) každé zem. šířce odpovídá jakási střední úmrtnost, ta závisí na zeměpisné šířce lineárně

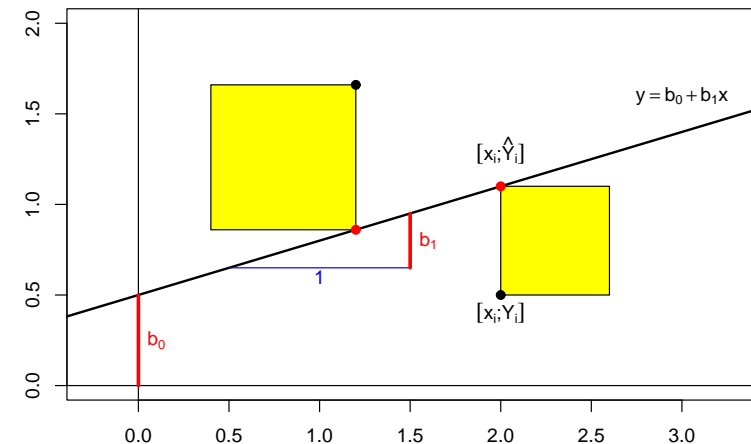
$$E Y_i = \beta_0 + \beta_1 x_i, \quad i = 1, \dots, n$$

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

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

- výsledné minimum (pro b_0, b_1) – **reziduální součet čtverců** S_e

191



192

náš příklad

[summary(lm(mortality~latitude))]

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

- odhad závislosti: $\text{est}(\text{mortality}) = 389,2 - 5,98 \text{ latitude}$
- s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 6 osob na 10 000 000 obyvatel
- na rovníku by úmrtnost měla být 389 jednotek, ale je to extrapolace mimo rozmezí známých hodnot – velmi nejisté
- závislost je průkazná, neboť v řádku pro x (latitude) je $p < 0,001$

193

obecně

- odhadovaná závislost $y = \beta_0 + \beta_1 x$, odhadnutá $y = b_0 + b_1 x$
- závislost na x prokazujeme testováním hypotézy $H_0: \beta_1 = 0$ (pak y pro všechna x stejné: $y = \beta_0$) proti oboustranné alternativě pomocí

$$T = \frac{b_1}{\text{S.E.}(b_1)} = \frac{b_1}{s} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{zamítáme pokud } |T| \geq t_{n-2}(\alpha)$$

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

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

194

náš příklad a tabulka analýzy rozptylu [anova(lm(mortality~latitude))]

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

- kolísání úmrtnosti vysvětlíme závislostí z 68 %, neboť je

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

- na 30. stupni očekáváme úmrtnost $389,19 - 5,98 \cdot 30 = 209,86$,
na 40. stupni očekáváme úmrtnost $389,19 - 5,98 \cdot 40 = 150,08$

195

můžeme predikci zlepšit?

[summary(lm(mortality~latitude+longitude))]

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

- není průkazné, že by koeficient u longitude byl nenulový (nezamítneme hypotézu, že koeficient je nulový)
- koeficient determinace $R^2 = 0,684$ (původně 0,680)

196

podrobnější rozbor – vliv oceánu

jen vnitrozemské státy ($R^2 = 59,6 \%$):

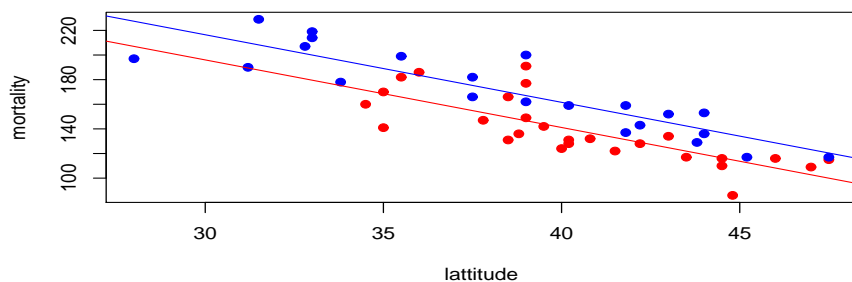
koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,55	36,70	9,82	<0,001
latitude	- 5,485	0,904	- 6,07	<0,001

jen přímořské státy ($R^2 = 78,6 \%$):

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

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

197



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

199

můžeme predikci zlepšit? [summary(lm(mortality~ocean+latitude))]

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

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

198

pozor na interpretaci odhadů (na dalším příkladu)

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

200

regrese v MS Excelu 2000, 2003

	Excel 2000	označení
absolutní člen odhad	Hranice	b_0
střední chyba odhadu koeficient	Koeficienty	b_i
(mnohonásobné) korelace koeficient determinace	Chyba střední hodnoty	S.E.(b_j)
adjustovaný koef. det. resid. směr. odchylka	Násobné R	$\sqrt{R^2}$
počet pozorování	Hodnota spolehlivosti R	R^2
počet st. volnosti	Nastavená hodnota spol. R	R_{adj}^2
	Chyba střední hodnoty	s
	Pozorování	n
	Rozdíl	

201

obecné předpoklady pro regresní model

- **tvar závislosti:** známe jak vysvětlovaná veličina závisí na vysvětlujících
- **homoskedasticita:** pro všechny kombinace hodnot vysvětlujících veličin je rozptyl vysvětlované veličiny konstantní
- **nezávislost:** náhodné složky vysvětlovaných veličin jsou nezávislé
- **normalita:** náhodná složka má normální rozdělení
- předpoklady lze ověřovat (regresní diagnostika)
- někdy pomohou transformace

203

regrese v MS Excelu 2000, 2003

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

202

použití reziduí

- pomocí regrese hledáme model pro závislost nebo predikci (střední hodnoty) příštích pozorování
- celkovou schopnost vysvětlit chování (variabilitu) závisle proměnné hodnotíme pomocí **koeficientu determinace**

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

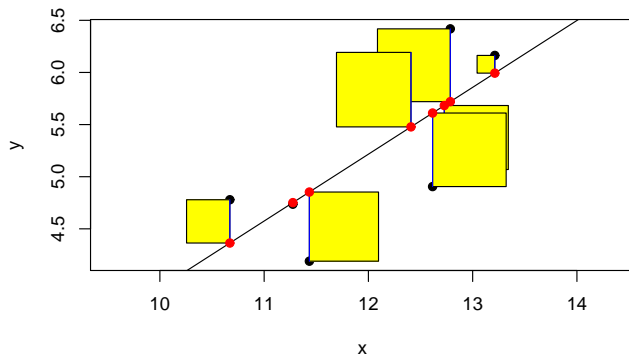
- v čitateli posledního výrazu **rezidua**

$$u_i = Y_i - \hat{Y}_i$$

(rozdíl **naměřená** - **vyrovnaná** hodnota vysvětlované proměnné)

- rezidua lze použít k hodnocení (diagnostice) regrese

204



Y_i, \hat{Y}_i

205

diagnostika pomocí reziduí

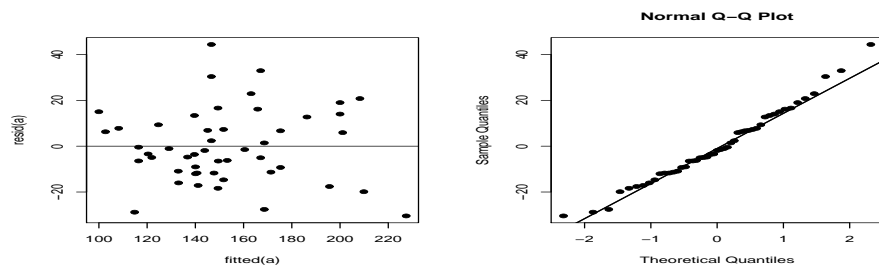
- grafické znázornění bodů $[u_i, \hat{Y}_i]$ nebo $[u_i, x_i]$ (k ověření konstantního rozptylu či tvaru závislosti)
- histogram reziduí nebo normální diagram (k ověření normálního rozdělení), pomocí u_i nebo v_i (upravená rezidua)
- diagramy k ověření stability rozptylu, např. $[\sqrt{|v_i|}, \hat{Y}_i]$
- Cookova vzdálenost pro vyjádření relativního vlivu jednotlivých pozorování
- `[plot(lm(mortality ocean+latitude))]`

206

ukázky diagnostiky

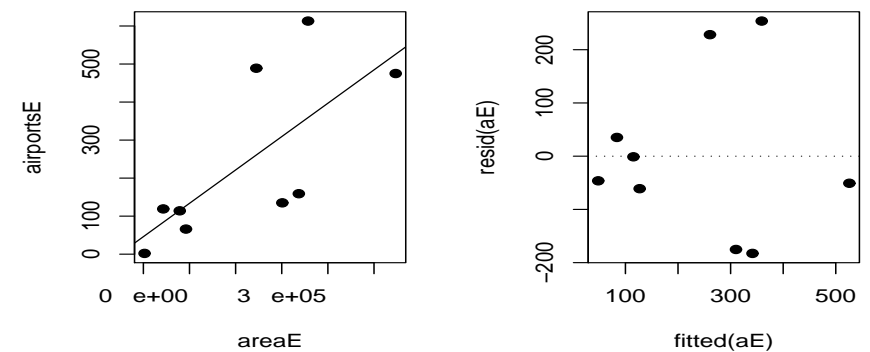
vlevo: rezidua spíše kladná než záporná, možná jsme měli raději vysvětlovat odmocninu z mortality

vpravo: normální diagram, ukazuje, že s předpokladem o normálním rozdělení není problém (body těsně kolem přímky)



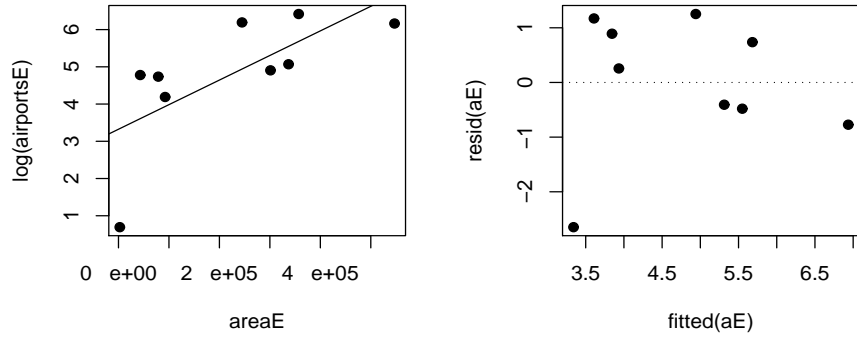
207

nekonstantní rozptyl (trychtýřovité rozšiřování mraku reziduí)



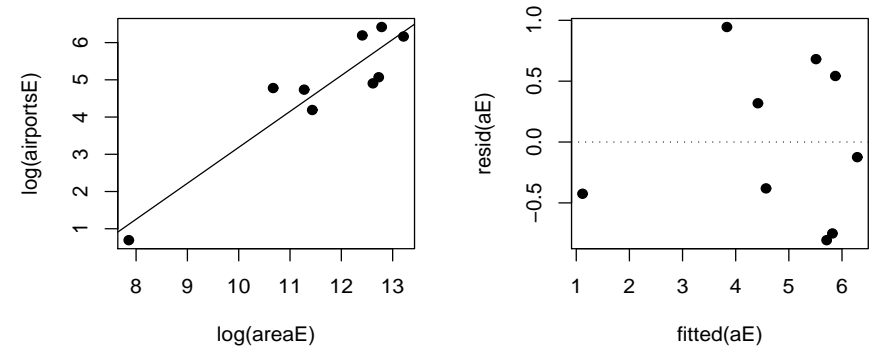
208

odlehle pozorování (první pozorování daleko od vodorovné osy)



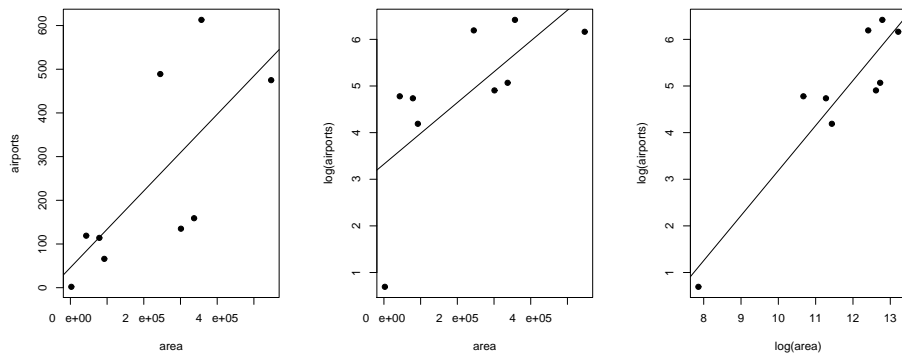
209

vlivné pozorování (první pozorování daleko ve vodorovném směru)

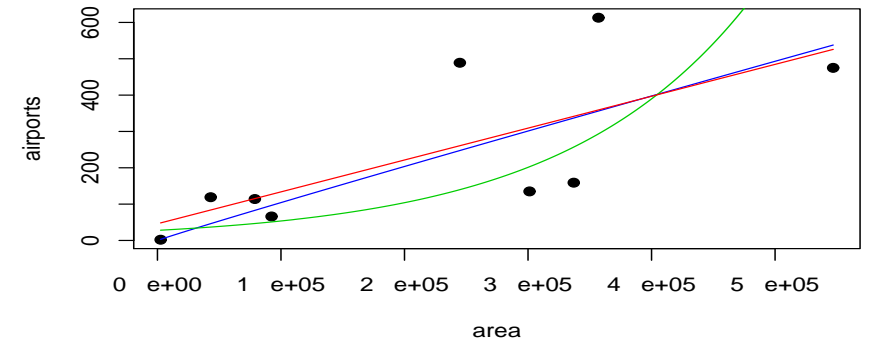


210

ukázka transformace: počet letišť na ploše evropského státu



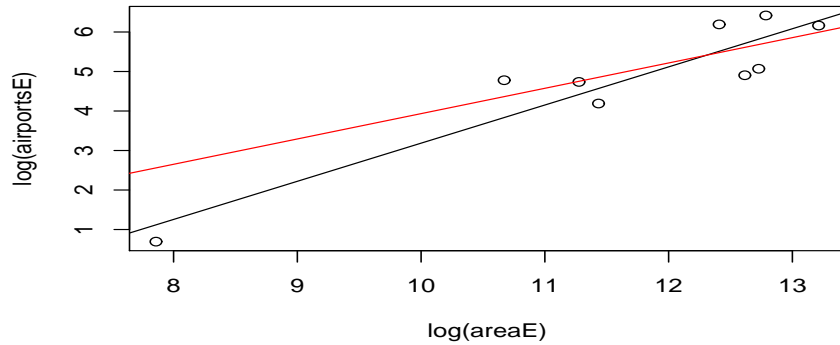
211



$y=46+0,0009 x$; $\log(y)=3,3+0,000007 x$; $\log(y)=-6,5+0,97 \log(x)$
 $R^2 = 51,4 \%$; $R^2 = 48,0 \%$; $R^2 = 86,1 \%$

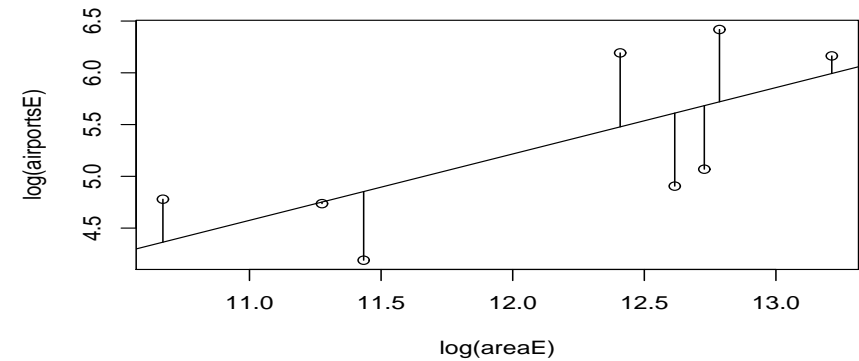
212

počet letišť: vynechat Lucembursko?



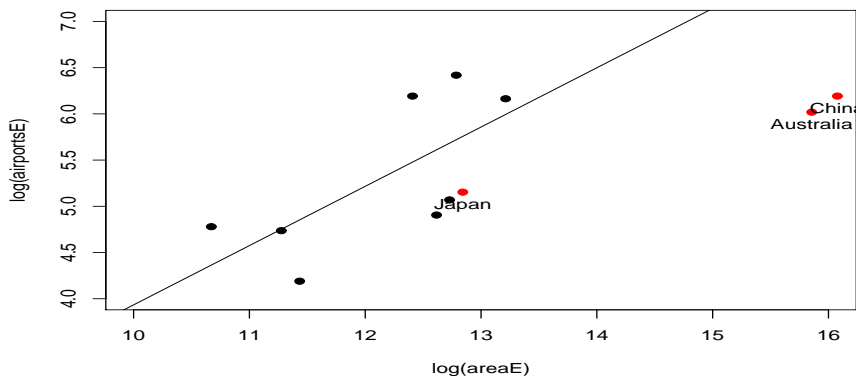
213

počet letišť: po vynechání Lucemburska



214

platí stejná závislost pro Japonsko, Čínu, Austrálii?



215

shrnutí

- regrese slouží pro
 - predikci (středních hodnot) budoucích pozorování
 - prokazování závislosti na zvoleném regresoru
 - ověřování modelu o závislosti
- nejsou-li splněny základní předpoklady \Rightarrow pochybné závěry
 - obtížně lze predikovat mimo obor měření
 - je-li malé R^2 , nespolehlivá předpověď, ale závislost může být prokazatelná
 - vysvětlovaná proměnná může záviset na více nezávisle proměnných, nutná opatrnost (confounding)

216

poznámky

- **souvislost regresní přímky a korelačního koeficientu:** testovou statistiku pro $H_0: \beta_1 = 0$ lze spočítat také jako

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

- je-li $|T|$ velké, závislost je prokázána, lze použít (nutno předpokládat normální rozdělení)
- metoda nejmenších čtverců je velmi citlivá na mimořádně umístěná pozorování
- příklad: počet letišť vers. velikost státu (obojí v logaritmech)
 - všech 9 států: $r = 0,717$; $T = 2,720$; $p = 3,0 \%$
 - bez Lucemburska: $r = 0,654$; $T = 2,226$; $p = 7,9 \%$

217

příklad: počet letišť (přesně $p = 2,9 \%$)

plocha	78,9	43,1	337,0	547,0	357,0	301,2	92,4	244,8
letišť	114	119	159	475	613	135	66	489
R_i	2	1	6	8	7	5	3	4
Q_i	2	3	5	6	8	4	1	7
$R_i - Q_i$	0	-2	1	2	-1	1	2	-3
$(R_i - Q_i)^2$	0	4	1	4	1	1	4	9

H_0 : počet letišť **nezávisí** na velikosti státu

H_1 : počet letišť **roste** s velikostí státu (jednostranná alternativa)

$$r_S = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (R_i - Q_i)^2 = 1 - \frac{6 \cdot 24}{8(64 - 1)} = 1 - \frac{144}{504} = 0,714$$

$$Z_0 = r_S \sqrt{n - 1} = 0,714 \cdot \sqrt{7} = 1,89$$

$$p = P(Z \geq Z_0) = 1 - \Phi(Z_0) = 1 - 0,971 = 0,029$$

na 5% hladině jsme (při jednostranné alternativě) závislost prokázali

219

Spearmanův korelační koeficient

- vlastně korelační koeficient pořadí
- citlivě reaguje i na nelineární, ale monotonní závislost
- k prokazování závislosti netřeba normální rozdělení, slabší test
- pro $n \geq 10$ lze při nezávislosti předpokládat $r_S \sqrt{n - 1} \sim N(0, 1)$

- závislost (proti oboustranné alternativě) prokázána, pokud

$$|r_S \sqrt{n - 1}| \geq z(\alpha/2)$$

- závislost (proti jednostranné alternativě) prokázána, pokud

$$r_S \sqrt{n - 1} \geq z(\alpha) \text{ resp. } r_S \sqrt{n - 1} \leq -z(\alpha)$$

218

vyrovnávání

- mechanismus k vyhlazování dat, spíše technický
- cílem je např. nahradit nedostupné pozorování (budoucí pozorování – **predikce**, chybějící pozorování – **interpolace**) nebo odstranit nahodilé výkyvy
- často (naš příklad) porušen předpoklad nezávislosti pozorování, proto by obyčejná regrese vedla k nesprávnému odhadu přesnosti odhadů (příliš optimistické!) a tedy nesprávně počítala testy o parametrech
- koeficienty nelze snadno statisticky hodnotit, někdy vůbec
- nejen metoda nejmenších čtverců

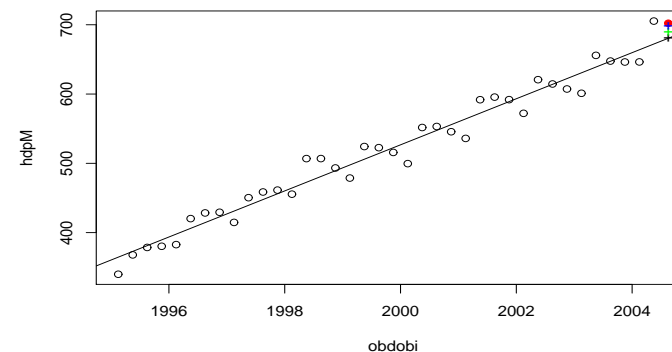
220

časové řady

- spojitý znak měřený v pravidelných časových intervalech
- mysladná hodnota složena z několika složek
 - **trend** (dlouhodobý vývoj)
 - **sezónní složka** (periodická složka se známou periodou, např. denní/roční chod teplot, čtvrtletní chod ekonomických veličin)
 - **periodická složka** (s neznámou periodou)
 - **autokorelace** (chybové složky na rozdíl od regrese nejsou mezi sebou nezávislé)
- první dvě složky lze vedle regrese **odhalit** pomocí
 - klouzavých průměrů
 - exponenciálního vyrovnávání
- **prokázat** pomocí regrese, která přihlédne k autokorelaci . . .

221

příklad: HDP v ČR po čtvrtletích (běžné ceny)



222

ilustrativní příklad: HDP [miliard Kč]

- predikce pomocí regrese (bez kvartálů): $360,4 + 33,3(\text{rok}-1995)$ pro 3. čtvrtletí 2004 tedy předpověď 680,5

- predikce pomocí regrese s ohledem na kvartály

$$\text{est}(HDP) = 339,4 + 33,1(\text{rok} - 1995)$$

$$\text{est}(HDP) = (339,4 + 38,5) + 33,1(\text{rok} - 1995)$$

$$\text{est}(HDP) = (339,4 + 30,2) + 33,1(\text{rok} - 1995)$$

$$\text{est}(HDP) = (339,4 + 18,1) + 33,1(\text{rok} - 1995)$$

předpověď tedy $339,4 + 30,2 + 33,1 \cdot (2004,625 - 1995) = 688,6$

- predikce s ohledem na autokorelaci (odhad autokorelačního koeficientu $\hat{\rho} = 0,59$): $688,6 + 0,59 \cdot 16,7 = 698,4$
- skutečnost: 702,2

223

autokorelace, periodicita

- speciální postupy pro zbývající složky (periodická složka s neznámou periodou, autokorelace), nelze mechanicky použít lineární regresi či lineární vyhlazování
- náhodné (chybové) členy v regresním modelu by měly být nezávislé, někdy ale daný člen závisí na předchozím; síla závislosti popsána pomocí autokorelačního koeficientu ρ
 - kladné ρ : po sobě jdoucí členy podobné (častý případ)
 - záporné ρ : po sobě jdoucí členy nepodobné
 - $\rho = 0$: po sobě jdoucí členy nezávislé (v regresi se požaduje)
- autokorelaci a periodicitu lze prokazovat (odhadovat, hodnotit) až po odstranění trendu a sezónního kolísání

224

klouzavé průměry (moving averages)

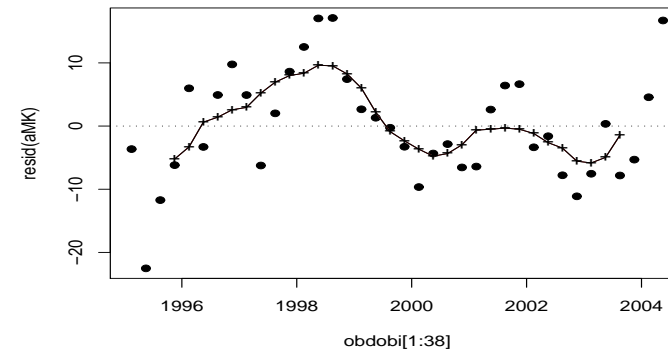
- pozorování Y_i porovnáváme s průměrem z m sousedních pozorování (včetně Y_i samotného), např. pro $m = 5$

$$\frac{1}{5}(Y_{i-2} + Y_{i-1} + Y_i + Y_{i+1} + Y_{i+2})$$

- snaha vyhladit nahodilé výchyly, zachovat „průměrný“ vývoj
- vhodné zejména k interpolaci, k nalézání dosud přehlížených systematických vlivů
- u HDP vezmeme nejprve v úvahu lineární trend a čtvrtletní periodicitu, teprve pak počítáme rezidua (to, co nevysvětlíme lineárním trendem a čtvrtletními sezonními výkyvy)

225

příklad: klouzavé průměry reziduí ($m = 7$)



226

exponenciální vyrovnávání

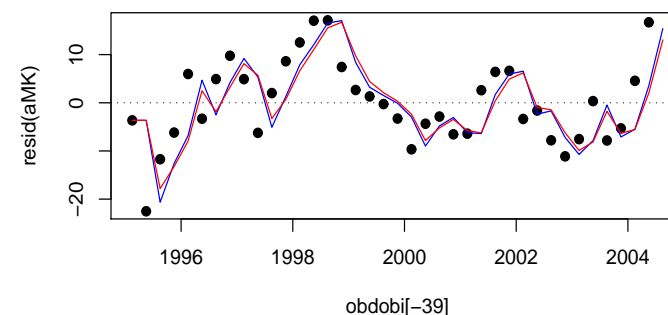
- představa: v posledním pozorování se projevuje vliv všech předchozích
- tento vliv je postupně utlumován: největší vliv má bezprostředně předcházející pozorování, nejmenší vliv pozorování v čase nejvzdálenější
- vážený průměr mezi předpovědí předchozího pozorování a skutečným předchozím pozorováním

$$w\hat{Y}_{i-1} + (1-w)Y_{i-1}$$

- w – váha „historie“, čím je větší, tím méně kolísání
- použitelné k předpovědi

227

exponenciální vyrovnávání reziduí (704,0 pro $w=0,1$, 701,6 pro $w=0,25$)



228

multinomické rozdělení

- zobecnění binomického rozdělení na k -tici náhodných veličin X_1, \dots, X_k
- parametry n, π_1, \dots, π_k ($0 < \pi_j < 1, \pi_1 + \dots + \pi_k = 1$)
- n **nezávislých** pokusů
- v každém pokusu **právě jeden** z k možných výsledků
- j -tý výsledek s pravděpodobností π_j
- X_j – počet pokusů, v nichž nastal j -tý možný výsledek, tedy nutně

$$X_1 + \dots + X_k = n$$

229

vlastnosti multinomického rozdělení

- každá složka má binomické rozdělení: $X_j \sim \text{bi}(n, \pi_j)$
- střední hodnota: $\mu_{X_j} = n\pi_j$, rozptyl: $\sigma_{X_j}^2 = n\pi_j(1 - \pi_j)$
- (pro zajímavost) kovariance: $\text{cov}(X_j, X_t) = -n\pi_j\pi_t \quad j \neq t$
- asymptotická vlastnost **chí-kvadrát** (velká n , $n\pi_j \geq 5$)

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

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

231

příklady multinomického rozdělení

- předvolební průzkum
 - n – počet tázaných
 - π_j – skutečný podíl voličů j -té strany v populaci
 - X_j – počet (četnost) voličů j -té strany ve výběru
- hody hrací kostkou
 - n – počet hodů
 - π_1, \dots, π_6 – pravděpodobnosti jednotlivých stran kostky
 - X_1, \dots, X_6 – absolutní četnosti jednotlivých stran kostky

230

příklad: hrací kostka A

- test **jednoduché** hypotézy
- $n = 100$ hodů kostkou
- $X_1 = 12, X_2 = 21, X_3 = 14, X_4 = 15, X_5 = 21, X_6 = 17$
- hypotéza $H_0 : \pi_1 = \dots = \pi_6 = 1/6$ dá očekávané četnosti
 $n\pi_1 = \dots = n\pi_6 = 100/6 = 16,67$

$$\chi^2 = \frac{(12 - 16,67)^2}{16,67} + \dots + \frac{(17 - 16,67)^2}{16,67} = 4,16 < \chi_5^2(0,05) = 11,07$$

$$p = 52,7 \%$$

- `[chisq.test(c(12,21,14,15,21,17),p=rep(1,6)/6)]`

232

příklad: hrací kostka B (1)

- $n = 100$ hodů kostkou
- $X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15, X_6 = 41$
- hypotéza $H_0 : \pi_1 = \dots = \pi_6 = 1/6$ dá očekávané četnosti
 $n\pi_1 = \dots = n\pi_6 = 100/6 = 16,67$

$$\chi^2 = \frac{(15 - 16,67)^2}{16,67} + \dots + \frac{(41 - 16,67)^2}{16,67} = 48,32 > \chi_5^2(0,05) = 11,07$$

- zřejmě je nutno zamítnout hypotézu, že kostka je symetrická

233

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

- $n = 100$ hodů kostkou
- $(X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15), X_6 = 41$
- nulová hypotéza: $\pi_6 = 5/10 = 1/2$
- hypotéza o psti jediného z možných výsledků (pst šestky) – binomické rozdělení
- na 6. přednášce 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)]

235

příklad: hrací kostka B (2), jiná H_0

- $n = 100$ hodů kostkou
- $X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15, X_6 = 41$
- nulová hypotéza: $\pi_1 = \dots = \pi_5 = 1/10, \pi_6 = 5/10 = 1/2$
- očekávané četnosti za hypotézy:
 $n\pi_1 = \dots = n\pi_5 = 100/10 = 10, n\pi_6 = 100/2 = 50$

$$\chi^2 = \frac{(15 - 10)^2}{10} + \dots + \frac{(15 - 10)^2}{10} + \frac{(41 - 50)^2}{50} = 12,72 > \chi_5^2(0,05) = 11,07$$

- 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)]

234

test homogenity r výběrů

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

236

test homogenity r výběrů

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

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

- platí-li hypotéza, má výsledná statistika χ^2 -rozdělení $\chi^2_{(r-1)(k-1)}$
- hypotézu o shodě pravděpodobností v r populacích zamítáme, je-li $\chi^2 \geq \chi^2_{(r-1)(k-1)}(\alpha)$

237

mají obě kostky stejné šestice pravděpodobností?

`[chisq.test(rbind(kostkaA,kostkaB))]`

- 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): $27 \cdot 100 / 200 = 13,5, \dots$

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

$$\chi^2 = \frac{(12 - 13,5)^2}{13,5} + \frac{(21 - 18,5)^2}{18,5} + \dots + \frac{(41 - 29)^2}{29} = 18,13 > 11,07 = \chi^2_5(0,05)$$

hypotézu o shodě psťi na obou kostkách **zamítáme** ($p = 0,3 \%$)

238

příklad: vzdělání matek `[chisq.test(table(Porodnice,Vzdelani))]`

porodnice	vzdělání			celkem
	základní	střední	VŠ	
Praha	23 (24,0)	30 (33,2)	17 (12,7)	70
venkov	11 (10,0)	17 (13,8)	1 (5,3)	29
celkem	34	47	18	99

- v závorce jsou očekávané četnosti za hypotézy, že podíly tří vzdělanostních skupin jsou v obou populacích shodné

$$\frac{(23 - 24)^2}{24} + \frac{(30 - 33,2)^2}{33,2} + \frac{(17 - 12,7)^2}{12,7} + \frac{(11 - 10)^2}{10} + \frac{(17 - 13,8)^2}{13,8} + \frac{(1 - 5,3)^2}{5,3}$$

$$\chi^2 = 6,12 > \chi^2_2(0,05) = 5,99, \quad p = 4,7 \%$$

- bylo třeba $o_{ij} \geq 5$ pro všechna i, j

239

test nezávislosti

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

240

příklad: plánovaná těhotenství `[chisq.test(table(Plan,Vzdelani))]`

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

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

- všechny očekávané četnosti dostatečně velké

$$\chi^2 = 6,68 > 5,99 = \chi_2^2(0,05), \quad p = 3,5 \%$$

241

čtyřpolní tabulka

`[chisq.test(matrix(c(1097,1365,362,354),2,2),correct=FALSE)]`

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

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

úmrtní na vnější příčiny v roce 2003

příčina	muži	ženy	celkem
dopr. nehody	1097 (1130,3)	362 (328,7)	1459
sebepoškození	1365 (1331,7)	354 (387,3)	1719
celkem	2462	716	3178

$$\chi^2 = \frac{3178(1097 \cdot 354 - 362 \cdot 1365)^2}{1459 \cdot 1719 \cdot 2462 \cdot 716} = 8,05 > \chi_1^2(0,05) = 3,84 \quad p = 0,5 \%$$

prokázali jsme neshodu mezi muži a ženami, souvislost s pohlavím vlastně prokázali neshodu podílů při nehodách (odhady 44,6 %, 50,6 %)

242