Statistika

(D360P03Z, D360P03U)

akademický rok 2004/2005

Karel Zvára

3. ledna 2005
opraveno 3. února 2005

literatura

- Z. Pavlík, K. Kühnl: Úvod do kvantitativních metod pro geografy, 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

cvičení, zápočet, zkouška

- cvičení v počítačových učebnách, MS Excel
- aktivní účast na cvičení, maximálně dvě absence, napsání zápočtového testu $\Rightarrow$ zápočet
- obsah cvičení více přizpůsoben studovanému oboru
- přednášky formulovány obecněji
- zkouška nejspíš písemná, kombinovaná s ústní, zápočet **musí** zkoušce **předcházet**

přehled témat (1)

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

přehled témat (2)

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

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
- 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?

co měříme (zjišťujeme) a kde

- měříme na mnoha statistických **jednotkách** (osoba, obec, 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**
- můžeme porovnávat vlastnosti znaku mezi soubory

měřítka (1)

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (země původu, barva očí)
- **ordinální** (dosažené vzdělání, stupeň bolesti)
  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

měřítka (2)

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

veličina

- číselně vyjádřený výsledek měření
- 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 hromadných vlastností (**charakteristiky polohy, variability**)

příklad: 100 hodů kostkou

počet ok – nominální znak

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

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

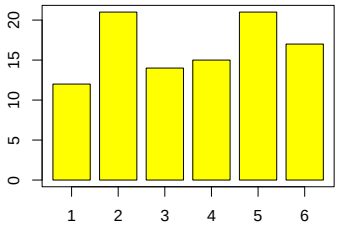
hody kostkou jako hromadný jev

- chceme 100 hodnot (počet ok, mohl to být obrázek) vyjádřit názorně, aby vypovídaly o vlastnostech kostky
- zjistíme (absolutní) **četnosti** hodnot
- lze dopočítat **relativní četnosti** hodnot, možno v %
- tabulka četností (absolutních, relativních)
- grafické vyjádření četností – **histogram** (velikost plochy je úměrná četnosti)
- rozhodování o kvalitě kostky (zda symetrická) je úlohou statistické indukce

četnosti výsledků hodů kostkou A

četnosti

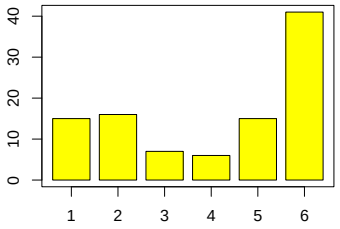
	n_j	$f_j = n_j/n$
1	12	0,12
2	21	0,21
3	14	0,14
4	15	0,15
5	21	0,21
6	17	0,17



četnosti výsledků hodů kostkou A

četnosti

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



možné úlohy statistické indukce

- je pravděpodobnost šestky rovna 1/6? (teorie pravděpodobnosti odvodí teoretickou hodnotu, matematická statistika odhadne, prověří představu teorie)
- je kostka symetrická, mají všechny stěny kostky stejnou pravděpodobnost?
- kolik potřebujeme nezávislých hodů, abychom s dostatečnou spolehlivostí poznali, že je kostka nesymetrická?
- liší se mezi sebou kostky A a B?
- vše založeno na modelu populace – výběr

populace a výběr

- možnost zobecnění z hodnot zjištěných na souboru měření: model populace – výběr
- **populace (základní soubor)** – velký soubor, jehož je zpracovávaný soubor **(výběr) reprezentativním** vzorkem (výskyt důležitých doprovodných znaků ve výběru odpovídá jeho výskytu v populaci
- reprezentativnosti nejlépe dosáhneme tak, že použijeme **prostý náhodný výběr**, kdy každá n-tice prvků populace má stejnou šanci (pravděpodobnost) do výběru se dostat
- na základě výběru tvrdíme něco o populaci

příklad: věk 99 matek

99 zjištěných hodnot – soubor 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	

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	

variační řada, pořadí

- původní (neuspořádaná) data – hodnoty v původním pořadí, bez ohledu na případná opakování $x_1, x_2, \ldots, x_n$
- **variační řada** $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$
  data uspořádána tak, aby hodnoty neklesaly
- **pořadí** – umístění pozorování ve variační řadě; shodným hodnotám dáváme průměrné pořadí

$x_{(j)}$	22	15	17	15	21	13	18
pořadí R_j	7	2,5	4	2,5	6	1	5

třídění, třídní četnosti

- spojitá veličina s velkým počtem hodnot
- obor hodnot rozdělíme na nepřekrývající se třídy (intervaly), nejlépe stejné délky
- všechna pozorování z daného intervalu nahradíme zástupnou hodnotou (středem x_j^*)
- zjistíme četnosti n_j jednotlivých tříd
- **kumulativní četnosti** udávají počet hodnot v dané třídě a třídách předcházejících

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

věk matek – třídní četnosti

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

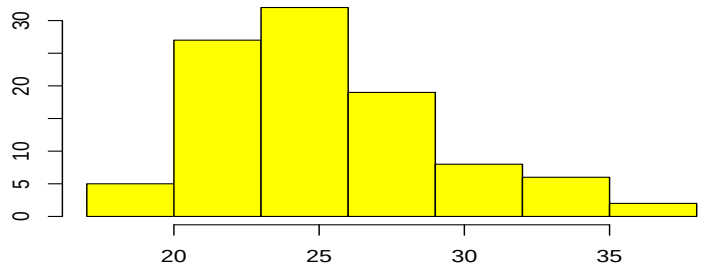
grafické znázornění třídních četností

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

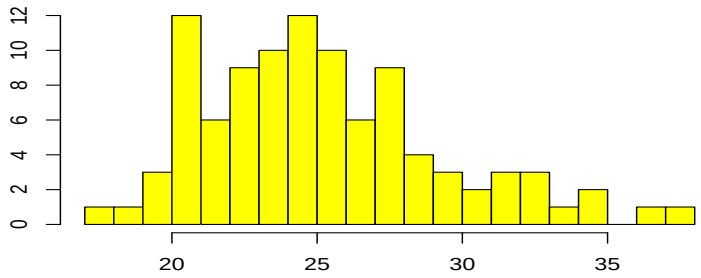
$$k \approx 1 + 3{,}3 \cdot \log_{10} n = 1 + \log_2 n$$

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

histogram, h =3 (k =7)



histogram, h =1 (nevhodné h)



populace

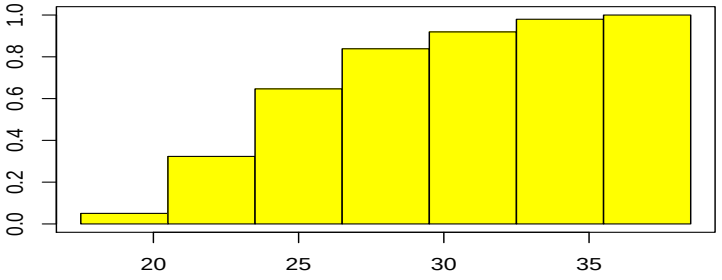
- velká populace, spojitá veličina – intervaly mohou být krátké, obálce histogramu relativních četností odpovídá **hustota** $f_X(x)$
- podobně kumulativním relativním četnostem odpovídá **distribuční funkce**
- hodnota distribuční funkce $F_X(x)$ je pravděpodobnost, že náhodná veličina X nepřekročí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

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

- souvislost: hustota je derivace distribuční funkce:

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

příklad: věk matek (kumulativní relativní četnosti)



parametry – odhady, statistiky

- podle toho, jakou roli hraje hodnocený soubor, rozlišujeme charakteristiky
  - **populační**: vztažené k populaci, mnohdy jen ideální, námi představované, **parametry** modelu
  - **výběrové**: vztažené k výběru z nějaké populace, takže jde o **odhady** nějakých populačních parametrů, **statistiky** spočítané z výběru
- příkladem dvojice odhad – parametr je relativní četnost – pravděpodobnost
- statistiky se používají při statistické indukci

charakteristiky polohy (1)

- **medián** (prostřední hodnota)

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

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

- medián dělí data na dvě stejné části (velkých hodnot a malých hodnot)
- **populační medián**:

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

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	

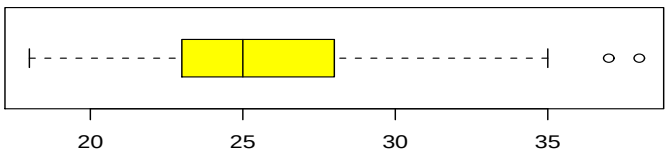
charakteristiky polohy (2)

- **dolní (horní) kvartil** Q_1 (Q_3) vyděluje čtvrtinu nejmenších (největších) hodnot
- speciální případ **percentilu** x_p pro $p = 0{,}25$ ($p = 0{,}75$), přičemž x_p vyděluje $100p$ % nejmenších hodnot od ostatních
- výpočet percentilů – mnoho vzorečků
- medián je také percentilem, totiž $x_{0,5}$
- **kvantil** = populační percentil

$$F_X(\mu_p) = \mathrm{P}(X \leq \mu_p) = p$$

grafické znázornění spojité veličiny

- **krabicový diagram** (box-plot) zobrazuje kvartily, medián, minimum, maximum, případně odlehlá pozorování: od bližšího kvartilu dál než 3/2·$(Q_3 - Q_1)$
- příklad: věk matek (dvě odlehlá pozorování, $Q_1 = 23$, $Q_3 = 28$)



charakteristiky polohy (3)

- **průměr** (kdyby bylo všech n hodnot stejných)

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

- **vážený průměr**: založen na četnostech

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

- **populační průměr**: značíme μ
- u nula-jedničkového měřítka: průměr = relativní četnost jedniček, populační průměr = pravděpodobnost jedničky

charakteristiky polohy (4)

- **modus** $\hat{x}$ nejčastější hodnota (lze počítat také pro nominální či ordinální měřítko)
- modus nemusí být určen jednoznačně
- **populační modus** pro spojitou veličinu – hodnota, kde je hustota maximální
- **populační modus** pro diskrétní veličinu (četnosti) – nejpravděpodobnější hodnota

příklad – věk matek

- průměr

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

- vážený průměr založený na třídění

$$\bar{x} = \frac{1}{99}(5 \cdot 19 + 27 \cdot 22 + 32 \cdot 25 + 19 \cdot 28 + 8 \cdot 31 + 6 \cdot 34 + 2 \cdot 37)$$
$$= \frac{2547}{99} \doteq 25{,}7$$

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

charakteristiky polohy (5)

- **alfa-useknutý průměr**: nejprve se oddělí (usekne) 100α % nejmenších a 100α % nejmenších hodnot, ze zbytku se spočítá průměr
- je robustní vůči odlehlým hodnotám
- volí se $\alpha = 0{,}1$ (0,15)
- věk matek

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

příklad: věk 99 matek

vyloučí se [0,1·99]=[9,9]=9 (celá část)
nejmenších a 9 největších hodnot

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

vlastnosti charakteristik polohy

- 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

$$\text{průměr}(a + b \cdot x) = a + b \cdot \bar{x}, \qquad b > 0$$

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, docentů UK 2002:
  41 profesorů, průměrný věk 51,1
  77 docentů, průměrný věk 47,8
  celkový průměr:

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

  nikoliv

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

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)

charakteristiky variability (1)

- obecně $s(a + b \cdot x) = b \cdot s(x)$
- přičtením stejné konstanty a (posunutím) se charakteristika variability nezmění
- vynásobení kladnou konstantou znamená, že stejnou konstantu nutno vynásobit charakteristiku variability
- měří nestejnost hodnot spojité veličiny
- **rozpětí** (jen pro výběr) $R = x_{(n)} - x_{(1)}$
- **kvartilové rozpětí** $R_Q = Q_3 - Q_1$

charakteristiky variability (2)

- (výběrový) **rozptyl**
  (nevyhovuje obecnému požadavku přesně: $s^2_{a+b\cdot x} = b^2 \cdot s^2_x$)

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

- **populační rozptyl** σ^2 má ve jmenovateli n

charakteristiky variability (3)

- rozptyl měří průměrný čtverec vzdálenosti od průměru
- **směrodatná odchylka**: odmocnina z rozptylu

$$s_x = \sqrt{s^2_x} \qquad \sigma = \sqrt{\sigma^2}$$

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

$$\text{odečti } \frac{h^2}{12}$$

příklad – věk matek

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

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

- směrodatná odchylka je 4,1

příklad – věk matek

- pomocí třídních četností

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

- navíc Sheppardova korekce

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

charakteristiky variability (4)

- **střední odchylka**: průměr odchylek od mediánu (někdy od průměru)

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

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

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

normované charakteristiky rozptýlenosti

- dosud zavedené míry závisejí na volbě měřítka (např. délka v m nebo v km)
- hledáme charakteristiky nezávislé na měřítku
- nutně *poměrové* měřítko, *kladné* hodnoty
- **variační koeficient**

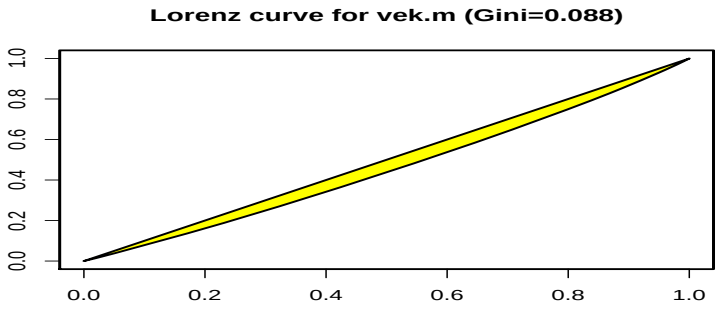
$$v = \frac{s_x}{\bar{x}}$$

- **(Giniho) koeficient koncentrace**

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

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

Lorenzova křivka (věk matek)



Lorenzova křivka

- variační řada: $0 < x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$
- kumulativní součty pro $j = 0, 1, \ldots, n$

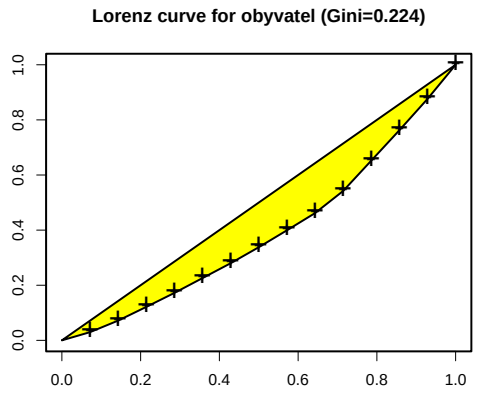
$$t_0 = 0 \qquad t_j = x_{(1)} + x_{(2)} + \ldots + x_{(j)} = \sum_{i=1}^{j} x_{(i)}$$

- úsečkami spojit body $(j/n; t_j/t_n), \quad 0 \le j \le n$
- zajímá nás plocha nad touto lomenou čarou a pod úhlopříčkou jednotkového čtverce
- Giniho koeficient koncentrace je dvojnásobkem této plochy
- lze vzít v úvahu i váhy (četnosti)

příklad: obyvatelé krajů

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

Lorenzova křivka (obyvatelé – kraje)



z-skór, standardizace

- variační koeficient, Giniho koeficient – příklady bezrozměrných veličin
- z-skór

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

- dostaneme nulový průměr, jednotkový rozptyl
- bezrozměrný
- umožní srovnávat různě obtížné testy, ...

šikmost, špičatost

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

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

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

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

dvojice znaků (veličin)

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

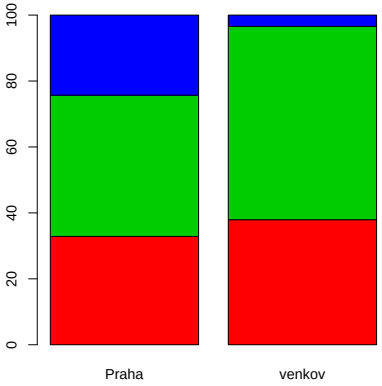
kvalitativní – kvalitativní

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

příklad – vzdělání matek (pozor na orientaci)

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

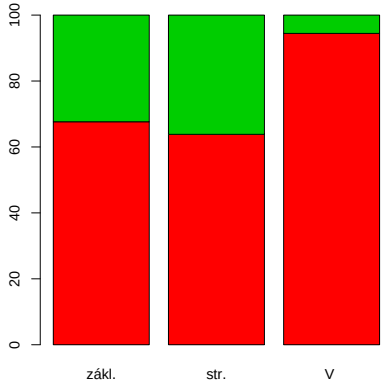
vzdělání	porodnice		celkem
	Praha	venkov	
základní	32,9 %	37,9 %	34,3 %
střední	42,8 %	58,6 %	47,5 %
VŠ	24,3 %	3,5 %	18,2 %
celkem	100 %	100 %	100 %



příklad – vzdělání matek (pozor na orientaci)

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

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



příklad – vzdělání matek (očekávané četnosti)

- kdyby rozdělení vzdělání bylo všude stejné, očekáváme tři možnosti v poměru 34:47:19, celkem 99
- pražských 70 matek by stejný poměr dalo při **očekávaných** četnostech 70·34/99=24,0, resp. 70·47/99=33,2 resp. 70·18/99=12,7
- podobně pro matky z venkova dostaneme 9,96, po zaokrouhlení 10,0, pro další četnosti 13,8 resp. 5,3

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

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

příklad – vzdělání matek (očekávané četnosti)

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

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

empirické a **očekávané** četnosti porovnáme pomocí statistiky **chí-kvadrát**:

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

velká hodnota χ^2 svědčí o velké neshodě

příklad: plánovaná těhotenství

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

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

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

příklad: plánovaná těhotenství (očekávané četnosti)

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

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

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

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

příklad: předvolební výzkum

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

muž	strana A	strana B	celkem
muž	**11**	**4**	15
žena	**6**	**9**	15
celkem	17	13	50

muž	strana A	strana B	celkem
muž	73 %	27 %	100 %
žena	40 %	60 %	100 %
celkem	57 %	43 %	100 %

muž	strana A	strana B	celkem
muž	65 %	31 %	50 %
žena	35 %	69 %	50 %
celkem	100 %	100 %	100 %

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 **čtyřpolním korelačním koeficientem**

$$r_{2,2} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

- $r_{2,2}$ je mezi -1 a 1
- např. pro

11	4	15
6	9	15
17	13	30

vyjde $r_{2,2} = \frac{11 \cdot 9 - 4 \cdot 6}{\sqrt{15 \cdot 15 \cdot 17 \cdot 13}} = 0{,}34$

příklad: předvolební průzkum

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

muž	strana A	strana B	celkem
muž	**11**	**4**	15
žena	**6**	**9**	15
celkem	17	13	50

vychází 0,34>0,

protože je 11·9 > 6·4

čtyřpolní tabulka – prokazování závislosti

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

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

- příklad (předvolební průzkum)

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

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

$r_{2,2}$=0,055

město	A	B	celkem
muž	5	2	7
žena	11	5	16
celkem	8	10	18

$r_{2,2}$=0,027

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

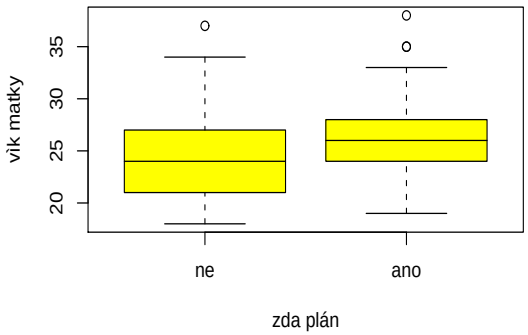
$r_{2,2}$=-0,07

kdyby byl stejný poměr mezi muži a ženami oslovenými ve městě a na venkově, problém by nevznikl

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)

příklad: věk matek



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

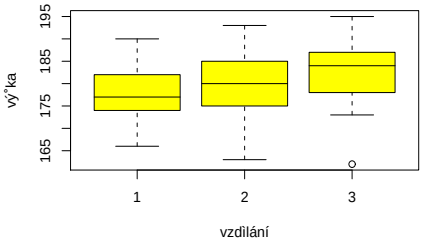
závislost kvalitativní – kvantitativní

- pro nula-jedničkové x sílu závislosti x, y vyjadřuje **bodově biseriální korelační koeficient**

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

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

příklad: výška otce ∼ vzdělání matky

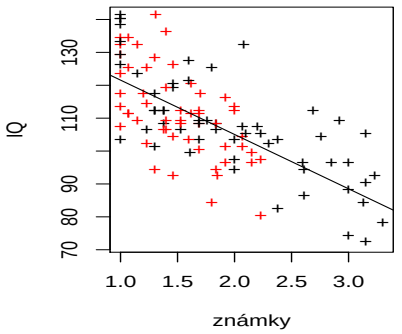


vzdělání	rozsah	průměr	sm. odch.
základní	34	177,1	6,0
střední	47	179,5	6,4
VŠ	18	182,8	7,8
celkem	99	179,3	6,8

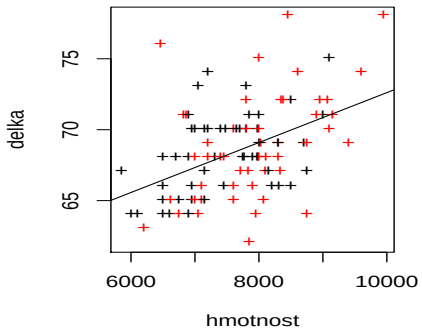
$$\bar{x} = \frac{34 \cdot 177{,}1 + 47 \cdot 179{,}5 + 18 \cdot 182{,}8}{34 + 47 + 18} = 179{,}3$$

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

dvojice kvantitativních veličin



$r = -0{,}69$



$r = 0{,}45$

závislost spojitých veličin

- (výběrová) **kovariance**

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

- (Pearsonův) **korelační koeficient** (z-skóry)

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

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

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

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

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 (tabulky)
- takto lze závislost prokazovat jen někdy (normální rozdělení)
- špatně zachytí křivočarou (nelineární) závislost

Spearmanův korelační koeficient

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

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

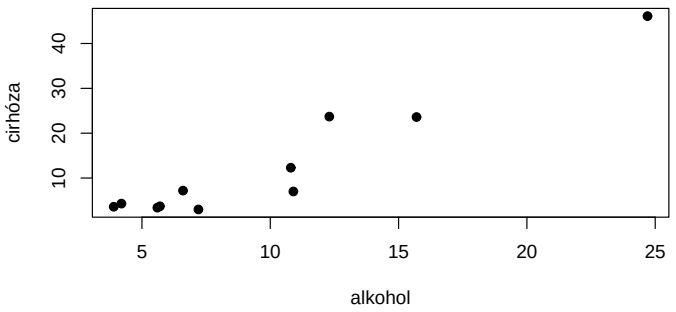
- vhodný pro nelineární monotonní **závislost**, nevadí odlehlé hodnoty
- při testování nemusí být normální rozdělení

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

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

příklad: spotřeba alkoholu a cirhóza jater



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 jeho pravděpodobnosti

klasická pravděpodobnost

- nechť každý jev se skládá ze stejně pravděpodobných **elementárních jevů** (symetrie)
- celkem M stejně pravděpodobných disjunktních elementárních jevů
- z nich je M_A **příznivých** jevu A
- **klasická pravděpodobnost**

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

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 $\mathsf{P}(A) = 1/6$
- $M_B = 3$, tedy $\mathsf{P}(B) = 3/6 = 1/2$

faktoriál

- **faktoriál** $n! = n \cdot (n-1) \cdots 2 \cdot 1$
- kolika způsoby lze uspořádat za sebou n rozlišitelných prvků
- příklady:
  - $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$
  - $1! = 1,\ 0! = 1$
- kolika způsoby lze uspořádat za sebou 14 krajů:
  $14! = 14 \cdot 13 \cdot 12 \cdots 2 \cdot 1 =$ 87 178 291 200

počet kombinací

- **kombinační číslo** n nad k
- počet k-prvkových podmnožin množiny o n prvcích nezávisle na jejich pořadí

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

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

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

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

příklad: losování otázek (1)

- student *neumí* 5 otázek, *umí* 10 otázek
- losuje se dvojice otázek z oněch 15
- 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

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

příklad: losování otázek (2)

- pravděpodobnost, že zná *právě* jednu otázku

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

- pravděpodobnost, že zná *právě* dvě otázky

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

- pravděpodobnost, že zná *aspoň jednu* otázku

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

pravidla pro pravděpodobnost (1)

- **sjednocení** jevů $B \cup C$: platí B **nebo** C
- **průnik** $B \cap C$: platí B a **současně** C

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

pravidla pro pravděpodobnost (2)

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

$$\mathsf{P}(B \cup C) = \mathsf{P}(B) + \mathsf{P}(C)$$

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

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

- **nezávislé jevy**: výskyt jednoho jevu **neovlivní** pravděpodobnost výskytu druhého:

$$\mathsf{P}(B) = \mathsf{P}(B|C) = \frac{\mathsf{P}(B \cap C)}{\mathsf{P}(C)} \Leftrightarrow \mathsf{P}(B \cap C) = \mathsf{P}(B)\mathsf{P}(C)$$

idealizovaný příklad

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

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

- pravděpodobnost, že aspoň jedna jednička:

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

rozdělení náhodné veličiny

- **náhodná veličina** – číselné vyjádřený výsledek náhodného pokusu
- **diskrétní rozdělení** určeno seznamem možných hodnot a jejich pravděpodobnostmi:

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

- **spojité rozdělení** určeno **distribuční funkcí**

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

nebo **hustotou**

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

příklad diskrétního rozdělení: známka u zkoušky

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

Jak jedním číslem charakterizovat úroveň známek?
Průměr by X, Y nerozlišil $\Rightarrow$ **vážený průměr**
vahami známek budou jejich pravděpodobnosti
dostaneme tak **populační průměry**

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

charakteristiky rozdělení náhodné veličiny (1)

- **střední hodnota** náhodné veličiny X (populační průměr)
  - vážený průměr možných hodnot
  - vahami pravděpodobnosti hodnot
  - $\mu_X = \mathsf{E}\, X = x_1\mathsf{P}(X = x_1) + x_2\mathsf{P}(X = x_2) + \ldots$
- když se použije operátor E na náhodnou veličinu X, spočítá vážený průměr jejích hodnot, vahami jsou u diskrétního rozdělení pravděpodobnosti těchto hodnot
- pro spojité rozdělení

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

charakteristiky rozdělení náhodné veličiny (2)

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

$$\sigma_X^2 = \mathsf{E}\,(X - \mu_X)^2 = (x_1 - \mu_X)^2\mathsf{P}(X = x_1) + (x_2 - \mu_X)^2\mathsf{P}(X = x_2) + \ldots$$
$$= \sum_k (x_k - \mu_X)^2\mathsf{P}(X = k)$$
$$\sigma_X^2 = \mathsf{E}\,(X - \mu_X)^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\mathrm{d}x$$

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

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

příklad diskrétního rozdělení: známka u zkoušky

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

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

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

vlastnosti střední hodnoty a rozptylu

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

$$\mu_{a+bX} = \mathsf{E}\,(a + bX) = a + b\ \mathsf{E}\, X = a + b\mu_X$$
$$\mu_{X+Y} = \mathsf{E}\,(X + Y) = \mathsf{E}\, X + \mathsf{E}\, Y = \mu_X + \mu_Y$$
$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{X,Y}$$
$$\sigma_{X,Y} = \mathsf{E}\,(X - \mu_X)(Y - \mu_Y) \text{ **kovariance** } X, Y$$
$$= (x_1 - \mu_X)(y_1 - \mu_Y)\mathsf{P}(X = x_1, Y = y_1)$$
$$+ (x_1 - \mu_X)(y_2 - \mu_Y)\mathsf{P}(X = x_1, Y = y_2) + \ldots \text{ (všechny dvojice)}$$

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

alternativní rozdělení

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

$$\mu_X = \mathsf{E}\, X = 1 \cdot \mathsf{P}(X = 1) + 0 \cdot \mathsf{P}(X = 0) = \pi$$

- (populační) **rozptyl**

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

binomické rozdělení bi(n, π) (1)

- diskrétní rozdělení s parametry $n, \pi \qquad (0 < \pi < 1)$
- n **nezávislých** pokusů
- v každém zdar s pravděpodobností π, nezdar s pstí $1 - \pi$
- X – celkový počet zdarů X má binomické rozdělení s parametry n a π
- X je součet n nezávislých náhodných veličin X_i s alternativním rozdělením 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

binomické rozdělení bi(n, π) (2)

-

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

-

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

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

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

příklad: zkoušky

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

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

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

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

příklad: kouření

- víme, že mezi dvacetiletými muži je (řekněme) 35 % kuřádů (je-li 70 tisíc dvacetiletých, pak je kuřáků 24500, jen nevíme kteří)
- vybereme náhodně 60 dvacetiletých mužů, X – počet kuřáků mezi nimi, tedy $X \sim \mathrm{bi}(60, 0{,}35)$
- $$\mu_X = 60 \cdot 0{,}35 = 21 \qquad \sigma_X^2 = 60 \cdot 0{,}35 \cdot 0{,}65 = 13{,}65 = (3{,}7)^2$$
- ukázky pravděpodobností možných hodnot

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

Poissonovo rozdělení $\mathrm{Po}(\lambda)$ (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 průměru vyskytuje ve zvolené jednotce)

$$\mathrm{P}(Y = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$$
$$\mu_Y = \lambda \qquad \sigma_Y^2 = \lambda$$

Poissonovo rozdělení $\mathrm{Po}(\lambda)$ (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 . . .), parametrem bude 3λ
- analogicky pro jiné kladné násobky
- $X \sim \mathrm{bi}(n, \pi)$, n velké, π malé, pak pravděpodobnosti hodnot X lze aproximovat (přibližně vyjádřit) pomocí pravděpodobností hodnot $Y \sim \mathrm{Po}(n\pi)$

příklady Poissonova rozdělení

- do pasti padne za noc v průměru 8 brouků
- s jako pravděpodobností jich tam ráno najdeme 10?

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

- vezmeme-li past s poloviční obvodem, očekáváme poloviční průměr za noc

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

příklady

- s jakou pravděpodobností **neudělá** 12 z 50 stejně připravených studentů zkoušku?
- binomické rozdělení bi$(50, 0{,}2)$

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

- Poissonovo rozdělení Po$(50 \cdot 0{,}2)$=Po(10)

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

normální (Gaussovo) rozdělení $\mathsf{N}\left(\mu, \sigma^2\right)$

- spojité rozdělení, symetrické okolo střední hodnoty μ
- model vzniku: součet velkého počtu nepatrných příspěvků
- pro $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$ platí

$$\mu_X = \mathsf{E}\, X = \mu \qquad \sigma_X^2 = \mathsf{E}\,(X - \mu_X)^2 = \sigma^2$$
$$\mathsf{P}(|X - \mu| < 1{,}00\sigma) = 0{,}68, \text{ tj. } 68\ \%$$
$$\mathsf{P}(|X - \mu| < 1{,}96\sigma) = 0{,}95, \text{ tj. } 95\ \%$$
$$\mathsf{P}(|X - \mu| < 3{,}00\sigma) = 0{,}997, \text{ tj. } 99{,}7\ \%$$
$$X \sim \mathsf{N}\left(\mu, \sigma^2\right) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim \mathsf{N}(0, 1)$$

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

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

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

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

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

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

- 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

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

výpočet distribuční funkce $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$

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

$$\begin{aligned}\mathsf{P}(X \leq x) &= \mathsf{P}\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\ &= \mathsf{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)\end{aligned}$$

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

příklad: výšky hochů

- nechť výšky desetiletých (v cm) přibližně $\mathsf{N}(140, 36)$
- s jakou pravděpodobností je výška náhodně vybraného chlapce mezi 140 a 145 cm (včetně), když vezmeme v úvahu zaokrouhlování?

$$\begin{aligned}\mathsf{P}(140 \leq X \leq 145) &= \mathsf{P}(139{,}5 < X < 145{,}5 \\ &= \Phi\left(\frac{145{,}5-140}{6}\right) - \Phi\left(\frac{139{,}5-140}{6}\right) \\ &= \Phi(0{,}917) - \Phi(-0{,}083) \\ &= 0{,}82 - 0{,}47 = 0{,}35 \text{ tedy } 35\ \%\end{aligned}$$

$$\mathsf{P}(X > 145) = 1 - \Phi\left(\frac{145{,}5-140}{6}\right) = 0{,}18 \text{ tedy } 18\ \%$$

příklad: uchazeči o studium

- 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)?
- $X \sim \mathsf{bi}(500, 0{,}45)$ má $\mu_X = 500 \cdot 0{,}45 = 225$, $\sigma_X^2 = 500 \cdot 0{,}45 \cdot 0{,}55 = 123{,}75$, tedy $\sigma_X = 11{,}1$

$$\mathsf{P}(200 \leq X \leq 220) = \Phi\left(\frac{220{,}5-225}{11{,}1}\right) - \Phi\left(\frac{199{,}5-225}{11{,}1}\right) = 0{,}34 - 0{,}01 = 0{,}33$$

chování výběrového průměru

- nechť $X_1, X_2, \ldots X_n$ jsou nezávislé náhodné veličiny s libovolným rozdělením se střední hodnotou μ a rozptylem σ^2 - náhodný výběr z onoho rozdělení
- pro průměr z těchto veličin platí

$$\mathsf{E}\left(\bar{X}\right) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu \qquad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

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

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

výběrový průměr z normálního rozdělení

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

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

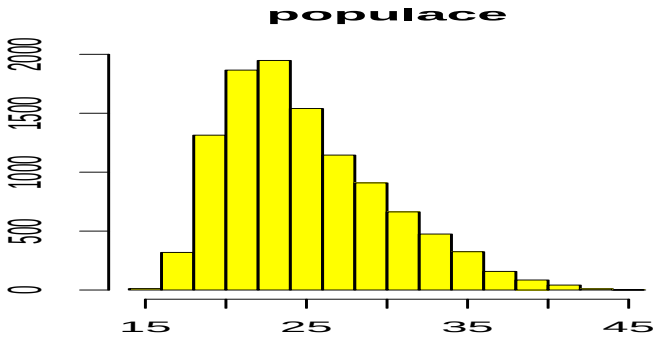
- opět je střední chyba $\bar{X}$ rovna $\frac{\sigma}{\sqrt{n}}$
- proto je

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

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

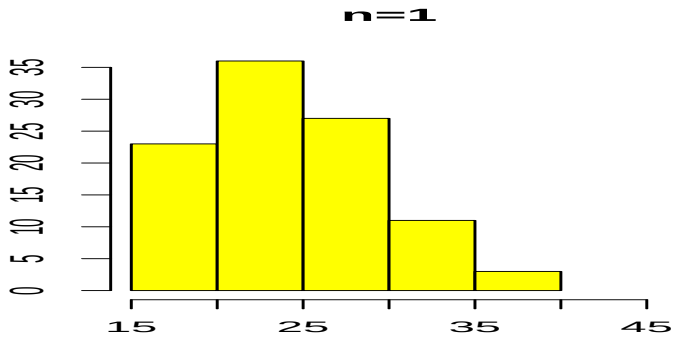
příklad: věk matek

- velká populace rodičů (11 tisíc)



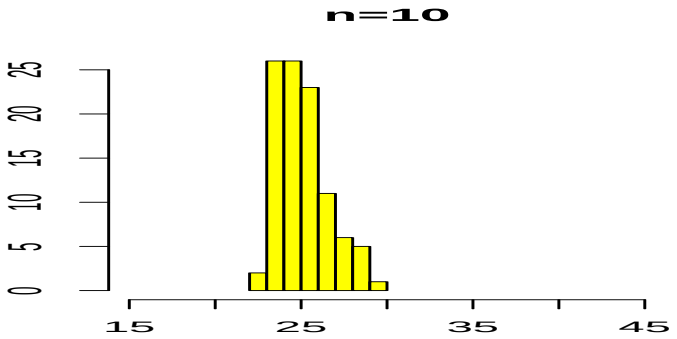
příklad: věk matek

- náhodně vybráno 100 matek



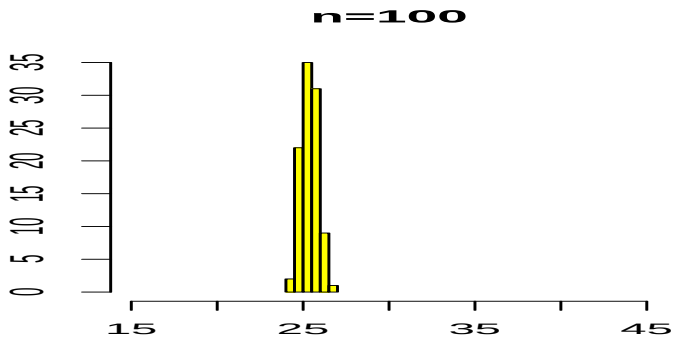
příklad: věk matek

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



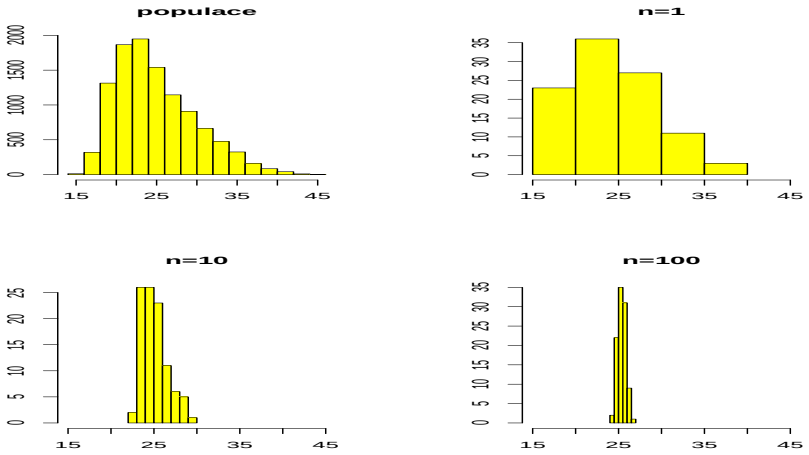
příklad: věk matek

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



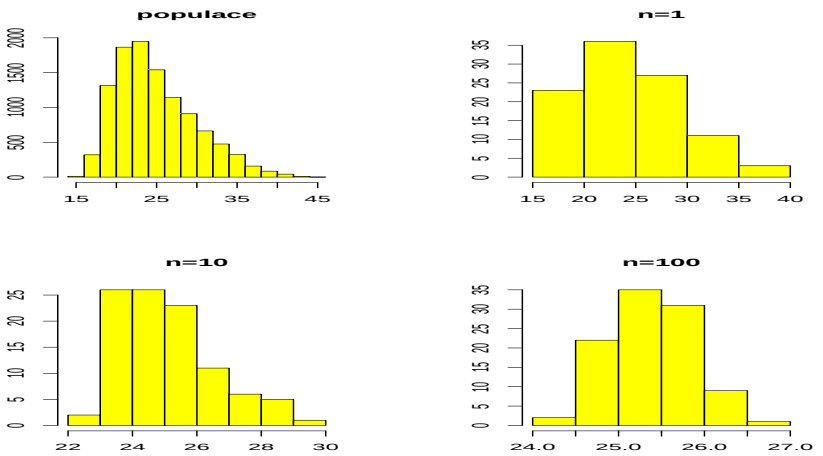
příklad: věk matek

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



centrální limitní věta

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



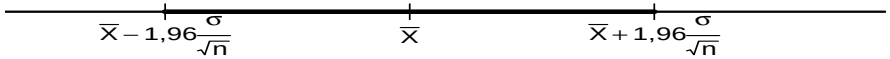
interval spolehlivosti (1)

- protože je $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, platí

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

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

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



interval spolehlivosti (2)

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

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

interval spolehlivosti (3)

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

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

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

příklad: věk matek

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

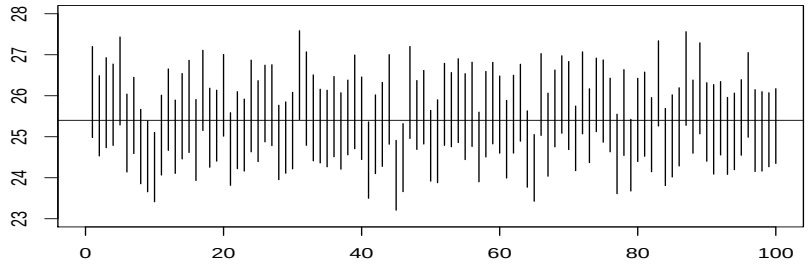
$$\left(25{,}7 - 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}9; 26{,}5)$$

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

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

- větší jistota $\Leftrightarrow$ větší šířka

příklad: simulované výběry pro $n = 100$



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

centrální limitní věta pro četnosti

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

interval spolehlivosti pro podíl (1)

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

interval spolehlivosti pro podíl (2)

- střední chyba relativní četnosti = směrodatná odchylka relativní četnosti = odmocnina z rozptylu je tedy $\sqrt{\frac{\pi(1-\pi)}{n}}$
- pravděpodobnost π neznáme, odhadneme ji pomocí relativní četnosti 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

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

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

připomeňme příklad s hrací kostkou

- odhadujeme pravděpodobnost šestky
- kostka A: $n = 100, n_A = 17, f_A = 0{,}17, \Rightarrow 95\%$ int. spol.$(0{,}10; 0{,}24)$
- kostka B: $n = 100, n_B = 41, f_B = 0{,}41 \Rightarrow 95\%$ int. spol.$(0{,}31; 0{,}51)$
- důležitý rozdíl: u kostky A patří $1/6 = 0{,}167$ do intervalu spolehlivosti; u kostky B nikoliv
- co se dá z tohoto zjištění usoudit?
- použijeme **testování hypotéz**

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 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

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 (falešně něco věcně prokazovat), proto zvolíme nízkou hladinu testu α (nejčastěji $\alpha = 5\%$)
- **hladina testu** α = maximální přípustná pravděpodobnost chyby 1. druhu
- **síla testu** = pravděpodobnost správného zamítnutí neplatné hypotézy

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 β)

postup při rozhodování

- zvolit hypotézu H_0, alternativu H_1
- zvolit hladinu testu α
- zvolit metodu rozhodování (test)
- z dat spočítat testovou statistiku T a porovnat ji s tabelovanou kritickou hodnotou
- když padne 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 zamítat

příklad: padá na kostce šestka příliš často?

- chceme na 5% hladině prokázat, že pravděpodobnost šestky je velká (tj. větší než 1/6)
- $\mathsf{H}_0 : \mathsf{P}(\text{padne šestka}) = 1/6 \qquad (= \pi_0)$
- $\mathsf{H}_1 : \mathsf{P}(\text{padne šestka}) > 1/6 \qquad (\neq \pi_0)$
- co svědčí pro neplatnost hypotézy?
  „šestka padá mnohem častěji, než by měla“
- provedeme $n = 100$ pokusů, Y počet šestek
- hypotézu budeme zamítat, když $Y > y_0$
- za platnosti H_0 má počet šestek Y rozdělení $\mathrm{bi}(n, 1/6)$
- y_0 zvolit tak, aby za hypotézy bylo $\mathsf{P}(Y > y_0) \leq \alpha$

příklad přesné volby kritického oboru

y_0	19	20	21	22	23
$\mathsf{P}(Y > y_0)$	0,220	0,152	0,100	0,063	0,038

- podmínku $\mathsf{P}(Y > y_0) \leq 0{,}05$ splňuje $y_0 = 23$
- padne-li ve 100 nezávislých hodech kostkou více než 23 šestek, budeme na **5% hladině zamítat hypotézu**, že pst šestky je 1/6 **ve prospěch alternativy**, že pst šestky je větší než 1/6 (dáno zvolenou alternativou)
- padlo nám $Y = Y_0 = 17$ šestek, hypotézu nezamítáme, což neznamená, že bychom hypotézu prokázali
- pro $\alpha = 10$ % bychom zvolili $y_0 = 21$

příklad: volba kritického oboru (přibližně)

- použijme přibližné tvrzení za H_0 $Y \sim \mathsf{N}(n\pi_0, n\pi_0(1-\pi_0))$, potom

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

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

p-hodnota

- spočítat k T_0 odpovídající p-hodnotu a porovnat ji s α
- **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)

příklad: rozhodování pomocí p-hodnoty

- snažíme se prokázat, že šestka padá příliš často
- padlo nám $Y_0 = 17$, proto (vzorec pro psti binomického rozdělení)

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

- protože 50,6 % $>$ 5 %, hypotézu nemůžeme na 5% hladině zamítnout, netvrdíme, že pst šestky je větší než 1/6
- neprokázali jsme však, že by hypotéza platila

příklad: kostka a oboustranná alternativa

- chceme ověřit, zda je kostka v pořádku
- pokusíme se prokázat, že šestka padá příliš často nebo příliš zřídka
- $\mathsf{H}_0 : \mathsf{P}(\text{padne šestka}) = 1/6$
- $\mathsf{H}_1 : \mathsf{P}(\text{padne šestka}) \neq 1/6$
- 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

příklad: kostka, oboustranná alternativa

y_0	9	10	11	...	23	24	25
$\mathsf{P}(Y < y_0)$	0,010	0,021	0,042	...	0,937	0,962	0,978
$\mathsf{P}(Y > y_0)$	0,979	0,957	0,922	...	0,038	0,022	0,012

- H_0 zamítneme, když bude $Y < 10$ *nebo* když bude $Y > 24$
- skutečná pst chyby 1. druhu bude 0,021 + 0,022 = 0,043
- hodnoty v rozmezí 10 až 24 (včetně obou mezí) nesvědčí proti H_0

oboustranná alternativa přibližně

- $\mathsf{H}_0 : \mathsf{P}(\text{padne šestka}) = 1/6 \qquad (= \pi_0)$
- $\mathsf{H}_1 : \mathsf{P}(\text{padne šestka}) \neq 1/6 \qquad (\neq \pi_0)$
- proti alternativě svědčí Y hodně daleko od $\mathsf{E}\,Y$, tj. rel. četnost $f = Y/n$ daleko od π_0:

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

- zamítáme tedy, je-li

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

nebo

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

znovu hodnocení četností

23	11	34
30	17	47
17	1	18
70	29	99

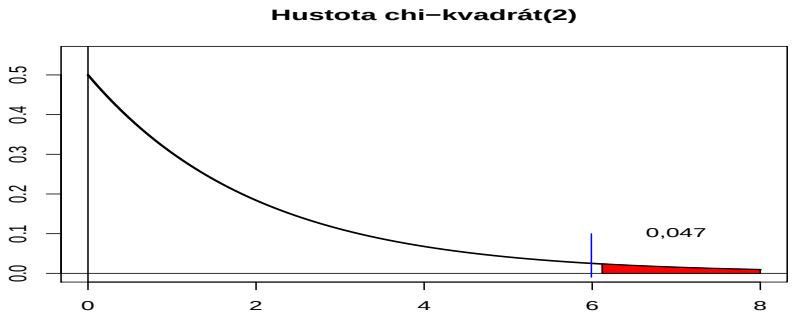
24,0	9,9	34
33,3	13,8	47
12,7	5,3	18
70	29	99

- statistika χ^2 porovnává skutečné četnosti (vlevo) s očekávanými (vpravo) za hypotézy

$$\chi^2 = \frac{(23-24{,}0)^2}{24{,}0} + \frac{(11-9{,}9)^2}{9{,}9} + \ldots + \frac{(1-5{,}3)^2}{5{,}3} = 6{,}12 > 5{,}99 = \chi^2_2(0{,}05)$$

- statistika χ^2 je větší, než kritická hodnota $\chi^2_2(0{,}05)$, závislost je na 5% hladině prokázána ($p = 0{,}047$)

grafická představa



- červená plocha = p-hodnota, modrá čára = hodnota statistiky

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 β)

příklad: výšky desetiletých hochů

- velký výběr v roce 1951 dal průměr $136{,}1$ cm, rozptyl $6{,}4^2$ cm^2
- v roce 1961 naměřeno v náhodném výběru $n = 15$ hodnot s průměrem $\bar{X} = 139{,}13$ 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 $H_0 : \mu = \mu_0 = 136{,}1$ (nebo $\mu \leq 136{,}1$, postup by byl stejný)
- alternativa $H_1 : \mu > 136{,}1$

výšky desetiletých hochů

- alternativě nasvědčují průměry o hodně větší než $\mu = \mu_0 = 136{,}1$
- 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)

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

- proto hypotézu zamítáme, je-li $Z > z(0{,}05) = 1{,}645$
- v našem příkladu je $Z_0 = \frac{139{,}13 - 136{,}1}{6{,}4}\sqrt{15} = 1{,}82 > 1{,}645$, takže na 5% hladině hypotézu **zamítáme ve prospěch jednostranné alternativy, že populační průměr za deset roků vzrostl**

obecně (jednostranná alternativa)

- $\mathsf{H}_0 : \mu = \mu_0$
- $\mathsf{H}_1 : \mu > \mu_0$
- 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)

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

- proto hypotézu zamítáme na hladině α, je-li $Z > z(\alpha)$
- $\mathsf{H}_1 : \mu < \mu_0$, podobně jako výše hypotézu zamítáme na hladině α, je-li $Z < -z(\alpha)$

obecně

- $\mathsf{H}_0 : \mu = \mu_0$
- $\mathsf{H}_1 : \mu \neq \mu_0$
- kritický obor: $\bar{X}$ je příliš daleko od μ_0,
- platí (za platnosti hypotézy)

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

- protože hladinu musíme rozdělit na dvě části ($\bar{X} << \mu_0$ a $\bar{X} >> \mu_0$ hypotézu zamítáme na hladině α, je-li $|Z| > z(\alpha/2)$

výšky desetiletých hochů

- výpočet p-hodnoty: (v Z odečítáme vždy skutečně platnou střední hodnotu, uvedeme ji jako dolní index u P)
  za H_0 je $Z = \frac{\bar{X} - 136{,}1}{6{,}4}\sqrt{15} \sim \mathsf{N}(0,1)$

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

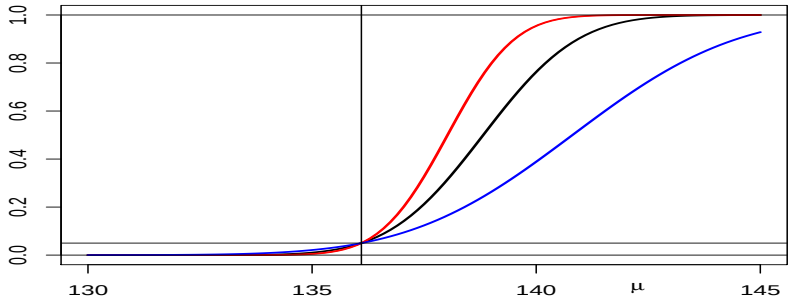
- na 5% hladině jsme zamítli hypotézu ve prospěch jednostranné alternativy (kterou jsme zvolili předem, bez znalosti dat!)
- prokázali jsme na 5% hladině vzrůst populačního průměru

síla testu pro $\mu = 140$

- síla testu = pst(zamítnout hypotézu, když tato neplatí)
- musíme vzít v úvahu, že $\mu = 140$

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

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



shrnutí: $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$, nezávislé

- předpokládáme, že $\sigma > 0$ známe
- $\mathsf{H}_0 : \mu = \mu_0$ (μ_0 známá konstanta)

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

- kdy hypotézu H_0 zamítáme (kritický obor):
  - $\mathsf{H}_1 : \mu \neq \mu_0$ (oboustranná alternativa) $\quad |Z| \geq z(\alpha/2)$
  - $\mathsf{H}_1 : \mu > \mu_0$ (jednostranná alternativa) $\quad Z \geq z(\alpha)$
  - $\mathsf{H}_1 : \mu < \mu_0$ (jednostranná alternativa) $\quad Z \leq -z(\alpha)$

častěji: $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$, nezávislé

- neznámé $\sigma > 0$ odhadneme pomocí $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$
- $\mathsf{H}_0 : \mu = \mu_0$ (μ_0 známá konstanta)

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

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

souvislost s intervalem spolehlivosti

- připomeňme interval spolehlivosti pro μ

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

  což lze přepsat jako

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

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

výšky desetiletých hochů (σ^2 neznámé)

- kritický obor: $\bar{X}$ se příliš liší od μ_0 ve směru zvolené alternativy
- spočítáme

$$s = \sqrt{\frac{1}{15-1}((130 - 139{,}13)^2 + \ldots + (141 - 139{,}13)^2)} = \sqrt{42{,}98} = 6{,}56$$
$$T = \frac{\bar{X} - 136{,}1}{6{,}56} \sqrt{15} = 1{,}79$$

- 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: $(135{,}5; 142{,}8)$

nová úloha: porovnání dvou populací

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

$$X_i \sim \mathrm{N}\big(\mu_1, \sigma^2\big), \quad i = 1, \ldots, n_1$$

- lze předpokládat, že výšky dívek

$$Y_i \sim \mathrm{N}\big(\mu_2, \sigma^2\big), \quad i = 1, \ldots, n_2$$

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

porovnání středních hodnot nezávislých výběrů

- zřejmě $\mathrm{H}_0 : \mu_1 = \mu_2$ (není rozdíl, **nulová** hypotéza)
- alternativy
  - $\mathrm{H}_1 : \mu_1 \neq \mu_2$ (není-li důvod k jednostranné alternativě)
  - $\mathrm{H}_1 : \mu_1 > \mu_2$ (bylo cílem dokázat, že hoši jsou větší dívek)
  - $\mathrm{H}_1 : \mu_1 < \mu_2$ (bylo cílem dokázat, že hoši jsou menší dívek)
- rozhodování založeno na porovnání průměrů $\bar{X}$ a $\bar{Y}$; čím více se liší, tím spíše zamítnout hypotézu
- je třeba porovnat s mírou přesnosti, s jakou průměry $\bar{X}, \bar{Y}$ odhadnou skutečné populační průměry μ_1, μ_2

porovnání středních hodnot nezáv. výběrů (2)

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

$$s^2 = \frac{1}{n_1+n_2-2}\left(\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2\right)$$
$$= \frac{n_1-1}{n_1+n_2-2}s_X^2+\frac{n_2-1}{n_1+n_2-2}s_Y^2$$

  (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 $s^2 = 38{,}94 = 6{,}24^2$

kritické obory

- o $\mathsf{H}_0 : \mu_1 = \mu_2$ se rozhoduje pomocí

$$T = \frac{\bar{X}-\bar{Y}}{s}\sqrt{\frac{n_1 n_2}{n_1+n_2}} = \frac{\bar{X}-\bar{Y}}{\widehat{\mathsf{S.E.}}(\bar{X}-\bar{Y})}$$

- $\mathsf{H}_1 : \mu_1 \neq \mu_2$ zamítáme pokud $|T| \geq t_{n_1+n_2-2}(\alpha)$
- $\mathsf{H}_1 : \mu_1 > \mu_2$ zamítáme pokud $T \geq t_{n_1+n_2-2}(2\alpha)$
- $\mathsf{H}_1 : \mu_1 < \mu_2$ zamítáme pokud $T \leq -t_{n_1+n_2-2}(2\alpha)$
- výšky desetiletých: $T = -0{,}70 < 2{,}06 = t_{15+12-2}(0{,}05)$
- na 5% hladině jsme **neprokázali** rozdíl mezi výškami desetiletých hochů a dívek ($p = 48{,}8$ %)

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 delta je

  $(\bar{X}-\bar{Y}) - \widehat{\mathsf{S.E.}}(\bar{X}-\bar{Y})\cdot t_{n_1+n_2-2}(\alpha) < \delta < (\bar{X}-\bar{Y}) + \widehat{\mathsf{S.E.}}(\bar{X}-\bar{Y})\cdot t_{n_1+n_2-2}(\alpha)$

  H_0 zamítáme právě tehdy, když nula **není** v int. spol. pro δ
- při porovnání výšek hochů a dívek je 95% interval pro δ

$$\left(-1{,}7 - 6{,}24\sqrt{\frac{1}{15}+\frac{1}{12}}\cdot 2{,}06;\ -1{,}7 + 6{,}24\sqrt{\frac{1}{15}+\frac{1}{12}}\cdot 2{,}06\right)$$
$$(-3{,}3;\ 6{,}7)$$

příklad: přijímačky na MFF

- liší se úrovní znalosti matematiky uchazeči o studium matematiky a fyziky?
- $n_1 = 104$, $\bar{X} = 34{,}6$, $s_x = 11{,}4 = \sqrt{129{,}4}$,
  $n_2 = 114$, $\bar{Y} = 31{,}2$, $s_x = 10{,}8 = \sqrt{117{,}2}$
  $s = 11{,}1 = \sqrt{123{,}0}$
- dostaneme tedy

$$T = \frac{34{,}6 - 31{,}2}{11{,}1}\sqrt{\frac{104\cdot 114}{104+114}} = 2{,}25$$

- na 5% hladině jsme prokázali rozdíl mezi dvěma skupinami uchazečů ($p = 2{,}6$ %)
- 95% interval spolehlivosti pro rozdíl populačních průměrů je $(0{,}4;\ 6{,}3)$

porovnání středních hodnot nezávislých výběrů

- zřejmě $\mathsf{H}_0 : \mu_1 = \mu_2$ (není **žádný** rozdíl mezi populacemi, **nulová** hypotéza)
- alternativy
  - $\mathsf{H}_1 : \mu_1 \neq \mu_2$ (není-li důvod k jednostranné alternativě)
  - $\mathsf{H}_1 : \mu_1 > \mu_2$ (bylo cílem dokázat, že hoši X větší dívek Y)
  - $\mathsf{H}_1 : \mu_1 < \mu_2$ (bylo cílem dokázat, že hoši X menší dívek Y)
- rozhodování založeno na porovnání průměrů $\bar{X}$ a $\bar{Y}$; čím více se liší, tím spíše zamítnout hypotézu
- je třeba porovnat s mírou přesnosti, s jakou rozdíl průměrů $\bar{X} - \bar{Y}$ odhadne skutečný rozdíl populačních průměrů $\mu_1 - \mu_2$

kritické obory dvouvýběrového t-testu

- o hypotéze $\mathsf{H}_0 : \mu_1 = \mu_2$ se rozhoduje pomocí

$$T = \frac{\bar{X} - \bar{Y}}{\widehat{\mathsf{S.E.}}(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

- $\mathsf{H}_1 : \mu_1 \neq \mu_2$ zamítáme pokud $|T| \geq t_{n_1+n_2-2}(\alpha)$
- $\mathsf{H}_1 : \mu_1 > \mu_2$ zamítáme pokud $T \geq t_{n_1+n_2-2}(2\alpha)$
- $\mathsf{H}_1 : \mu_1 < \mu_2$ zamítáme pokud $T \leq -t_{n_1+n_2-2}(2\alpha)$
- s^2 je odhad σ^2 založený na obou výběrech

$$s^2 = \frac{n_1 - 1}{n_1 + n_2 - 2}s_X^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}s_Y^2$$

provedení v MS Excelu (stejné rozptyly)

	Excel	Soubor 1	Soubor 2
průměr	Stř. hodnota	139.133	140.833
rozptyl	Rozptyl	42.981	33.788
rozsah	Pozorování	15	12
spol. odhad rozpt.	Společný rozptyl	38.936	
$\mathsf{H}_0 : \mu_1 - \mu_2 =$	Hyp. rozdíl stř. hodnot	0	
stupně vol.	Rozdíl	25	
T	t stat	-0.733	
p jednostr. testu	P(T<=t) (1)	0.244	
$t_{n_1+n_2-2}(2\alpha)$	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

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 $\mathsf{H}_0 : \sigma_1^2 = \sigma_1^2$ se proti $\mathsf{H}_1 : \sigma_1^2 \neq \sigma_2^2$ 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
- Excel: uvádí kritickou hodnotu a p-hodnotu pro jednostrannou alternativu; p-hodnotu je třeba vynásobit dvěma

MS Excel: Dvouvýběrový F-test pro rozptyl

		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	

ve skutečnosti je $\mathsf{P}(F > 1{,}27) = 0{,}349$, takže $p = 2 \cdot 0{,}349 = 0{,}698$
pro oboustrannou alternativu bylo použito $F_{14,11}(0{,}025) = 3{,}359$

dvouvýběrový t-test při nestejných rozptylech

- není-li udržitelný předpoklad o stejných rozptylech, lze použít přibližný t-test (Welchův, jiný odhad S.E.$(\bar{X} - \bar{Y})$)

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

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

$$s_{\bar{X}-\bar{Y}} = \sqrt{v_1 + v_2} \qquad v_1 = \frac{s_X^2}{n_1} \qquad v_2 = \frac{s_Y^2}{n_2}$$

- H_0 se zamítá, je-li $|T| \geq t_f(\alpha)$, kde $f = \frac{s^4_{\bar{X}-\bar{Y}}}{\frac{v_1^2}{n_1-1} + \frac{v_2^2}{n_2-1}}$
- náš příklad $T = -0{,}713$, $f = 24{,}69$, $t_f(0{,}05) = 2{,}061$, $p = 0{,}482$

provedení v MS Excelu (nestejné rozptyly)

		Soubor 1	Soubor 2
průměr	Stř. hodnota	139.133	140.833
rozptyl	Rozptyl	42.981	33.788
rozsah	Pozorování	15	12
$\mathsf{H}_0 : \mu_1 - \mu_2 =$	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

souvislost s bodově biseriálním korel. koef.

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

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

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

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

- stejný vztah platí i pro (Pearsonův) korelační koeficient r
- náš příklad $r_{\mathsf{bis}} = -0{,}139$

porovnání podílů (příklad)

- podíl matek, které označily těhotenství za plánované:
  matky jen se základním vzděláním: 14 z 34 (41,2 %)
  matky aspoň s maturitou: 44 z 65 (67,7 %)
  matky bez ohledu na vzdělání: 58 z 99 (58,6 %)
- $n_1 = 34, f_1 = 0{,}412, n_2 = 65, f_2 = 0{,}677, f = 0{,}586$
- nutno odhadnout rozptyl $f_1 - f_2$:

$$\widehat{\mathrm{var}}\,(f_1 - f_2) = \widehat{\mathrm{var}}\, f_1 + \widehat{\mathrm{var}}\, f_2 = f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$Z = \frac{0{,}412 - 0{,}677}{\sqrt{0{,}586 \cdot 0{,}414\left(\frac{1}{34} + \frac{1}{65}\right)}} = -2{,}54 \qquad p = 2(1 - \Phi(|-2{,}54|)) = 1{,}1\ \%$$

porovnání podílů

- dva nezávislé výběry
  Y_1 absolutní četnost jevu v prvním výběru rozsahu n_1
  Y_2 absolutní četnost jevu ve druhém výběru rozsahu n_2
- hypotéza $\mathsf{H}_0 : \pi_1 = \pi_2$, tj. podíly jevu v obou populacích stejné
- statistika Z porovnává relativní četnosti $f_1 = Y_1/n_1, f_2 = Y_2/n_2$

$$Z = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad f = \frac{Y_1 + Y_2}{n_1 + n_2}$$

$\mathsf{H}_1 : \pi_1 \neq \pi_2$ zamítat pro $|Z| \geq z(\alpha/2)$
$\mathsf{H}_1 : \pi_1 < \pi_2$ zamítat pro $Z \leq -z(\alpha)$
$\mathsf{H}_1 : \pi_1 > \pi_2$ zamítat pro $Z \geq z(\alpha)$

souvislost s čtyřpolním korel. koef.

	výskyt jevu		
výběr	ano	ne	celkem
1	Y_1	$n_1 - Y_1$	n_1
2	Y_2	$n_2 - Y_2$	n_2
celkem	$Y_1 + Y_2$	$n_1 + n_2 - Y_1 - Y_2$	$n_1 + n_2$

- pomocí čtyřpolního korelačního koeficientu $r_{2,2}$ lze Z zapsat jako $Z = \sqrt{n_1 + n_2}\ r_{2,2}$
- příklad

14	20	34
44	21	65
58	41	99

$r_{2,2} = \frac{14 \cdot 21 - 44 \cdot 20}{\sqrt{34 \cdot 65 \cdot 58 \cdot 41}} = -0{,}256,$

$Z = \sqrt{99} \cdot (-0{,}256) = -2{,}54,\ \chi^2 = 99 \cdot (-0{,}256)^2 = 6{,}47,\ p = 1{,}1\ \%$

Mannův-Whitneyův (Wilcoxonův) test

- co když nelze předpokládat normální rozdělení?
- nechť $X_1, \ldots, X_{n_1}$ a $Y_1, \ldots, Y_{n_2}$ 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 průměrná pořadí v obou výběrech podobná

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
- 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

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 $Z \sim \mathrm{N}(0,1)$
- hypotézu zamítáme, je-li $|Z| \geq z(\alpha/2)$
- náš příklad:

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

- na 5% hladině jsme prokázali rozdíl

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 $1 + 2 + \ldots + n_1 + n_2$)
- stačí zabývat se jedinou ze statistik W_1, W_2, zpravidla tou pro menší výběr

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 $p = 24/715 = 3{,}4$ %

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

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

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

párový t-test:

- nechť $(Y_1, Z_1) \ldots, (Y_n, Z_n)$ nezávislé dvojice, $X_i = Y_i - Z_i$
- nechť $X_i \sim \mathsf{N}\left(\mu, \sigma^2\right)$
- neznámé $\sigma > 0$ odhadneme pomocí $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- $\mathsf{H}_0 : \mu = \mu_0$ (μ_0 známá konstanta, zpravidla 0)

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

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

příklad: klesá potratovost?

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)

párový ilcoxonův test

- nechť $(Y_1, Z_1) \ldots, (Y_n, Z_n)$ nezávislé dvojice, $X_i = Y_i - Z_i$
- $H_0 : Y_i, Z_i$ mají stejné rozdělení (populace jsou stejné)
- mají-li Y_i, Z_i stejné rozdělení, pak $X_i = Y_i - Z_i$ jsou symetricky rozdělena 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

rozhodování

- na základě centrální limitní věty lze použít

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

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

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

připomenutí příkladu: klesá potratovost?

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)

párový ilcoxonův (ilcoxon signed rank) test

- nechť $(Y_1, Z_1) \ldots, (Y_n, Z_n)$ nezávislé dvojice, $X_i = Y_i - Z_i$
- $H_0 : Y_i, Z_i$ 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

rozhodování

- na základě centrální limitní věty lze použít

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

- 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 $p = 2{,}6$ %))

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

poznámky k výpočtu

- nezapomenout vyloučit nulové rozdíly
- shodným absolutním hodnotám rozdílům přiřadíme jejich průměrné pořadí:
- Excel 2000 řeší problém shod nestandardně (bohužel)
- jednoduchá ukázka

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

párový znaménkový (sign) test

- hodnotí pouze **počet** kladných a záporných rozdílů, nezáleží na tom, jak jsou rozdíly veliké (slabší test než Wilcoxonův)
- $\mathsf{H}_0 : Y_i, Z_i$ 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 $Y \sim \mathsf{bi}(n, 1/2)$
- přibližné rozhodování (centrální limitní věta)

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

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

poznámky

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

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

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

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

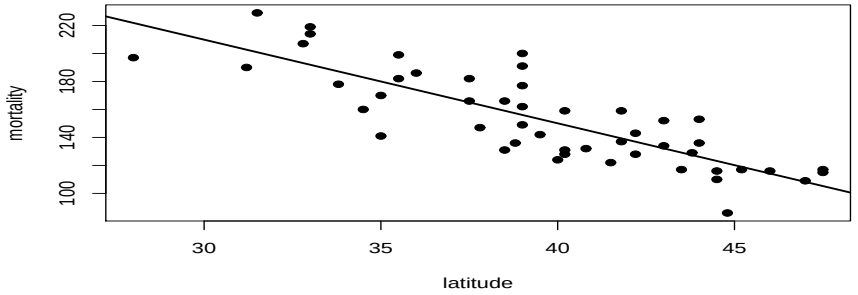
- náš příklad (Yatesova korekce, jiným způsobem přesně p =0,194)

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

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

příklad: souvisí úmrtnost se zeměpisnou šířkou?



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

regresní přímka

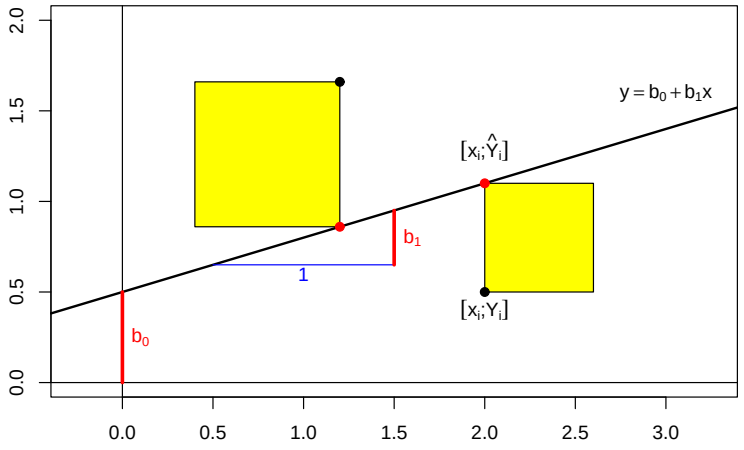
- chování Y (úmrtnost, mortality) vysvětlit lineární závislostí na x (zeměpisná šířka, latitude)
- každé zem. šířce odpovídá jakási střední úmrtnost, ta závisí na zeměpisné šířce lineárně

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

- 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



náš příklad

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 est(mortality) = 389,2 − 5,98 latitude
- s každým stupněm sev. šířky klesá úmrtnost v průměru 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

obecně

- odhadnutá závislost $y = b_0 + b_1 x$, modelová $y = \beta_0 + \beta_1 x$
- závislost na x prokazujeme testováním hypotézy $\mathsf{H}_0 : \beta_1 = 0$ proti oboustranné alternativě pomocí

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

- **reziduální součet čtverců** – **nevysvětlená** variabilita odezvy $S_e = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 x_i))^2$ 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

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

náš příklad a tabulka analýzy rozptylu

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 %

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

- na 30. stupni očekáváme úmrtnost 389,19 − 5,98 · 30 = 209,86, na 40. stupni očekáváme úmrtnost 389,19 − 5,98 · 40 = 150,08

můžeme predikci zlepšit?

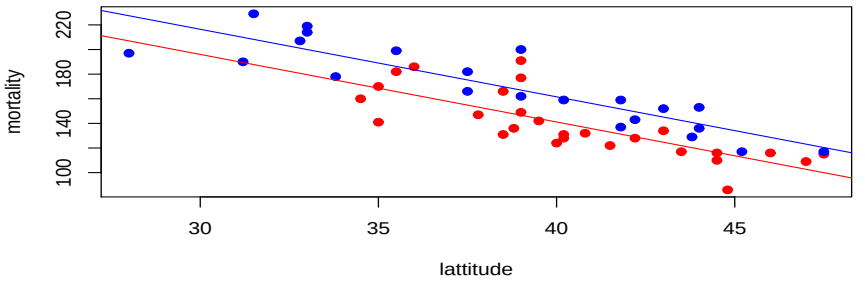
koef.	odhad	stř. chyba	t-stat.	p
abs. člen	401,17	28,04	14,31	<0,001
latitude	− 5,93	0,60	− 9,82	<0,001
longitude	0,15	0,19	0,82	0,418

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

můžeme predikci zlepšit?

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



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

pozor na interpretaci odhadů (příklad)

- 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) = − 55,91 + 0,391 height R^2 = 14,0 %
- est(fat) = 13,29 − 0,273 height + 0,627 weight R^2 = 69,5 %
- ve všech případech jsou koeficienty na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky (a **nezměněné hmotnosti** pro druhý model)

regrese v MS Excelu 2000

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

obecné předpoklady

- **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

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

použití reziduí

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

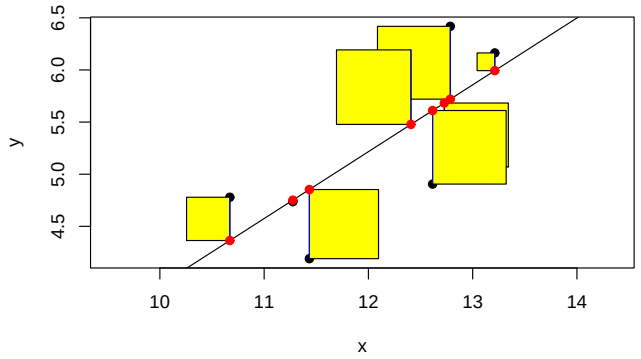
$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- 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



Y_i, $\hat{Y}_i$

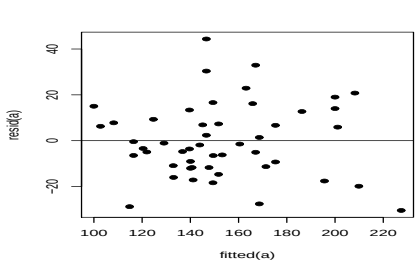
diagnostika pomocí reziduí

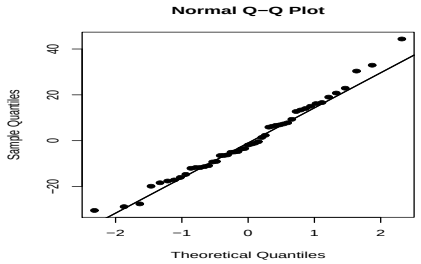
- histogram reziduí nebo normální diagram (k ověření normálního rozdělení)
- grafické znázornění bodů $[\hat{Y}_i, u_i]$ nebo $[x_i, u_i]$ (k ověření konstantního rozptylu či tvaru závislosti)

ukázky diagnostiky

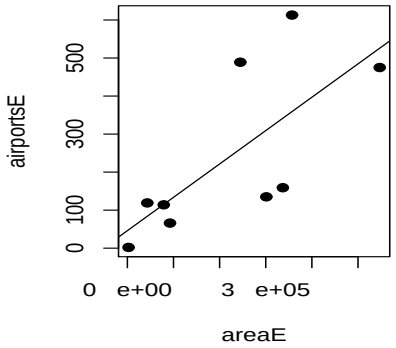
vlevo: rezidua spíše kladná než záporná, možná jsme měli raději vysvětlovat odmocninu z mortality

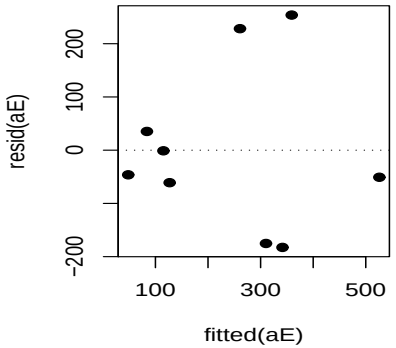
vpravo: normální diagram, ukazuje, že s předpokladem o normálním rozdělením není problém (body těsně kolem přímky)



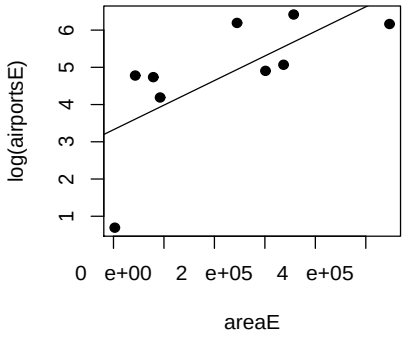


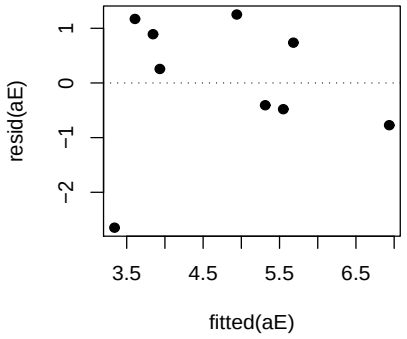
nekonstantní rozptyl (trychtýřovité rozšiřování mraku reziduí)



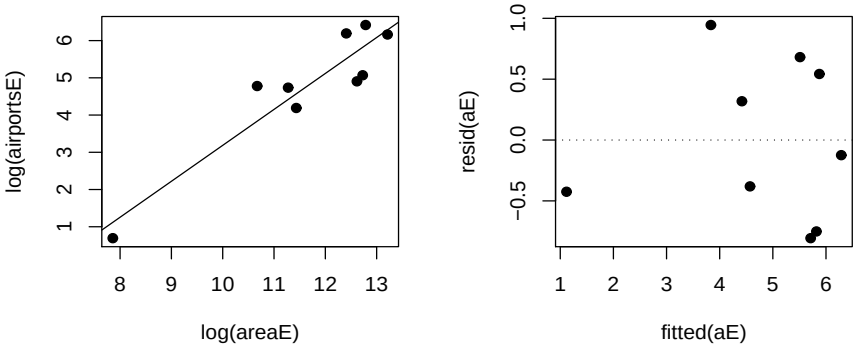


odlehlé pozorování (první pozorování daleko od vodorovné osy)

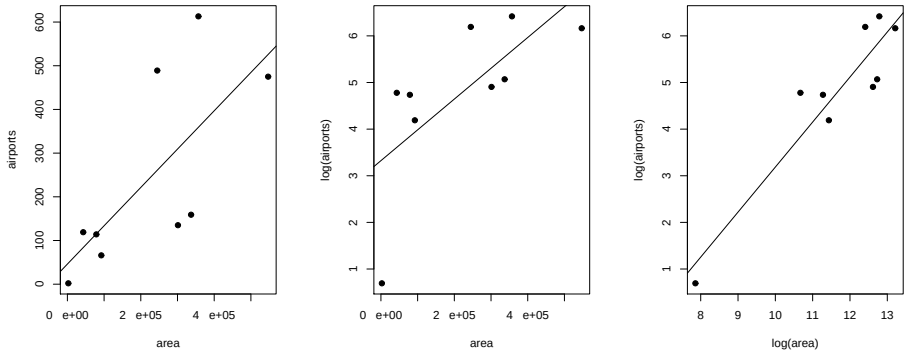


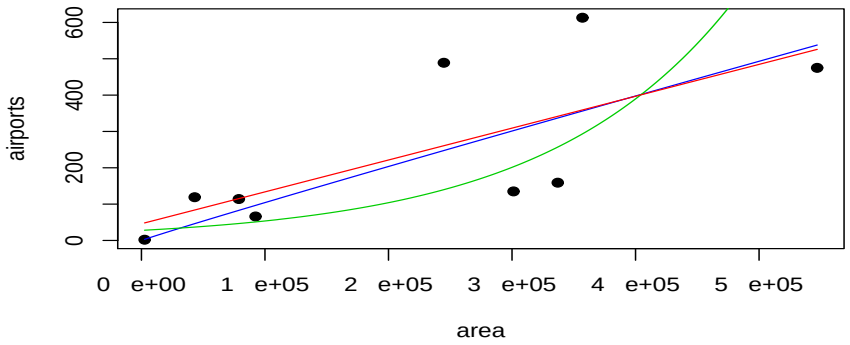


vlivné pozorování (první pozorování daleko ve vodorovném směru)



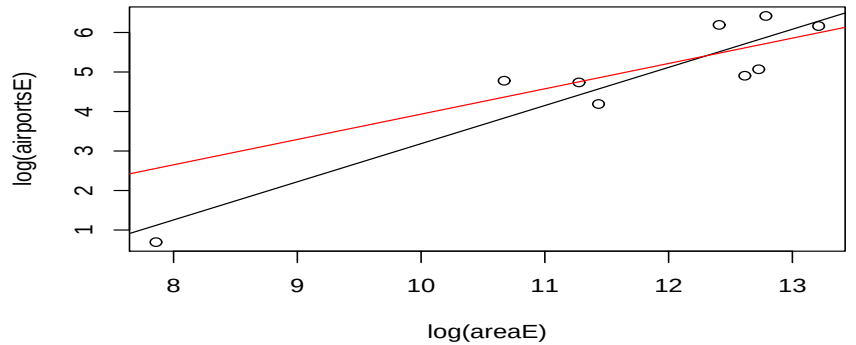
ukázka transformace: počet letišť na ploše evropského státu



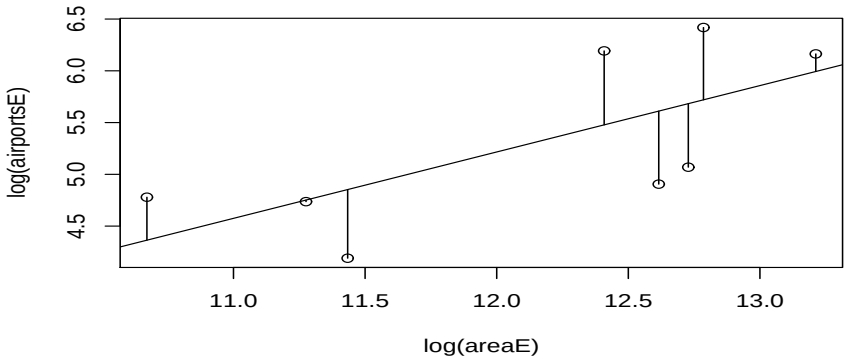


y=46+0,0009 x; log(y)=3,3+0,000007 x; log(y)=-6,5+0,97 log(x)
$R^2 = 51{,}4$ %; $R^2 = 48{,}0$ %; $R^2 = 86{,}1$ %

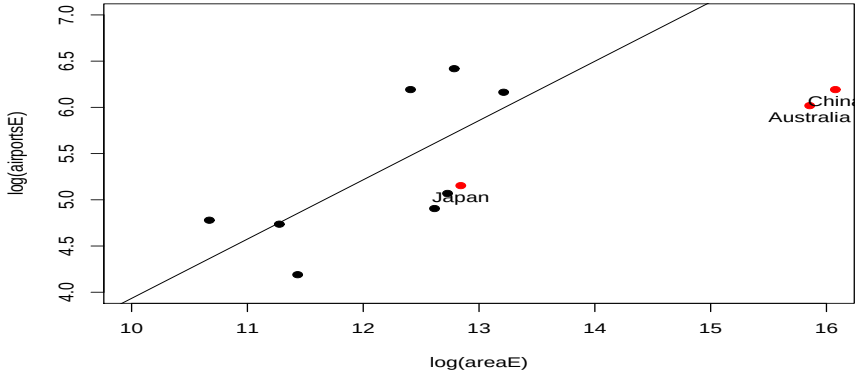
počet letišť: vynechat Lucembursko?



počet letišť: po vynechání Lucemburska



stejná závislost pro Japonsko, Čínu, Austrálii?



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)

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 %

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 předpokládat $r_S\sqrt{n-1} \sim \mathsf{N}(0,1)$
- závislost (proti oboustranné alternativě) prokázána, pokud

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

- závislost (proti jednostranné alternativě) prokázána, pokud

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

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)

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

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

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

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

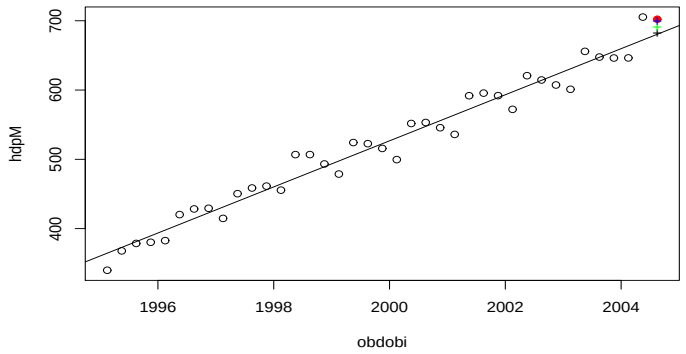
vyrovnávání

- mechanismus k vyhlazování dat, spíše technický
- cílem je např. nahradit chybějící pozorování (budoucí pozorování – **predikce**, chybějící pozorování - **interpolace**) nebo odstranit nahodilé výkyvy
- často (náš příklad) porušen předpoklad nezávislosti pozorování, proto by obyčejná regrese dala nesprávně přesnost odhadů, testy o parametrech,
- koeficienty nelze snadno statisticky hodnotit, někdy vůbec
- nejen metoda nejmenších čtverců

časové řady

- spojitý znak měřený v pravidelných časových intervalech
- složeno 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

příklad: HDP v ČR po čtvrtletích (běžné ceny)



příklad: HDP [mil. Kč]

- predikce bez ohledu na kvartály: 360,4 + 33,3 (rok-1995)
  pro 3. čtvrtletí 2004 tedy předpověď 680,5
- predikce s ohledem na kvartály

$$\begin{aligned} est(HDP) = & \quad 339,4 + 33,1(\text{rok} - 1995) \\ est(HDP) = & \quad (335,3 + 38,5) + 33,1(\text{rok} - 1995) \\ est(HDP) = & \quad (335,3 + 30,2) + 33,1(\text{rok} - 1995) \\ est(HDP) = & \quad (335,3 + 18,1) + 33,1(\text{rok} - 1995) \end{aligned}$$

  předpověď tedy 335,3 + 30,2 + 33,1 · (2004,75−1995) = 688,6
- predikce s ohledem na autokorelaci (odhad autokorelačního koeficientu $\hat{\rho}$ =0,59): 688,6 + 0,59 · 16,7 = 698,4
- skutečnost: 702,2

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
  - ρ =0: po sobě jdoucí členy nezávislé (v regresi se požaduje)
- autokorelaci a periodicitu lze prokazovat (odhadovat, hodnotit) až po odstranění trendu a sezónního kolísání

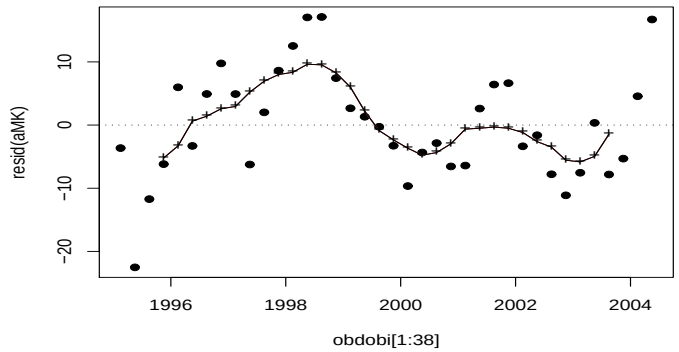
klouzavé průměry (moving averages)

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

příklad: klouzavé průměry reziduí (m =7)



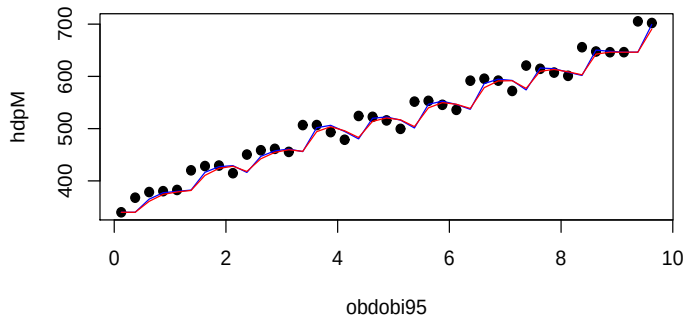
exponenciální vyrovnávání

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

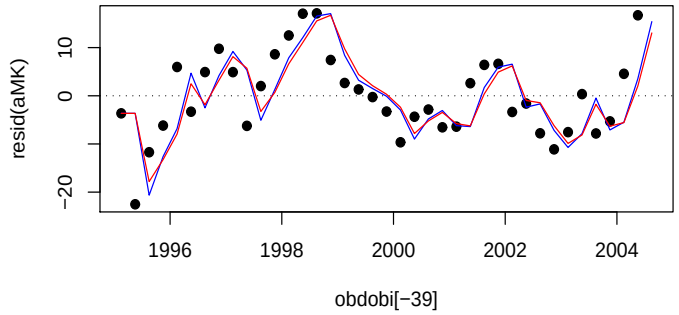
$$w\hat{Y}_{i-1} + (1-w)Y_{i-1}$$

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

příklad: exponenciální vyrovnávání (699,4 pro w=0,1, 690,6 pro w=0,25)



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



multinomické rozdělení

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

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

příklady multinomického rozdělení

- předvolební průzkum
  - n – počet tázaných
  - π_j – skutečný podíl voličů j-té strany v populaci
  - X_j – počet (četnost) voličů j-té strany ve výběru
- hody hrací kostkou
  - n – počet hodů
  - $\pi_1, \ldots, \pi_6$ – pravděpodobnosti jednotlivých stran kostky
  - $X_1, \ldots, X_6$ – absolutní četnosti jednotlivých stran kostky

vlastnosti multinomického rozdělení

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

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

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

příklad: hrací kostka A

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

$$\chi^2 = \frac{(12 - 16{,}67)^2}{16{,}67} + \ldots + \frac{(17 - 16{,}67)^2}{16{,}67} = 4{,}16 < \chi^2_5(0{,}05) = 11{,}07$$

$$p = 52{,}7\ \%$$

příklad: hrací kostka B (1)

- $n = 100$ hodů kostkou
- $X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15, X_6 = 41$
- očekávané četnosti za hypotézy $\pi_1 = \ldots = \pi_6 = 1/6$:
  $n\pi_1 = \ldots = n\pi_6 = 100/6 = 16{,}67$

$$\chi^2 = \frac{(15-16{,}67)^2}{16{,}67} + \ldots + \frac{(41-16{,}67)^2}{16{,}67} = 48{,}32 > \chi^2_5(0{,}05) = 11{,}07$$

- zřejmě je nutno zamítnout hypotézu, že kostka je symetrická

příklad: hrací kostka B (2)

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

$$\chi^2 = \frac{(15-10)^2}{10} + \ldots + \frac{(15-10)^2}{10} + \frac{(41-50)^2}{50} = 12{,}72 > \chi^2_5(0{,}05) = 11{,}07$$

- zřejmě je nutno zamítnout i tuto hypotézu

příklad: hrací kostka B (3)

- $n = 100$ 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 jediné četnosti – 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

test homogenity r výběrů

- $X_{i1}, \ldots, X_{ik}$ i-tý výběr z multinomického rozdělení s parametry $n_{i\bullet}, \pi_{i1}, \ldots, \pi_{ik} \qquad (i = 1, \ldots, r)$
- H_0 : pravděpodobnosti jsou ve všech srovnávaných populacích stejné: $\pi_{i1} = \pi_1, \ldots, \pi_{ik} = \pi_k$ (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
- očekávané četnosti tak budou $o_{ij} = n_{i\bullet}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet} n_{\bullet j}}{n}$

test homogenity r výběrů

- 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 výsledné chí-kvadrát rozdělení chí-kvadrát s $(r-1)(k-1)$ stupni volnosti
- hypotézu o shodě pravděpodobností v r populacích zamítáme, je-li $\chi^2 \geq \chi^2_{(r-1)(k-1)}(\alpha)$

mají obě kostky stejné šestice pravděpodobností?

- empirické četnosti (kontingenční tabulka)

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

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

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

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

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

příklad: vzdělání matek

	vzdělání			
porodnice	základní	střední	VŠ	celkem
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

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

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

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

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 psti stejné)
- výpočet χ^2 a jeho hodnocení stejné jako u homogenity

příklad: plánovaná těhotenství

- 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

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

čtyřpolní tabulka

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

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

úmrtí 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

$$\chi^2 = \frac{3178(1097\cdot 354 - 362\cdot 1365)^2}{1459\cdot 1719\cdot 2462\cdot 716} = 8{,}05 > \chi^2_1(0{,}05) = 3{,}84 \qquad p = 0{,}5\ \%$$

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

zobecnění dvouvýběrového t-testu

(požadoval normální rozdělení v aspoň dvou nezávislých výběrech)

- příklad: souvisí výška otce se vzděláním matky?
  (součet čtverců = součet čtverců odchylek od průměru)

vzdělání	rozsah	průměr	součet čtverců	směr. odch.
základní	34	177,1	1188,7	6,0
střední	47	179,5	1909,8	6,4
VŠ	19	182,8	1027,1	7,8
celkem	99	179,3	4511,2	6,8

- **součet** součtů čtverců 1188,7+1909,8+1027,1=4125,6, kdežto když nehledíme na vzdělání, tak 4511,2, rozdíl je 385,6
- je to dost k tomu, aby bylo prokázáno, že se tři skupiny liší populačním průměrem?

model analýzy rozptylu (r **nezávislých** výběrů)

$$X_{11}, X_{12}, \ldots, X_{1n_1} \sim \mathsf{N}\left(\mu_1, \sigma^2\right) \quad \text{první výběr (skupina)}$$
$$X_{21}, X_{22}, \ldots, X_{2n_2} \sim \mathsf{N}\left(\mu_2, \sigma^2\right) \quad \text{druhý výběr (skupina)}$$
$$\ldots$$
$$X_{r1}, X_{r2}, \ldots, X_{rn_r} \sim \mathsf{N}\left(\mu_r, \sigma^2\right) \quad r\text{-tý výběr (skupina)}$$

- zobecnění hypotézy dvouvýběrového t-testu $\mathsf{H}_0 : \mu_1 = \ldots = \mu_r$
- rozhoduje se na základě rozkladu součtu čtverců

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{i=1}^{r} n_i(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\bullet})^2$$

variabilita: celková = mezi výběry + uvnitř výběrů

označení: $\bar{X}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \qquad \bar{X}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} X_{ij}$

$$\begin{aligned}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{\bullet\bullet})^2 &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\Big((X_{ij}-\bar{X}_{i\bullet})+(\bar{X}_{i\bullet}-\bar{X}_{\bullet\bullet})\Big)^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\bullet})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{X}_{i\bullet}-\bar{X}_{\bullet\bullet})^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\bullet})^2 + \sum_{i=1}^{r} n_i(\bar{X}_{i\bullet}-\bar{X}_{\bullet\bullet})^2\end{aligned}$$

protože

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\bullet})(\bar{X}_{i\bullet}-\bar{X}_{\bullet\bullet}) = \sum_{i=1}^{r}(\bar{X}_{i\bullet}-\bar{X}_{\bullet\bullet})\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\bullet}) = 0$$

tabulka analýzy rozptylu

variabilita	součet čtv.	st. vol.	prům. čtv.	F	p
mezi výběry	385,6	2	192,8	4,49	0,014
uvnitř výběrů	4125,6	96	43,0		
celková	4511,2	98			

- součet čtverců uvnitř skupin = součet součtů čtverců, stupně volnosti v řádku uvnitř skupin: $n - r$
- průměrné čtverce: vždy součet čtverců/stupně volnosti; platí-li H_0, pak obojí průměrné čtverce podobné; čím je prům. čtverec mezi skupinami větší, tím spíš H_0 zamítnout
- H_0 zamítnout, je-li $F \geq F_{r-1,n-r}(\alpha)$ (tabelováno)
- příklad: $F_{2,96}(0{,}05) = 3{,}09$, na 5% hladině jsme **prokázali**, že výška otců souvisí se vzděláním matek (ale hladina testu)

analýza rozptylu – předpoklady

- r **nezávislých** výběrů (nelze obejít)
- stejné rozptyly ve všech skupinách (lze testovat)
- normální rozdělení (lze ověřit stejně jako u regrese, pomocí reziduí)
- někdy pomůže transformace, např. logaritmy hodnocené veličiny
- Kruskalův-Wallisův test – zobecnění dvouvýběrového Wilcoxonova testu: hodnotí se pořadí místo původních hodnot, zda jsou průměrná pořadí podobná

příklad: věk matky ∼ vzdělání matky

vzdělání	rozsah n_i	průměr $\bar{X}_{i\bullet}$	prům. pořadí	součet pořadí T_i
základní	34	23,4	30,1	1025
střední	47	26,3	55,7	2618
VŠ	18	28,5	72,6	1307

- rozhoduje se pomocí upravené F-statistiky z analýzy rozptylu:

$$Q = \frac{12}{n(n+1)}\sum_{i=1}^{r}\frac{T_i^2}{n_i} - 3(n+1)$$

- kde T_i je součet pořadí i-té skupiny
- kritický obor: $Q \geq \chi^2_{r-1}(\alpha)$ (dostatečně velká n_i)
- $Q = 29{,}48 > 5{,}99, \qquad p < 0{,}0001$