Matematická statistika

2012–2013

Šárka Hudecová

Katedra pravděpodobnosti a matematické statistiky
Matematicko-fyzikální fakulta UK

hudecova@karlin.mff.cuni.cz
http://www.karlin.mff.cuni.cz/~hudecova

Organizační pokyny k přednášce

- přednáškové slidy na adrese
  `http://www.karlin.mff.cuni.cz/~hudecova`
  k dispozici před přednáškou, může docházet k úpravám
- studijní literatura
- konzultace
- zkouška — písemná (důraz na pochopení látky, aplikace na reálné příklady)
- cvičení — nepovinné

Přehled témat

- úvod
  (co je to statistika, motivační příklady z chemie)
- popisná statistika
  (popis výsledku experimentálního měření)
- základ pravděpodobnosti (pravděpodobnost, náhodné veličiny, jejich charakteristiky, nezávislost)
- principy statistické indukce
- principy testování hypotéz
- vybrané statistické testy

Co je statistika?

Statistika = věda o získávání, zpracování a interpretaci informace obsažené v empirických pozorováních skutečného světa
(v naměřených datech, průzkumech apod.)

Statistika = věda o zkoumání reality na základě napozorovaných dat

Cíl přednášky= porozumět základním principům statistických metod a pochopit řešení vybraných jednoduchých problémů.

(Důležité je osvojení si hlavních principů, pojmů, základních metod. Nikoliv učení se vzorečků.)

Základní dělení statistiky

- popisná (deskriptivní)
  - popis konkrétních dat
  - několika čísly a obrázky stručně vystihnout důležité
  - závěry pouze o daných datech, nelze zobecňovat
- induktivní (konfirmatorní)
  - na základě dat umožňuje odpovídat na obecné otázky o populaci ⇝ závěry lze zobecnit
  - odhady populačních parametrů
  - předpoklady, znalost statistických metod
  - důležitá je interpretace

Kde, kdy a proč se používá statistika?

Zkoumáme složitý systém

- nelze jednoduše pochopit nebo popsat pouze na základě teorie (tj. potřebujeme empirické zkušenosti)
- za stejných nebo podobných podmínek se může projevovat odlišným způsobem ⇝ náhoda
- příklady: vědecký experiment (měření), lidská společnost, ekonomika, lidské tělo, ekosystém, sport, ...
- chceme odhalit souvislosti, zákonitosti, systematické chyby atd.

Oblasti aplikace statistiky

- Přírodní vědy
  - biologie, chemie, fyzika, meteorologie, klimatologie, environmentální vědy
  - medicína, genetika, farmakologie
- Ekonomie
  - makro & mikroekonomie, bankovnictví, pojišťovnictví, ...
- Technické vědy
  - telekomunikace, doprava, počítače, strojírenství, kontrola jakosti, řízení a organizace výroby, ...
- Společenské vědy
  - sociologie, behaviorální vědy, archeologie, lingvistika, antropologie ...
- A mnoho dalších (sport, marketing, průzkum veřejného mínění ...)

Druhy statistických úloh (úlohy statistické indukce)

- odhady parametrů ⇝ výpočet číselných charakteristik
- testování hypotéz ⇝ ověřování pravdivosti výroků
- predikce ⇝ předpovědi
- optimalizace ⇝ hledání optimálních parametrů

Příklad

Na základě údajů z předchozích let lze usuzovat

- že by tu dnes mělo být 60 % žen a 40% mužů
- přítomné studentky budou v průměru 168 cm vysoké, s hmotností 60 kg a velikostí bot asi 38,5
- přítomní studenti budou v průměru 183 cm vysocí s hmotností 76 kg a velikostí bot asi 43
- přes 30 % přítomných bude z Prahy, kolem 11 % ze středočeského kraje a jen velmi málo studentů bude ze Slovenska a Moravy (statisticky významně méně než např. na MFF)

Optimalizace: změna posluchárny z M1 na M2

Statistika v chemických oborech

Experiment

- důležitý nástroj výzkumu
- složité fyzikálně-chemické modely — experimentální zjištění, ověření
- prakticky veškerý moderní výzkum — statistické zpracování výsledků

Chyby měření

- náhodné chyby
  omezená přesnost měřících přístrojů, proměnlivost podmínek,. . . kolísají náhodně kolem skutečné hodnoty
- systematické chyby

Statistické úlohy

- plánování experimentů
- detekce systematických chyb
- kalibrační přímka
- analytická chemie
- optimalizace
- průmyslová výroba: kontrola kvality, atd.
- mnohorozměrná data (obor chemometrie)
- další: porovnání různých laboratoří, přístrojů, podmínek atd.

Příklady

- Kontrola čistoty (kvality) chemikálie
- Porovnání dvou (nebo více) metod měření
  - koncentrace oxidu fosforečného v hnojivu — využití citronanu nebo využití kyseliny sírové
  - stanovení obsahu dinitrokresolu v postřikovacím přípravku — polarografická metoda (pracná) nebo titrační stanovení (levnější, rychlejší)
  - stanovení zlata v klenotnických slitinách
- Porovnání výtěžku z chemické reakce za různých podmínek
- Porovnání čistoty vody na různých místech řeky
- Vliv různých hnojiv na růst rostlin
- . . .

Statistika v reálném životě

- volební průzkumy, průzkumy veřejného mínění
  volba prezidenta: určení platných podpisů
- zprávy v médiích („američtí vědci prokázali ...“, globální oteplování, procenta)
- statistika v medicíně (klinické studie, prevence, prenatální diagnostika, kouření, ...)
- ...

Reálný život studenta PřF UK

- odborné články (pojmy: p-hodnota, statistická významnost, interval spolehlivosti atd.)
- pravděpodobnostní modely ve fyzice (kinetická teorie plynů apod.)

Popisná statistika

- experimentální měření $\leftrightsquigarrow$ data
- chceme popsat výsledek měření stručně a výstižně
  - číselné charakteristiky, obrázky
  - závislost mezi měřenými veličinami
- deskriptivní charakter (popisuje pouze daný vzorek)
- za dodatečných předpokladů slouží jako odhady a lze je zobecnit (statistická indukce)
- popis konkrétního datového souboru je nedílnou součástí každé analýzy

Data

- výsledek pozorování (měření)
- pozorování provádíme na nezávislých *subjektech*
  - chemické vzorky, osoby, státy, pacienti, rostliny, opakování měření ...
- měříme (zjišťujeme) hodnoty *znaků* (veličin, vlastností)
  - koncentrace určité látky, hmotnost, teplota, zabarvení ...
- na jednom subjektu můžeme měřit více znaků
- datová tabulka (např. Excel): pozorování na jednotlivých subjektech jsou většinou v řádcích, jednotlivé měřené veličiny ve sloupcích
- statistická analýza pomocí specializovaných statistických softwarů (např. program R, Statistica, SPSS, SAS atd.)

Příklad datového souboru

id	pohl	vyska	vaha	n.sour	v.ot	v.mat	bydliste
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
23	1	183	70	3	49	50	Vysočina
24	1	192	85	2	51	53	Jižní Morava
25	1	178	90	1	45	41	Karlovy Vary
26	0	168	55	1	53	53	Praha
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Měřítka, na kterých měříme znaky

- nominální
  - hodnoty jsou pouze označení různých kategorií
  - pohlaví, politický názor, barva, odrůda, . . .
- ordinální
  - uspořádané nominální hodnoty
  - vzdělání, spokojenost v práci (stupnice 1 až 5), stupeň bolesti, . . .
- intervalové
  - lze uvažovat jejich rozdíly, ale nelze se ptát „kolikrát“
  - např. rok narození, teplota ve stupních Celsia, . . .
- poměrové
  - většina veličin, které měříme
  - hmotnost, koncentrace, velikost, čas, suma v Kč . . .

Jiné dělení měřítek

- kvalitativní ⇝ kategoriální ⇝ faktory
  - jen několik možných hodnot (kategorií)
  - zajímají nás četnosti jednotlivých kategorií
  - uvažovat charakteristiky jako průměr nemá smysl
- kvantitativní ⇝ spojité
  - hodnoty jsou čísla
  - zajímají nás charakteristiky polohy (průměr), variability atd.

➜ odlišné metody pro popis kvalitativních a kvantitativních veličin

Zařazení daného znaku nemusí být jednoznačné (např. počet sourozenců)

Kvalitativní veličiny

Příklad

Politický názor před 2. kolem prezidentských voleb ⇝ průzkum u 11 náhodně vybraných osob:
S, S, Z, N, S, Z, Z, N, S, Z, Z

Vhodné popisné charakteristiky

- tabulka četností jednotlivých kategorií
- tabulka relativních četností jednotlivých kategorií
- modus = nejčastější hodnota

Tabulka četností

S	Z	N	celkem
4	5	2	11

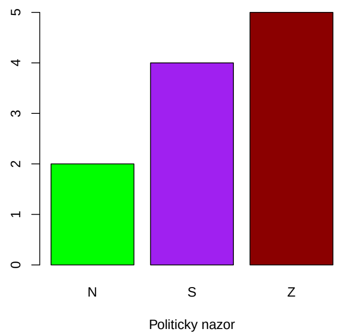
Tabulka relativních četností

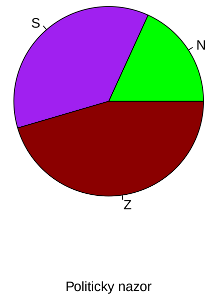
S	Z	N	celkem
0.364	0.455	0.181	1

Kvalitativní veličiny

Vhodné grafické znázornění

- sloupcový graf (obdelníkový diagram, barplot)
- koláčový graf (výsečová diagram, pieplot)





Kvalitativní veličiny

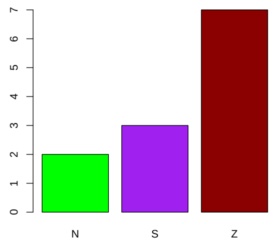
Stejný průzkum na jiném místě ČR: Z,Z,N,Z,S,Z,S,N,Z,Z,S,Z

Tabulka četností

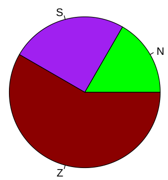
S	Z	N	celkem
3	7	2	12

Tabulka relativních četností

S	Z	N	celkem
0.250	0.583	0.167	1



Politicky nazor jinde v CR



Politicky nazor jinde v CR

Kvantitativní veličiny

Příklad

Experimentální měření koncentrace alkoholu ve 30 různých vzorcích vína:

13.20, 13.16, 14.37, 13.24, 14.20, 14.39, 14.06, 14.83, 13.86, 14.10, 14.12, 13.75, 14.75, 14.38, 13.63, 14.30, 13.83, 14.19, 13.64, 14.06, 12.93, 13.71, 12.85, 13.50, 13.05, 13.39, 13.30, 13.87, 14.02, 13.73

Chceme výstižně popsat výsledek měření

- míry polohy
  - charakteristika úrovně $\rightsquigarrow$ jakých hodnot veličina nabývá?
- míry variability
  - jak velmi se liší hodnoty veličiny u jednotlivých vzorků?
- grafické znázornění

Míry polohy — průměr

Pozorujeme hodnoty $x_1, \ldots, x_n$

- průměr

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- minimum, maximum

V některé aplikacích (ne velmi časté):

- vážený průměr: nezáporné váhy w_i

$$\overline{x}_W = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- příklad: vážený průměr známek (váhy = kredity)

Variančí řada

- původní hodnoty $x_1, \ldots, x_n$
- varianční řada

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

neklesající posloupnost vytvořená z naměřených hodnot
- $x_{(1)}$ je minimum, $x_{(n)}$ je maximum
- důležitý rozdíl mezi x_i a $x_{(i)}$

Příklad:

Naměřená data: 5,3,2,7,10

Varianční řada: 2,3,5,7,10

Míry polohy — medián

(Výběrový) medián $\tilde{x}$

- dělí data na dvě poloviny: polovina je menší (nebo rovna) než $\tilde{x}$ a polovina větší (nebo rovna) než $\tilde{x}$
- prostřední hodnota
- výpočet

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

Příklad:

- 5,3,2,7,10 $\rightsquigarrow \tilde{x} = 5$
- 5,3,2,7,10,1 $\rightsquigarrow \tilde{x} = 4$

Průměr vs. medián

- ČSÚ: medián platů v ČR, nikoliv průměrný plat
- Příklad: plat 5 osob (v tis. Kč)

$$18, 23, 35, 28, 21,$$

pak

$$\text{průměr } \bar{x} = 25, \quad \text{medián } \tilde{x} = 23$$

- Navíc jedna úspěšná osoba:

$$18, 23, 35, 28, 21, 160,$$

pak

$$\text{průměr } \bar{x} = 47.5, \quad \text{medián } \tilde{x} = 25.5$$

Míry polohy — kvantily

(Výběrové) kvantily (percentily):

- $\alpha\cdot$ 100% kvantil je hodnota taková, že $\alpha\cdot$ 100% hodnot v datech je menší nebo rovno a zbytek je větší nebo rovno
- např. 50 % kvantil je medián (polovina pod a polovina nad)
- dolní kvartil Q_1 = 25% kvantil
  čtvrtina hodnot je menších (nebo rovných) a tři čtvrtiny jsou větší (nebo stejné)
- horní kvartil Q_3 = 75% kvantil
  tři čtvrtiny hodnot jsou menší (nebo rovné) a čtvrtina je větší (nebo stejná)

Míry polohy — kvantily

Příklady využití:

- na VŠ budou brát pouze 10 % nejlepších studentů $\rightsquigarrow$ kolik musíte dosáhnout bodů v testu, abyste byli přijati?
- jaký obsah vápníku v krevním séru se považuje za nízký (výskyt max u 5 % u zdravých lidí)?
- růstové křivky u dětí — není dítě extrémně malé nebo extrémně velké?
- jak silné srážky lze očekávat v 1% extrémních případů?

Výpočet kvantilů

- pouze pro zajímavost
- více možných definic (např. v R devět různých metod výpočtu)

Hledáme $\alpha\cdot$ 100% kvantil $q(\alpha)$

- označíme
$$n_\alpha = 1 + (n-1)\alpha, \quad k = \lfloor n_\alpha \rfloor$$
(k je dolní celá část z n_α)
- $\alpha\cdot$ 100% kvantil leží mezi $x_{(k)}$ a $x_{(k+1)}$, spočítáme jej lineární interpolací
$$q = n_\alpha - \lfloor n_\alpha \rfloor,$$
$$q(\alpha) = (1-q)x_{(k)} + qx_{(k+1)}$$
- příklad: 30 pozorování, chceme 10% kvantil
  - logicky bychom chtěli vzít $1 + (30-1)\cdot 0.1 = 3.9$-tý člen varianční řady
  - vezmeme vážený průměr ze třetího a čtvrtého s vahami 0.1 a 0.9

Příklad víno

- průměr
$$\bar{x} = 13.814$$
- varianční řada
12.85, 12.93, 13.05, 13.16, 13.20, 13.24, 13.30, 13.39, 13.50, 13.63, 13.64, 13.71, 13.73, 13.75, 13.83, 13.86, 13.87, 14.02, 14.06, 14.06, 14.10, 14.12, 14.19, 14.20, 14.30, 14.37, 14.38, 14.39, 14.75, 14.83
- minimum 12.85, maximum 14.83
- medián
$$\tilde{x} = 13.845$$
- kvartily
$$Q_1 = 13.47, \quad Q_2 = 14.14$$
- 5% kvantil je 12.99
- 95% kvantil 14.55

Příklad hmotnost studentů v minulých letech

Data z let 2006-2011:

- 269 pozorování, 2 studenti hmotnost neuvedli
- průměrná hmotnost 66.2 kg, medián 64 kg, minimum 43 kg, maximum 113 kg
- 5% kvantil 50 kg, 95% kvantil 90 kg

Studenti:

- 109 hodnot a 1 chybějící
- průměrná hmotnost 76 kg, medián 75 kg, minimum 56 kg, maximum 113 kg
- 5% kvantil 60 kg, 95% kvantil 94 kg

Studentky:

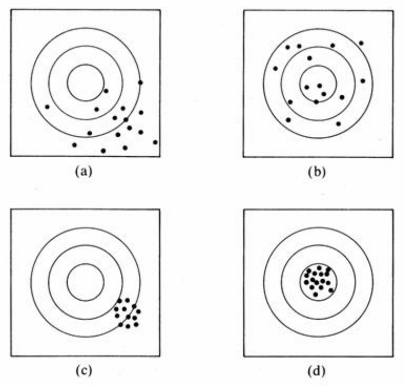
- 158 hodnot a 1 chybějící
- průměrná hmotnost 59.5 kg, medián 59 kg, minimum 43 kg, maximum 85 kg

Vlastnosti charakteristik polohy

- míry polohy charakterizují úroveň měřené spojité veličiny
- přičteme-li ke všem hodnotám stejnou konstantu a (posunutí) $\rightarrow$ změní se stejně i charakteristika polohy
- vynásobíme-li všechny hodnoty konstantou $b > 0$ $\rightarrow$ charakteristika polohy se zvýší b-krát
- je-li $m(x)$ míra polohy, pak
$$m(a+x) = a + m(x), \quad m(b\cdot x) = b\cdot m(x)$$
pro $a \in \mathbb{R}, b > 0$.

Míry variability

- měří rozptýlení (variabilitu, nestejnost)



Míry variability

(Výběrový) rozptyl

- průměrný čtverec vzdálenosti od průměru

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

- v jednotkách^2

(Výběrová) směrodatná odchylka

- odmocnina z rozptylu

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- stejný fyzikální rozměr jako původní data

Další míry variability

- rozpětí $x_{(n)} - x(1)$
- mezikvartilové rozpětí $R = Q_3 - Q_1$

Vlastnosti charakteristik variability

- posunutím se míra variability nezmění (nezávisí na poloze)

$$s(a + x) = s(x)$$

- reaguje na vynásobení kladnou konstantou

$$s(b \cdot x) = b \cdot s(x), \quad b > 0.$$

Příklad víno

- rozptyl

$$\sum_{i=1}^{30} x_i^2 = 5732.319, \quad \bar{x}^2 = 190.817$$

a tedy

$$s^2 = \frac{1}{29}(5732.319 - 30 \cdot 190.817) = 0.269$$

- směrodatná odchylka

$$s = \sqrt{0.269} = 0.519$$

- rozpětí

$$x_{(30)} - x_{(1)} = 14.83 - 12.85 = 1.98$$

- mezikvartilové rozpětí

$$Q_3 - Q_1 = 14.14 - 13.47 = 0.67$$

Příklad hmotnost studentů

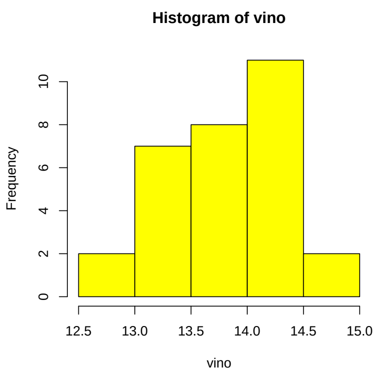
Charakteristika	Studenti	Studentky
rozptyl [kg^2]	127.51	54.57
směrodatná odchylka [kg]	11.29	7.39
rozpětí [kg]	57	42
mezikvart. rozpětí [kg]	14	10

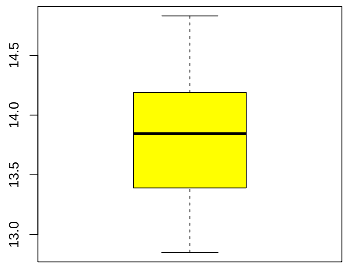
Poznámky

- existuje řada dalších popisných charakteristik (šikmost, špičatost, specializované popisné statistiky ...)
- ve statistické indukci slouží popisné statistiky jako odhady neznámých parametrů $\leftrightsquigarrow$ uvidíme později (je zavést předpoklady, zvážit reprezentativnost atd.)

Grafické nástroje popisné statistiky

- histogram
- krabicový diagram (boxplot)





Histogram

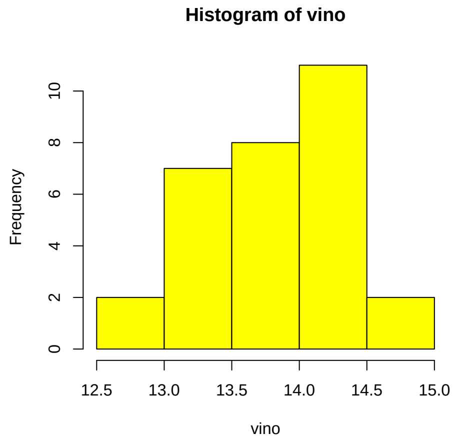
- dává nahlédnout, jak jsou jednotlivé hodnoty znaku v našich datech rozloženy (které hodnoty se objevují často a které ojediněle)
- interval $I = [a, b]$ pokrývá celé rozmezí dat
- rozdělíme jej na K navazujících stejně velkých podintervalů A_k, $k = 1, \dots, K$, všechny délky $h = \frac{b-a}{K}$ (bereme např. zprava uzavřené s výjimkou prvního)
- n_k počet pozorování, které padly do A_k
- histogram = grafické znázornění intervalových četností n_k: každému A_k odpovídá obdelník, jehož výška je rovna n_k

Příklad víno

12.85, 12.93, 13.05, 13.16, 13.20, 13.24, 13.30, 13.39, 13.50, 13.63, 13.64, 13.71, 13.73, 13.75, 13.83, 13.86, 13.87, 14.02, 14.06, 14.06, 14.10, 14.12, 14.19, 14.20, 14.30, 14.37, 14.38, 14.39, 14.75, 14.83

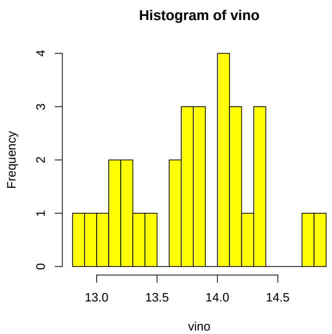
Zvolíme $a = 12.5,\ b = 15,\ K = 5\ \to h = 0.5$

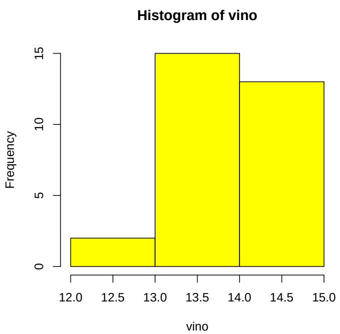
k	interval A_k	četnost n_k
1	$[12.5, 13]$	2
2	$(13, 13.5]$	7
3	$(13.5, 14]$	8
4	$(14, 14.5]$	11
5	$(14.5, 15]$	2



Histogram

Histogram se může lišit podle volby K



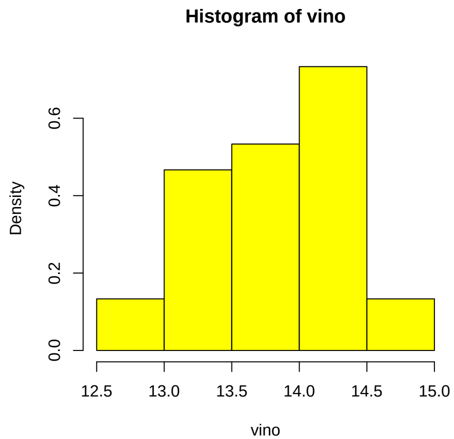


Sturgesovo pravidlo:

$$K \approx 1 + \log_2 n$$

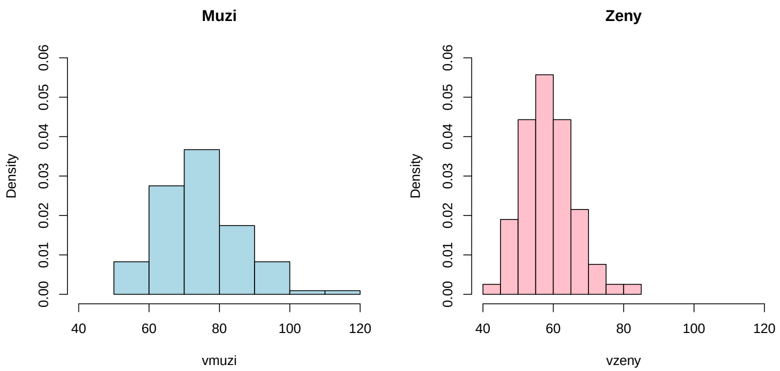
Histogram

Normovaná verze histogramu (plocha =1)



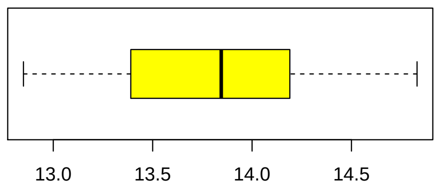
Histogram

Hmotnost studentů



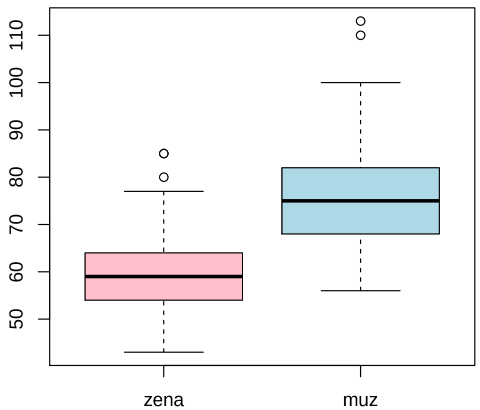
Krabicový diagram

- nemá úplně závaznou definici (může se lišit v různých programech
- obvykle zakreslen výběrový medián a kvartily
  - krabice: horní a dolní okraj určují výběrové kvartily Q_1 a Q_3
  - uprostřed čára určující výběrový medián
  - „vousy“ ukazují rozmezí dat $\leftrightsquigarrow$ od kvartilu k minimu/maximu (není-li odlehlé)
  - odlehlé pozorování $\leftrightsquigarrow$ je dál než $3/2 \cdot (Q_3 - Q_1)$ od bližšího kvartilu



Krabicový diagram

Hmotnost studentů



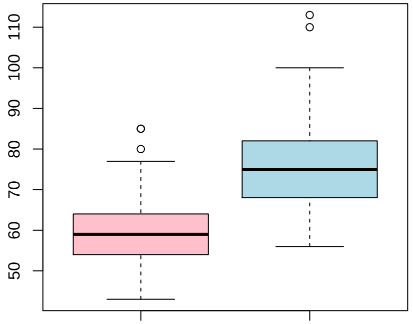
Popis závislosti dvou veličin

- jednou ze základních otázek je vyšetřování závislosti (vztahu) dvou veličin
- na každém subjektu měříme dva znaky
- statistická indukce: testování nezávislosti, modelování závislosti atd.
- první krok = popisná statistika
- metody závisí na měřítkách znaků

Vztah kategoriální a spojité veličiny

Příklad: vztah hmotnosti a pohlaví

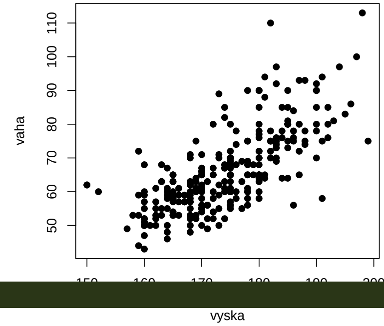
- číselný popis ve skupinách → porovnání
- odlišnosti svědčí pro závislost znaků



Vztah dvou spojitých veličin

Příklad: Vztah mezi výškou a hmotností

- bodový graf
- číselný popis – tzv. korelace (korelační koeficient) — bude později
- regresní přímka (kalibrace) — bude později (?)



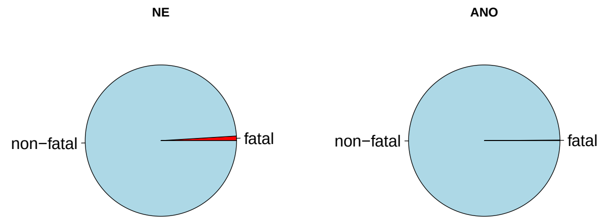
Vztah dvou kategoriálních veličin

Příklad: Používání bezpečnostních pásů a charakter zranění (výzkum z roku 1988 na Floridě)

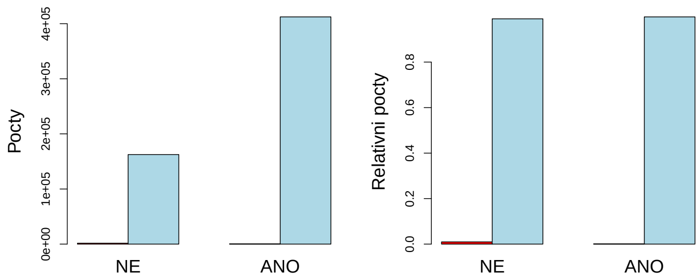
	Zranění		
Bezpečnostní pás	fatální	nefatální	celkem
ne	1 601	162 527	164 128
ano	510	412 368	412 878
celkem	2111	574 895	577 006

Relativní četnosti I

	Zranění		
Bezpečnostní pás	fatální	nefatální	celkem
ne	0.98 %	99.02 %	100 %
ano	0.12 %	99.88 %	100 %

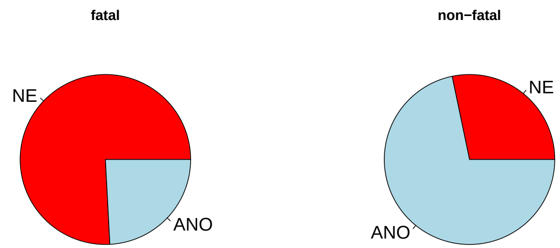


Relativní četnosti I



Relativní četnosti II

	Zranění	
Bezpečnostní pás	fatální	nefatální
ne	75.84 %	28.27 %
ano	24.16 %	71.73 %
celkem	100 %	100 %



Relativní četnosti II