PŘÍPRAVA BAKALÁŘSKÉ PRÁCE
4. METODIKA A VÝSLEDKY

Mgr. Markéta Pavlíková

www.biostatisticka.cz

VÝZKUM x BAKALÁŘKA

VÝZKUM

1. Pozorování
2. Vytvoření hypotéz
3. Design studie
4. Sběr dat
5. Analýza
6. Interpretace a srovnání
7. Publikace
8. Příprava nového experimentu

BAKALÁŘKA

1. Úvod a teoretická část
2. Výzkumná otázka + Hypotézy
3.+ 4. Metodika
4.+ 5. Výsledky
6. Diskuse
7. Závěr + obhajoba
8. Magisterské :), ale i praxe

VÝZKUM x BAKALÁŘKA

VÝZKUM

1. Pozorování
2. Vytvoření hypotéz
3. Design studie
4. Sběr dat
5. Analýza
6. Interpretace a srovnání
7. Publikace
8. Příprava nového experimentu

BAKALÁŘKA

1. Úvod a teoretická část
2. Výzkumná otázka + Hypotézy
3.+ 4. Metodika
4.+ 5. Výsledky
6. Diskuse
7. Závěr + obhajoba
8. Magisterské :), ale i praxe

VÝZKUM x BAKALÁŘKA

VÝZKUM

1. Pozorování
2. Vytvoření hypotéz
3. Design studie
4. Sběr dat
5. Analýza
6. Interpretace a srovnání
7. Publikace
8. Příprava nového experimentu

BAKALÁŘKA

1. Úvod a teoretická část
2. Výzkumná otázka + Hypotézy
3.+ 4. Metodika
4.+ 5. Výsledky
6. Diskuse
7. Závěr + obhajoba
8. Magisterské :), ale i praxe

VÝZKUM x BAKALÁŘKA

VÝZKUM

1. Pozorování
2. Vytvoření hypotéz
3. Design studie
4. Sběr dat
5. Analýza
6. Interpretace a srovnání
7. Publikace
8. Příprava nového experimentu

BAKALÁŘKA

1. Úvod a teoretická část
2. Výzkumná otázka + Hypotézy
3.+ 4. Metodika
4.+ 5. Výsledky
6. Diskuse
7. Závěr + obhajoba
8. Magisterské :), ale i praxe

VÝZKUM x BAKALÁŘKA

VÝZKUM

1. Pozorování
2. Vytvoření hypotéz
3. Design studie
4. Sběr dat
5. Analýza
6. Interpretace a srovnání
7. Publikace
8. Příprava nového experimentu

BAKALÁŘKA

1. Úvod a teoretická část
2. Výzkumná otázka + Hypotézy
3.+ 4. Metodika
4.+ 5. Výsledky
6. Diskuse
7. Závěr + obhajoba
8. Magisterské :), ale i praxe

ANALÝZA DAT + PREZENTACE VÝSLEDKŮ

- Analýza dat = to co technicky děláte, abyste
  - vyčistili data
  - popsali, jak data vypadají
  - otestovali hypotézy
  - prozkoumali vliv dalších proměnných
- Výsledky = ta část práce, kde popisujete a ukazujete
  - jak vypadají osoby v souboru, který jste nasbírali → popisná část
  - jak vypadají zájmové proměnné u osob v souboru → popisná část
  - výsledek testování hypotéz → analytická část
  - výsledky další doplňkových analýz → analytická část
  - používáme text, tabulky a grafy, logicky a informativně

METODIKA x VÝSLEDKY

- Metodika = popisuji jak jsem to dělal(a)
  - jak, kdy a kde jsem vybíral(a) soubor, proč takto, kdo byl plánovaně zahrnut a kdo vyloučen (obecná kriteria), povolení etické komise
  - jaké jsem sbíral(a) proměnné a jak jsem je měřila
  - popis všech nástrojů sběru:
    - přístroj (výrobce, nastavení přístroje)
    - dotazník (kdo ho vymyslel, validoval, vzor do přílohy)
    - laboratorní testy (kdo provedl, jaký kit …)
    - měření a testy (jaké škály, jak prováděno, za jakých podmínek)
  - popis statistických technik včetně užitého software, stanovená míra statistické významnosti (obvykle 5 % / 0.05)

METODIKA x VÝSLEDKY

- Výsledky, popisná část = popisuji kolik, koho a co jsem nasbíral(a)
  - kolik jsem získal(a) respondentů
  - kolik respondentů odmítlo, návratnost dotazníků, kolik studii nedokončilo, proč
  - popis respondentů: pohlaví, věk, relevantní anamnestické charakteristiky (BMI, typ / stupeň onemocnění, …)
  - používáme deskriptivní statistiky (viz dále)
  - ideálně stručný popis v textu (např průměr + SD) + přehledná tabulka
  - nedávat sem celá data! max ilustrační ukázku, pokud je to vhodné
  - pokud chceme ukázat celá data, pak do příloh!

METODIKA x VÝSLEDKY

- Výsledky, popisná část (pokračování)
  - popis nasbíraných charakteristik relevantních k výzkumu
  - používáme deskriptivní statistiky
  - ideálně stručný popis v textu + přehledná tabulka
  - možno doplnit informativními grafy, které například ukazují šikmé rozložení dat a naznačují použití nějaké transformace v analýze
  - někdy u části respondentů některé údaje chybí – je vhodné uvést do tabulky i skutečný počet odpovědí (používám pravidlo 5-10 %)
  - nezapomínejte na jednotky!

Cílem popisné části je dát čtenáři obrázek o datech, která máte k dispozici, aniž byste mu nutili hrubá data.

METODIKA x VÝSLEDKY

- Výsledky, analytická část
  - popis výsledků testů hypotéz
  - uvádíme testové statistiky:
    - průměry/mediány, odhad rozdílu, odds ratio (OR), ...
    - intervaly spolehlivosti
    - p-hodnoty
    - ...
  - popis v textu + přehledná tabulka nebo graf nebo oboje
  - nezapomínejte na jednotky!
  - zaokrouhlujte na rozumné desetiny

Cílem analytické části je ukázat, zda platí / neplatí hypotézy, které jste si stanovili, a jaké další faktory ovlivňující výsledek jste našli.

METODIKA x VÝSLEDKY

- Výsledky, analytická část
  - popis výsledků testů hypotéz
  - uvádíme testové statistiky:
    - průměry/mediány, odhad rozdílu, odds ratio (OR), ...
    - intervaly spolehlivosti
    - p-hodnoty
    - ...
  - popis v textu + přehledná tabulka nebo graf nebo oboje
  - nezapomínejte na jednotky!
  - zaokrouhlujte na rozumné desetiny

Cílem analytické části je ukázat, zda platí / neplatí hypotézy, které jste si stanovili, a jaké další faktory ovlivňující výsledek jste našli.

VÝSLEDKY - CHYBY

- × úplná absence statistického testování, jen „okometrický“ pohled, srovnávání hodnot bez otestování
- × použití nevhodného testu
- × použití průměru u nesymetrického rozdělení
- × chybějící parametr rozptylu (SD, IQR, rozsah)
- × uvádění celých dat v hlavní části (ukázka dat v příloze nebo elektronicky)
- × zvláštní tabulka na každou proměnnou
- × tabulky přeplněné k nesrozumitelnosti (✓raději rozdělit)
- × zbytečně mnoho desetinných míst (u % stačí jedno nebo i žádné)
- × grafy k proměnným, kde je informativnější tabulka nebo jen číslo v textu
- × nadužívání koláčových grafů

TABULKA - PŘÍKLAD

Table 1. Demographic, biochemical, and genetic characteristics of gout and hyperuricemia cohorts.

Characteristic		All patients (*N* = 250)		Gout patients (*N* = 182)		Hyperuricemia patients (*N* = 68)		*P*-value[#]
		N	%	*N*	%	*N*	%	
Gender	Male	214	85.6	166	91.2	48	70.6	0.0002
	Female	36	14.4	16	16.8	20	29.4	
Familial occurrence		97	38.8 (40.2*)	66	36.3 (36.5*)	31	45.6 (51.7*)	0.0480

Characteristic	*N*	Median (IQR)	Range	*N*	Median (IQR)	Range	*N*	Median (IQR)	Range	*P*-value[†]
Age at examination [years]	250	51.5 (25.0)	3–90	182	54.0 (21.0)	11–90	68	36.0 (42.0)	3–78	< 0.0001
BMI at examination	209	28.4(5.8)	16–50	151	28.4 (5.4)	19.5–50	58	28.1(6.4)	16–41	0.0822
Gout/hyperuricemia onset[$] [years]	236	40.0 (28.0)	1.2–84	181	40.0 (24.0)	8–84	55	27.0 (40.5)	1.2–76	0.0070
SUA at examination, with medication [µmol/L]	201	375.0 (134.0)	163–808	159	372.0 (128.0)	163–808	42	424.0 (140.0)	240–628	0.0515
FEUA at examination, with medication	194	3.4 (2.0)	0.9–14	158	3.4 (1.9)	0.9–14	36	3.8(2.1)	1.3–8	0.5862

[#] Fisher's exact test for categorical and [†] Wilcoxon two-sample sum rank test were used to compare the gout sub-group with the hyperuricemia sub-group; * relative frequencies when missing information about familial occurrence was excluded; [$] onset (gout) and age of ascertainment (hyperuricemia). IQR, interquartile range; SUA, serum uric acid; FEUA, fractional excretion of uric acid.

ANALÝZA DAT

CO SE SKRÝVÁ V DATECH

- data sbíráme proto, abychom porozuměli skutečnosti, principům, zákonitostem
- musíme umět data nějak popsat, zestručnit
  - rozsah hodnot
  - rozložení hodnot
  - jsou některé hodnoty významnější, častější než jiné?
- klademe si otázky po podstatě
  - jsou naměřené hodnoty stejné jako naměřené dříve / jinde?
  - existuje nějaké pravidlo, podle kterého měřené hodnoty vznikají?

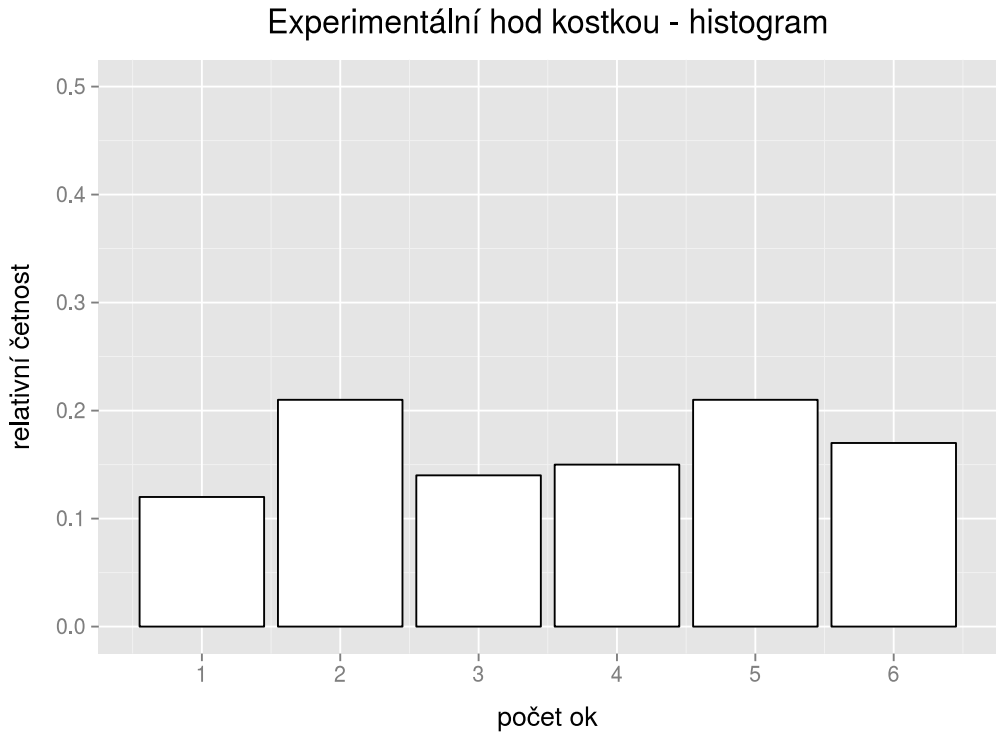
DESKRIPTIVNÍ

ANALYTICKÁ

CO SE SKRÝVÁ V DATECH

Házíme 100x kostkou

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



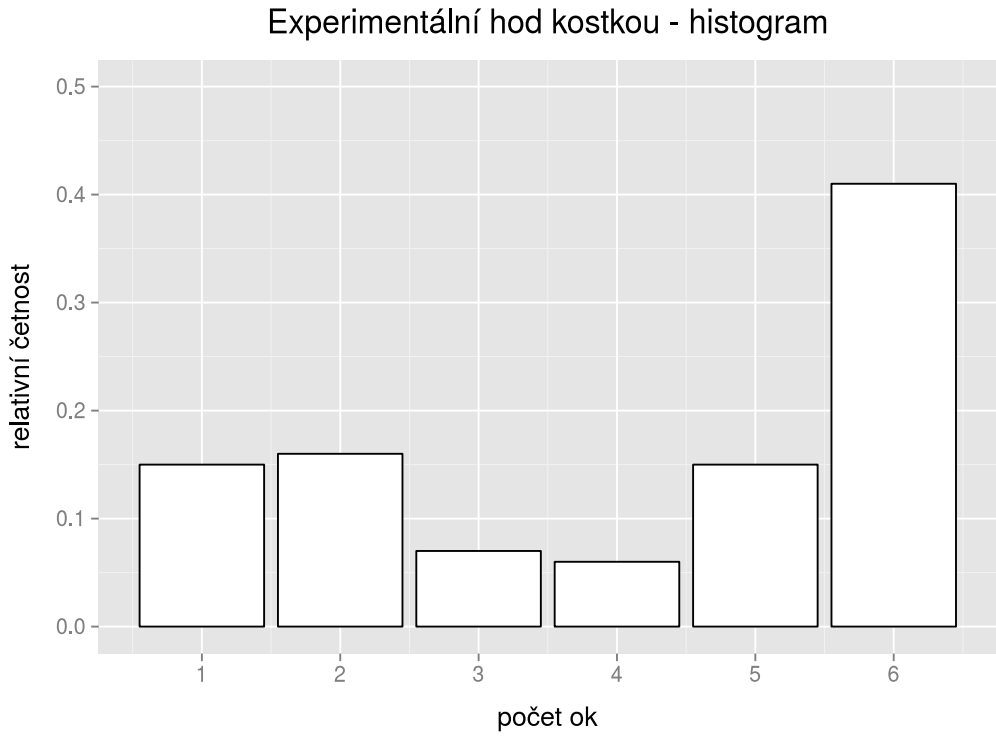
Popis: Absolutní četnost a relativní četnost, tabulka, histogram

Podstata: Existuje nějaké pravidlo za výsledky?

CO SE SKRÝVÁ V DATECH

Házíme 100x kostkou

j		n_j	$f_j = n_j/n$
1	𝍸	15	0,15
2	𝍸 |	16	0,16
3	𝍸 ||	7	0,07
4	𝍸 |	6	0,06
5	𝍸 𝍸	15	0,15
6	𝍸 𝍸 𝍸 𝍸 |	41	0,41
		$n = 100$	1,00

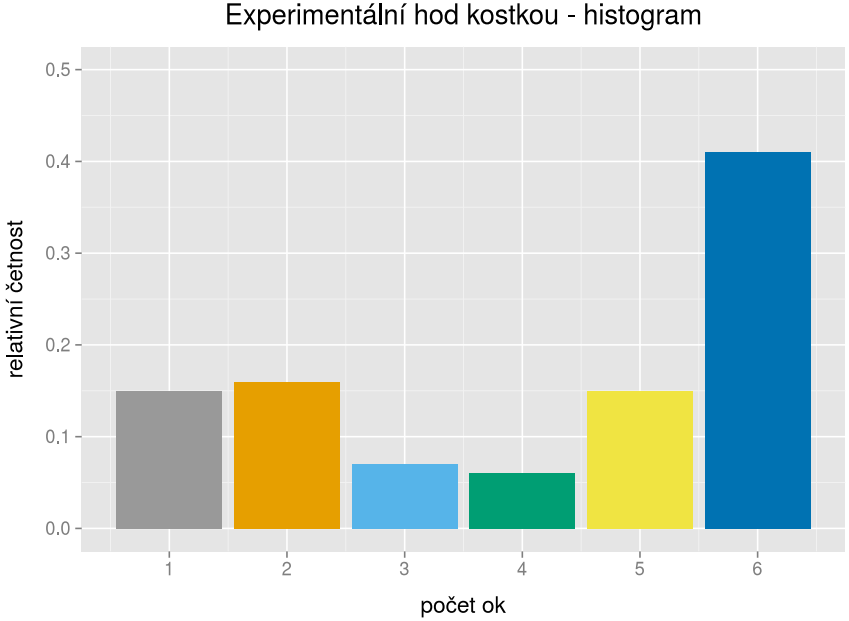


Popis: Absolutní četnost a relativní četnost, tabulka, histogram

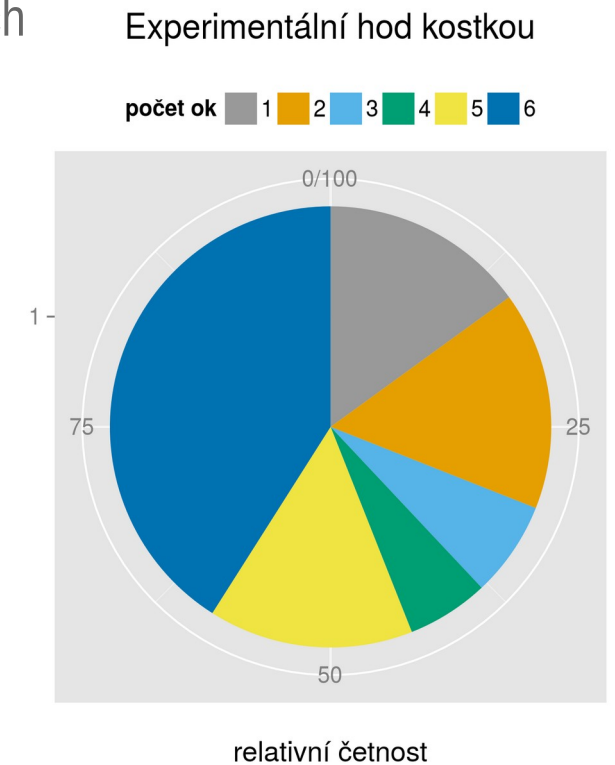
Podstata: Jsou kostky stejné? Liší se od předpokládaného principu rovnoměrného výskytu?

KATEGORIÁLNÍ PROMĚNNÁ

počet ok	absolutní četnost	relativní četnost
1	15	0.15
2	16	0.16
3	7	0.07
4	6	0.06
5	15	0.15
6	41	0.41
celkem	100	1.00



- tabulka ukazuje konkrétní data
- sloupcový graf / histogram vizualizuje četnosti
  - umožňuje okometrické srovnání
- koláčový graf
  - vypadá pěkně, ale nedává reálnou představu o konkrétních hodnotách
  - neumožní srovnání více faktorů
  - jen velmi výjimečně

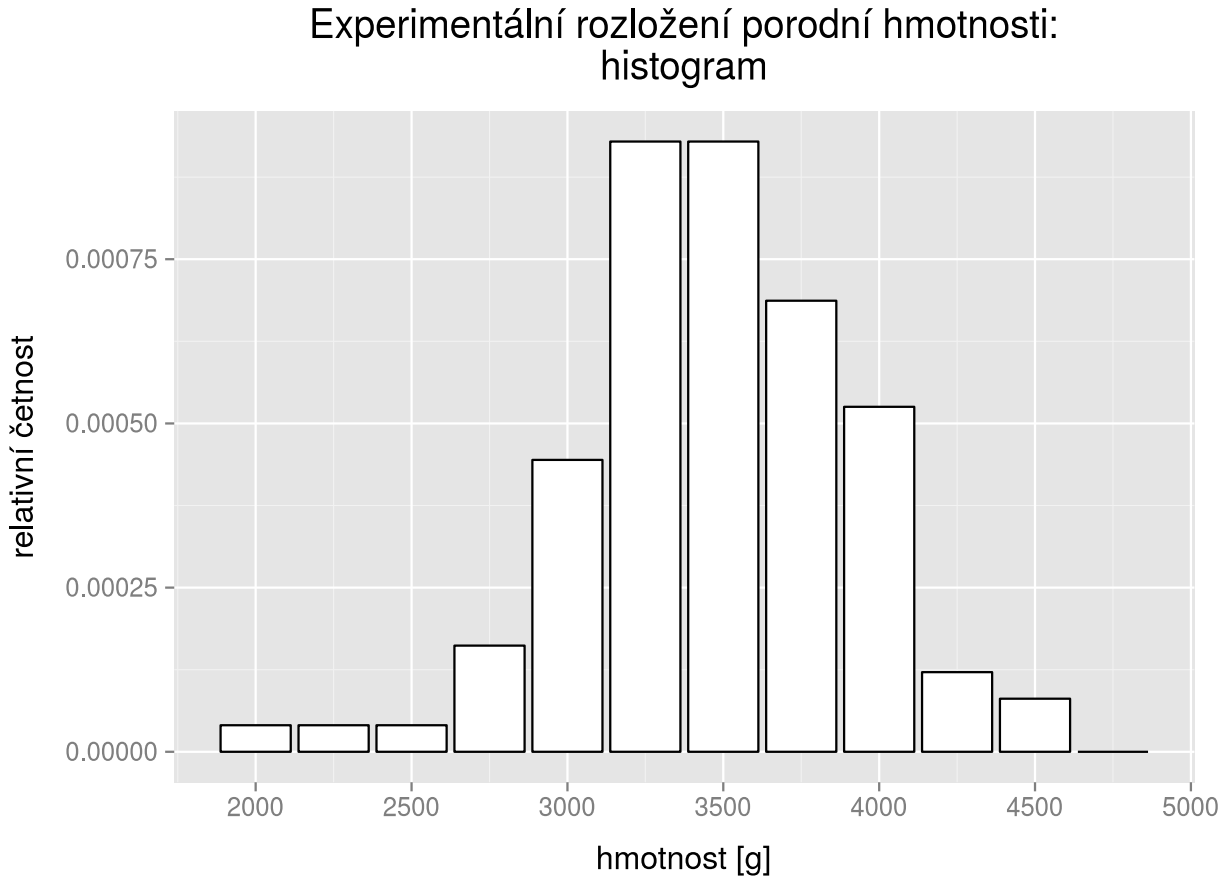


SPOJITÁ PROMĚNNÁ

Vzorek 99 novorozených dětí, porodní hmotnost

Jak data popsat?

- rozsah
- nejčastější hodnota
- „prostřední“ hodnota
- symetrie, „sešikmenost“
- „rozpláclost“
- ...



POPISNÉ STATISTIKY - POLOHA

- minimum, maximum
  - odkud kam jdou data
  - dávají hodnoty smysl? pokrývají „správný“ rozsah?
  - jsou tam nějaká extrémní měření? jsou správně?

- aritmerický průměr
  - součet všech hodnot / počet všech hodnot

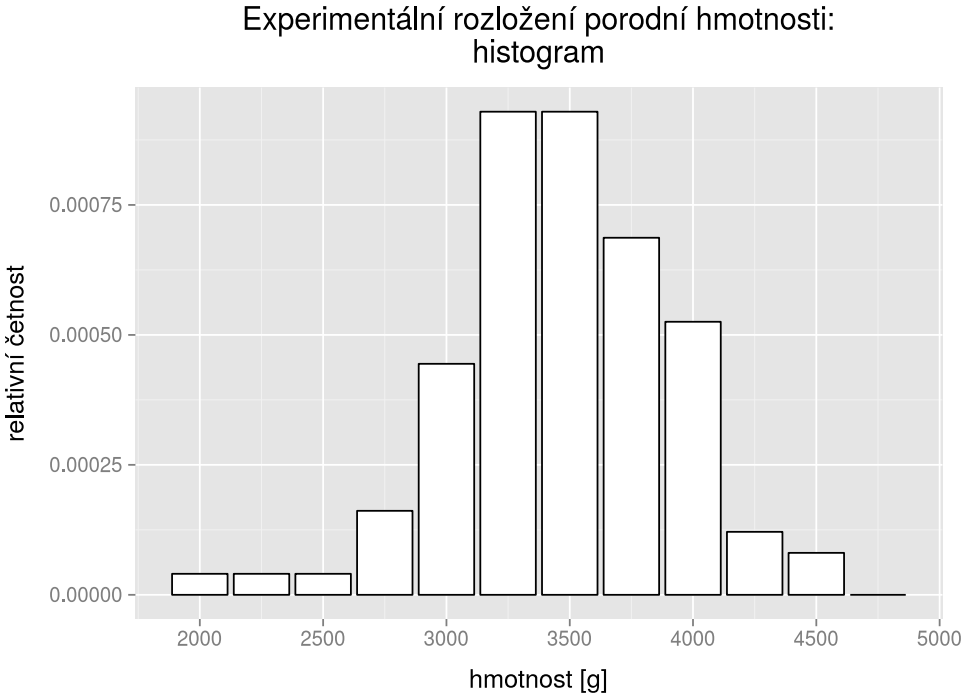
$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

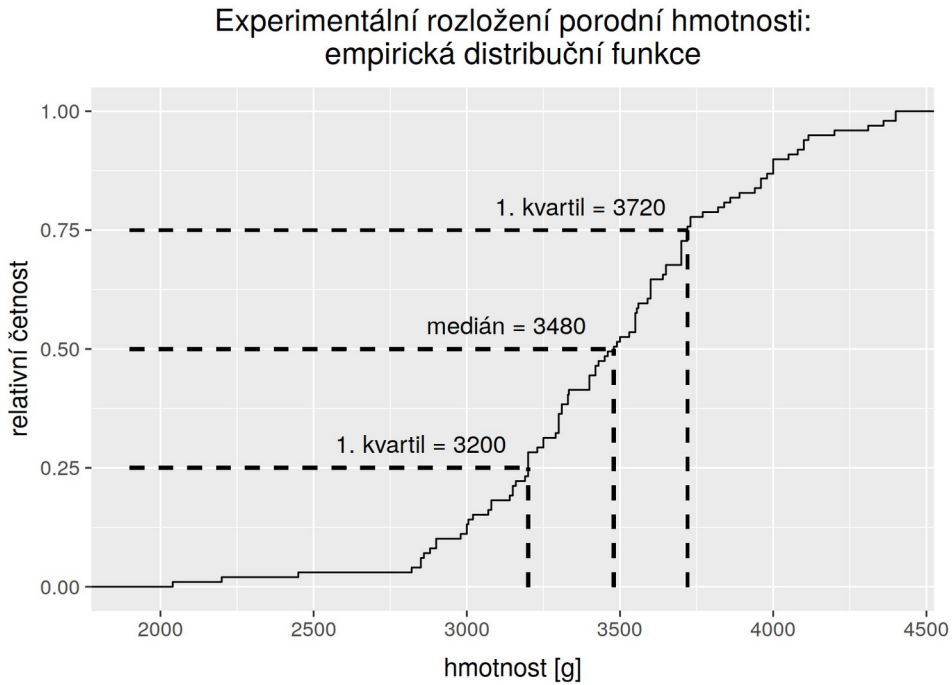
POPISNÉ STATISTIKY - POLOHA

- modus
  - nejčastější hodnota, může jich být i víc (více hrbů)
- medián
  - dělí soubor na polovinu
  - pod ním je polovina dat, nad ním je polovina dat
  - často výrazně spolehlivější informace než průměr
- kvartily
  - dělí soubor na čtvrtiny
  - pod 1. kvartilem leží čtvrtina dat
  - nad 3. kvartilem leží čtvrtina dat

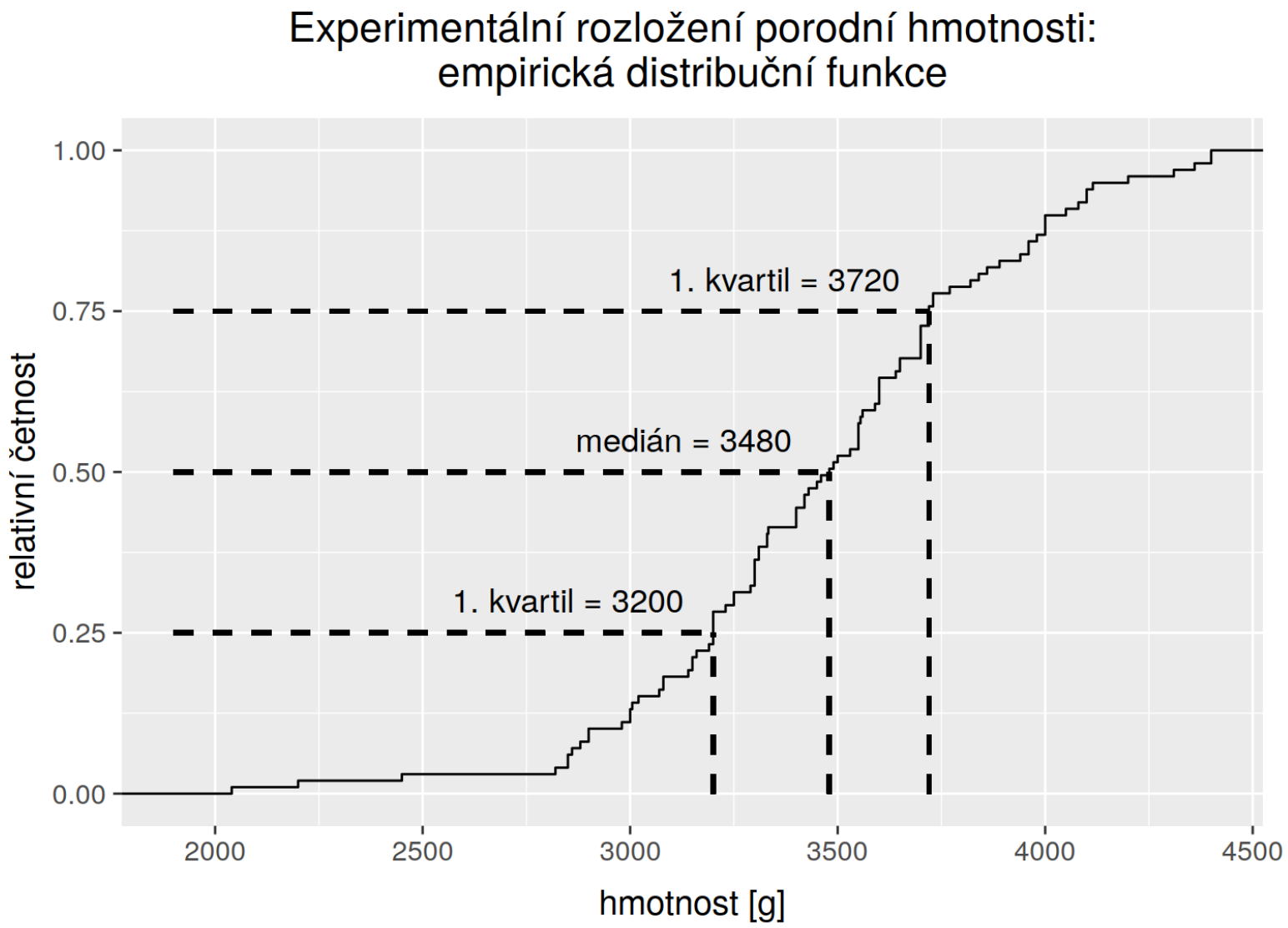
MEDIÁN, KVARTILY – JAK VYPOČÍST

- seřadíme data podle velikosti (novorozenci: 99 hodnot)
- najdeme polovinu (novorozenci: 50. hodnota)
- pokud sudý počet: vezmeme průměr z těch dvou uprostřed





MEDIÁN, KVARTILY – JAK VYPOČÍST

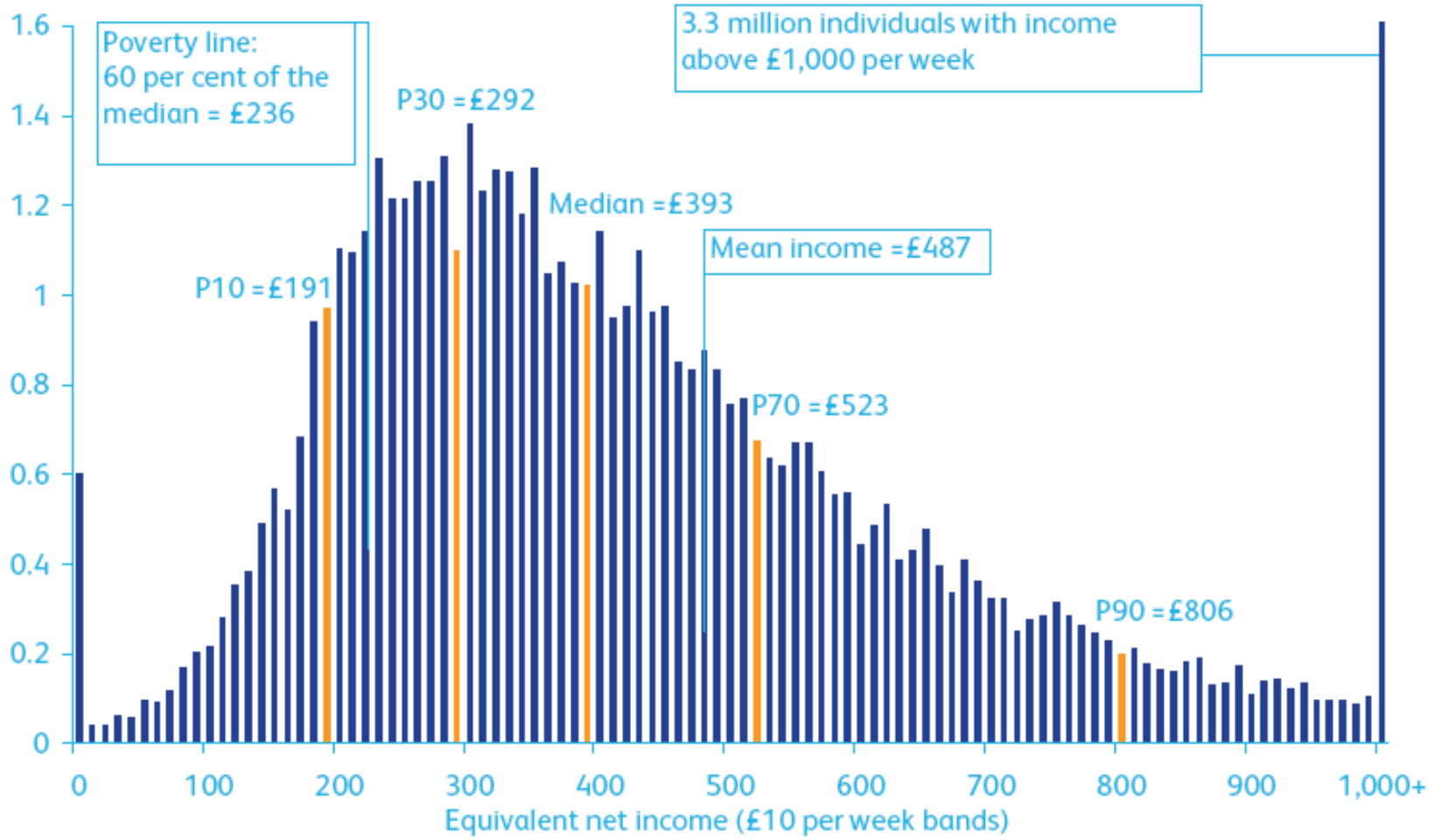


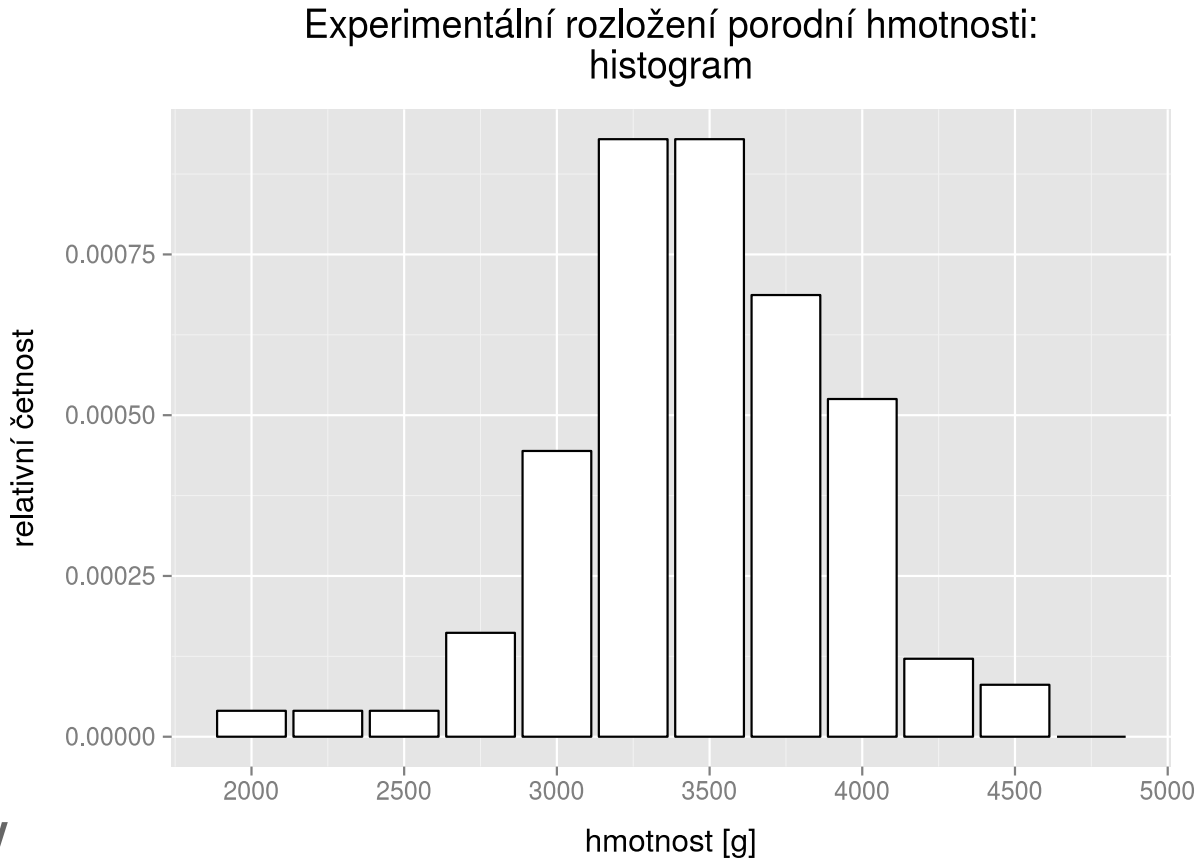
MEDIÁN vs. PRŮMĚR

- průměr
  - dobře určuje „prostředek“ u symetrických dat (novorozenci: 3470g)
  - nejlépe funguje, když je rozložení typu Gaussova křivka
  - nefunguje v případech typu A: 1 kuře B, C: žádné kuře, každý snědl třetinu kuřete
  - klasický příklad: 2/3 zaměstnanců mají podprůměrný plat
  - bohatí táhnou průměr svým směrem
- medián
  - u symetrických dat bude poblíž průměru (novorozenci: 3480g)
  - u nesymetrických dat bude blíže „těžší“ straně
  - ignoruje výši extrémů, bere v potaz jen jejich počet
  - plat pod mediánem má vždy 1/2 zaměstnanců
  - bohatých je málo, a tak hrají ve výpočtu menší roli
  - u kuřat: jaký bude medián?

MEDIÁN vs. PRŮMĚR

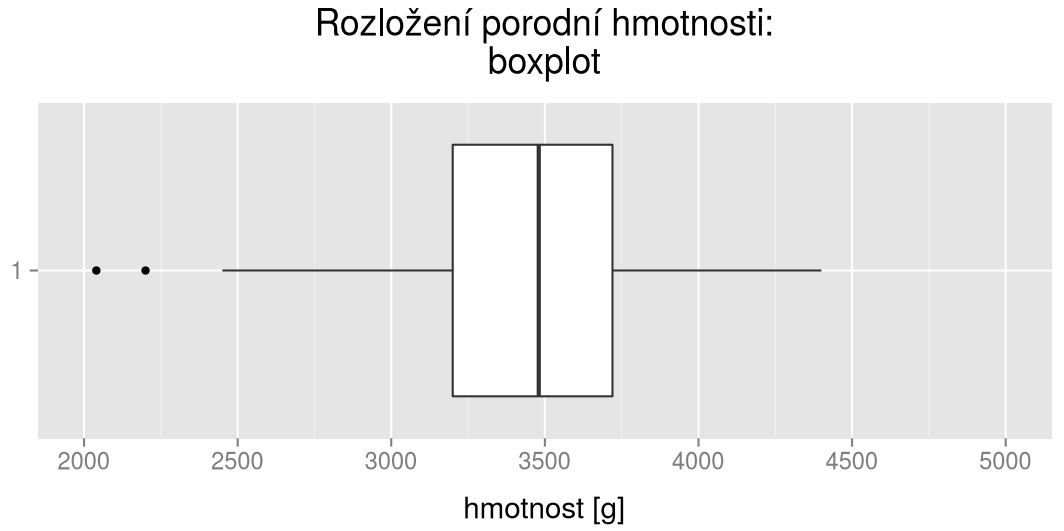
Half of the population has income below and half above £393 per week (adjusted for household size)





Boxplot

- shrnuje míry polohy
- uprostřed medián
- okraj boxu kvartily
- vousy 1.5 x IQR
- tečky jsou „outliers“ (odlehlá pozorování)



POPISNÉ STATISTIKY - VARIABILITA

- rozpětí (range)
  - maximum - minimum
  - základní informace o variabilitě dat
- interkvartilové rozpětí (interquartile range)
  - třetí - první kvartil
- výběrový rozptyl
  - čím dále od průměru tím větší váha
  - mocnina zdůrazňuje vzdálenější
- směrodatná odchylka (SD)
  - odmocnina rozptylu
  - má stejnou jednotku jako data

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

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

POPISNÉ STATISTIKY - SHRNUTÍ

- Spojitá proměnná
  - do tabulky
    - počet
    - parametry polohy: průměr, medián
    - parametry variability: minimum a maximum, SD, (IQR, 1. a 3. kvartil)
  - graf: histogram a boxplot
- Kategoriální proměnná
  - absolutní a relativní počty v jednotlivých kategoriích
  - graf: sloupcový graf jak pro počty v jednotlivých kategoriích, tak i pro relativní četnosti

TABULKA - PŘÍKLAD

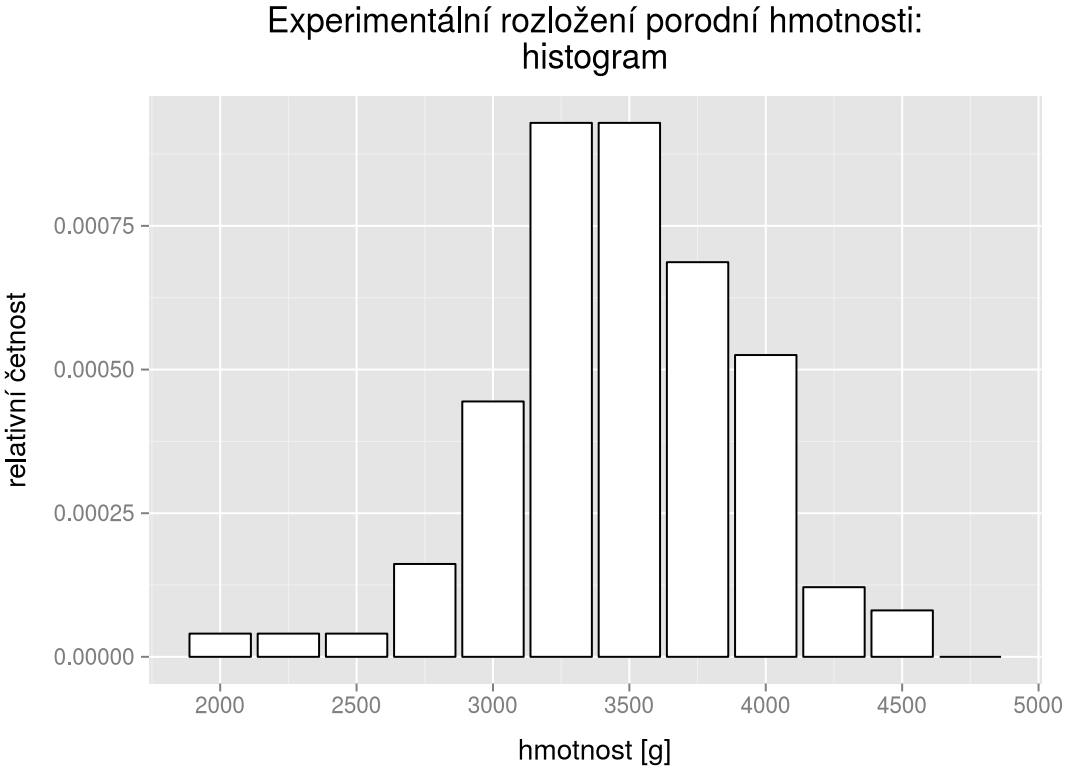
Table 1. Demographic, biochemical, and genetic characteristics of gout and hyperuricemia cohorts.

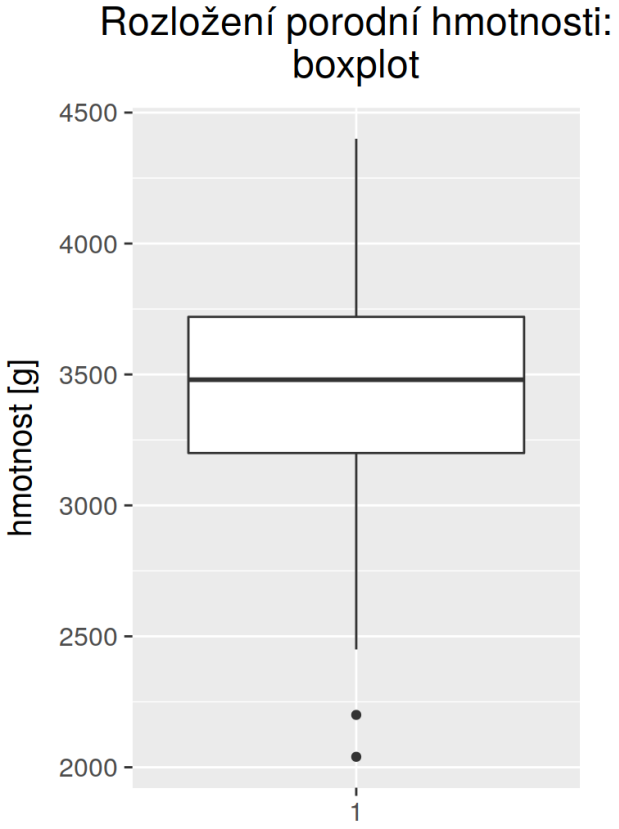
Characteristic		All patients (*N* = 250)		Gout patients (*N* = 182)		Hyperuricemia patients (*N* = 68)		*P*-value[#]
		N	%	*N*	%	*N*	%	
Gender	Male	214	85.6	166	91.2	48	70.6	0.0002
	Female	36	14.4	16	16.8	20	29.4	
Familial occurrence		97	38.8 (40.2*)	66	36.3 (36.5*)	31	45.6 (51.7*)	0.0480

Characteristic	*N*	Median (IQR)	Range	*N*	Median (IQR)	Range	*N*	Median (IQR)	Range	*P*-value[†]
Age at examination [years]	250	51.5 (25.0)	3–90	182	54.0 (21.0)	11–90	68	36.0 (42.0)	3–78	< 0.0001
BMI at examination	209	28.4(5.8)	16–50	151	28.4 (5.4)	19.5–50	58	28.1(6.4)	16–41	0.0822
Gout/hyperuricemia onset[$] [years]	236	40.0 (28.0)	1.2–84	181	40.0 (24.0)	8–84	55	27.0 (40.5)	1.2–76	0.0070
SUA at examination, with medication [µmol/L]	201	375.0 (134.0)	163–808	159	372.0 (128.0)	163–808	42	424.0 (140.0)	240–628	0.0515
FEUA at examination, with medication	194	3.4 (2.0)	0.9–14	158	3.4 (1.9)	0.9–14	36	3.8(2.1)	1.3–8	0.5862

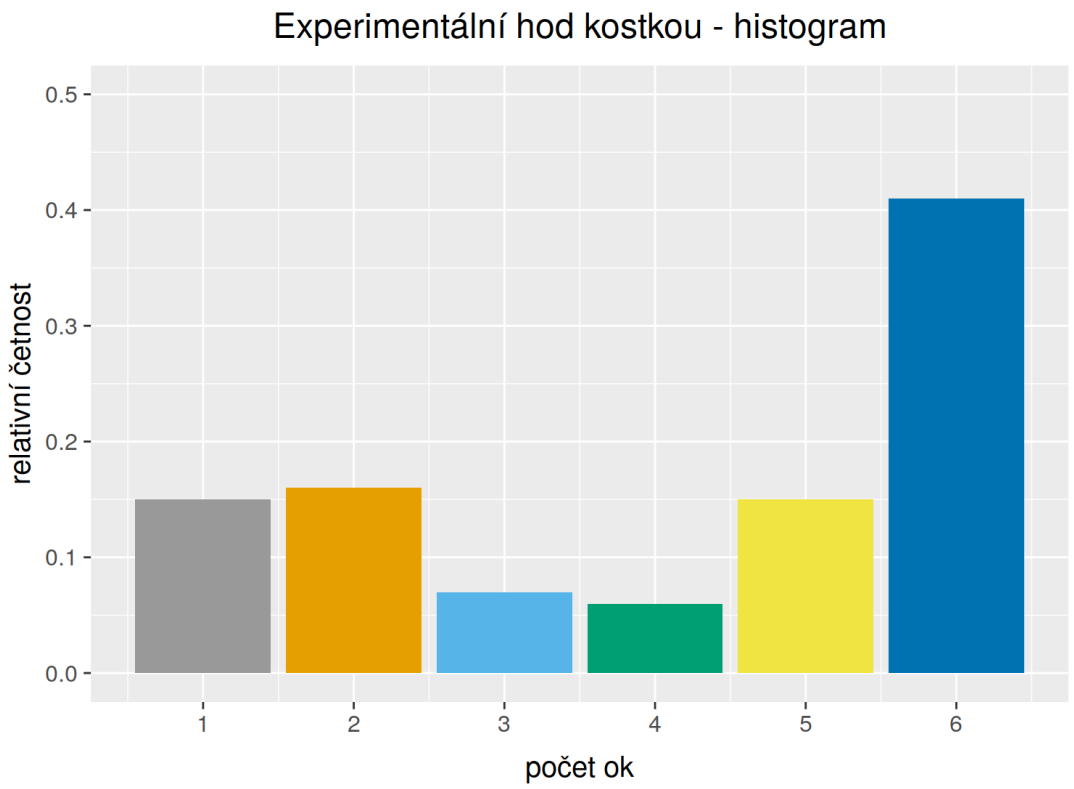
[#] Fisher's exact test for categorical and [†] Wilcoxon two-sample sum rank test were used to compare the gout sub-group with the hyperuricemia sub-group; * relative frequencies when missing information about familial occurrence was excluded; [$] onset (gout) and age of ascertainment (hyperuricemia). IQR, interquartile range; SUA, serum uric acid; FEUA, fractional excretion of uric acid.

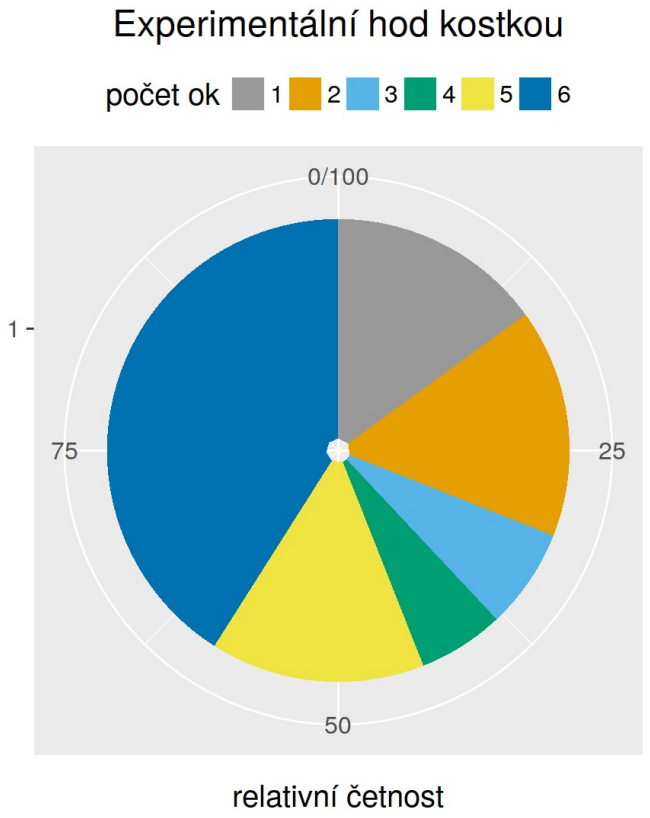
GRAFY – SPOJITÁ PROMĚNNÁ



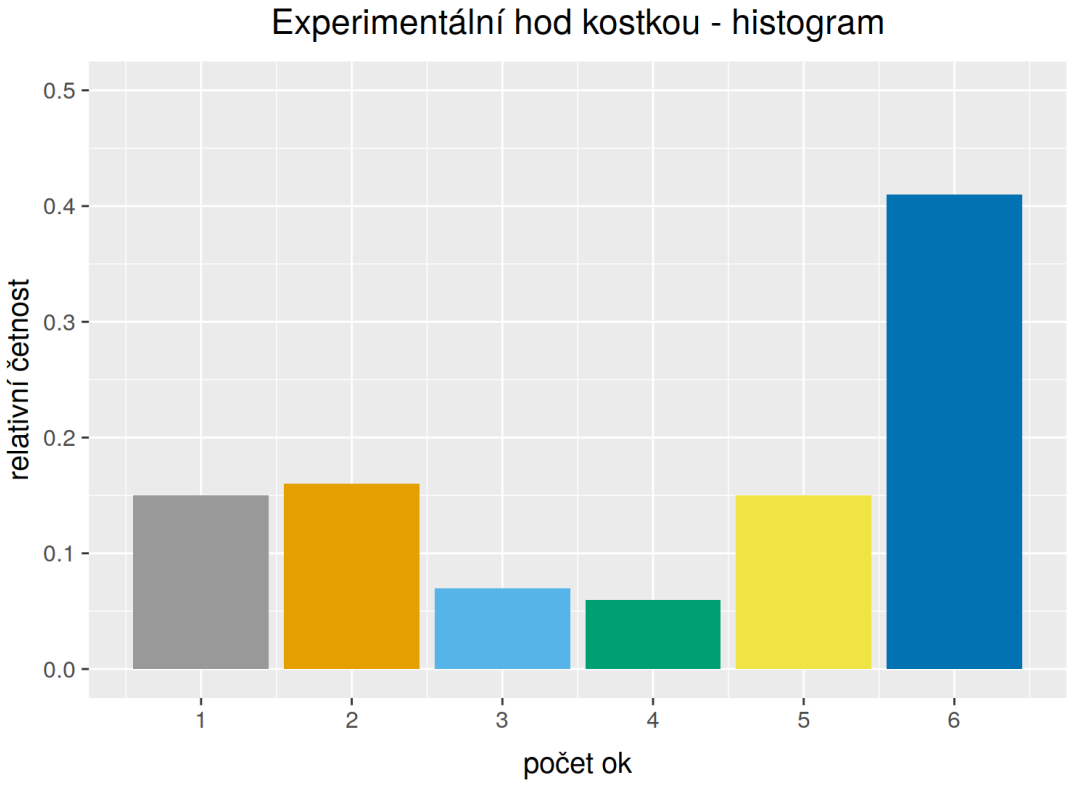


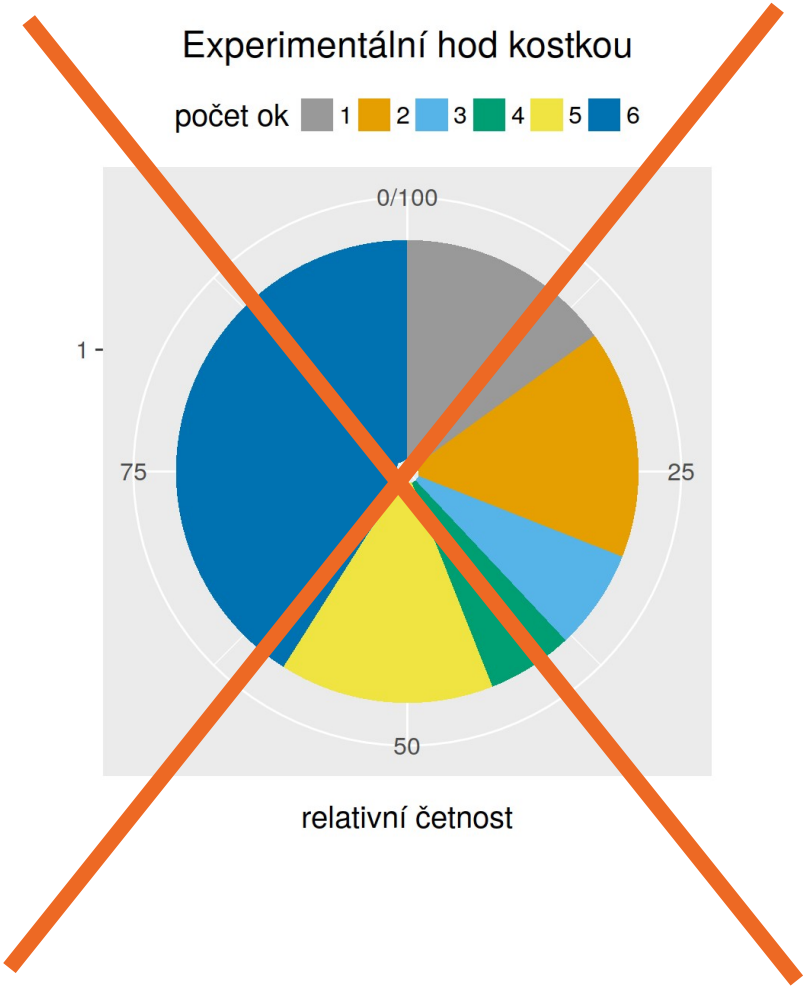
GRAFY – KATEGORIÁLNÍ PROMĚNNÁ



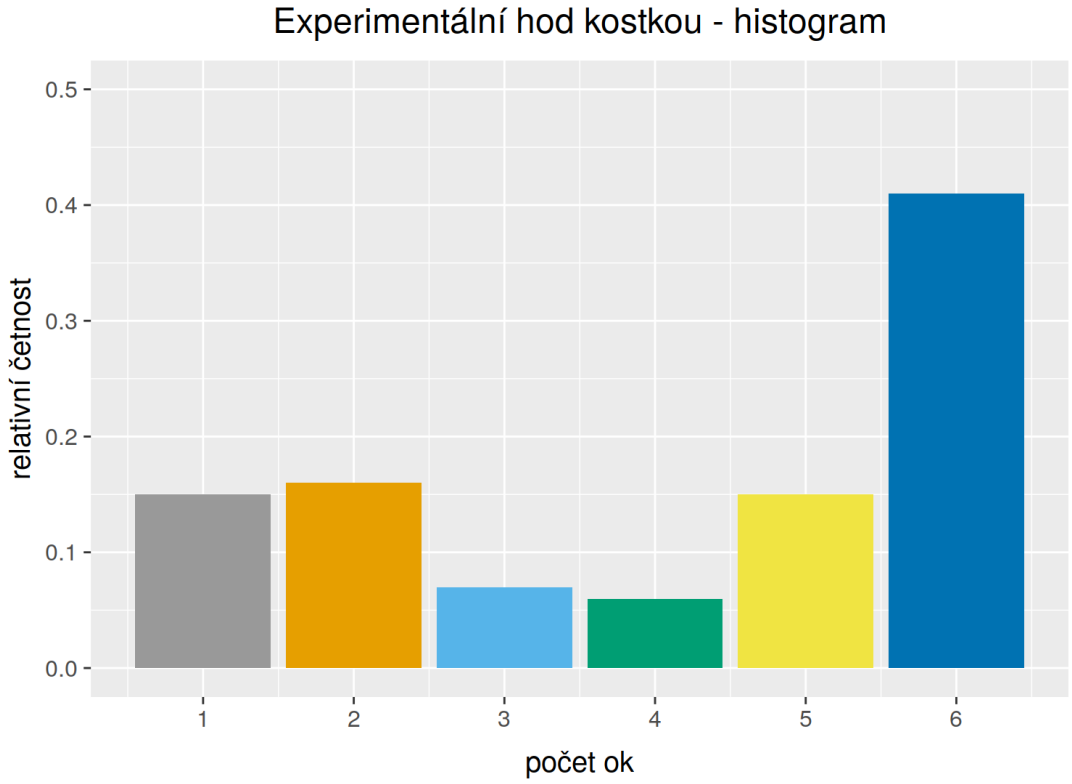


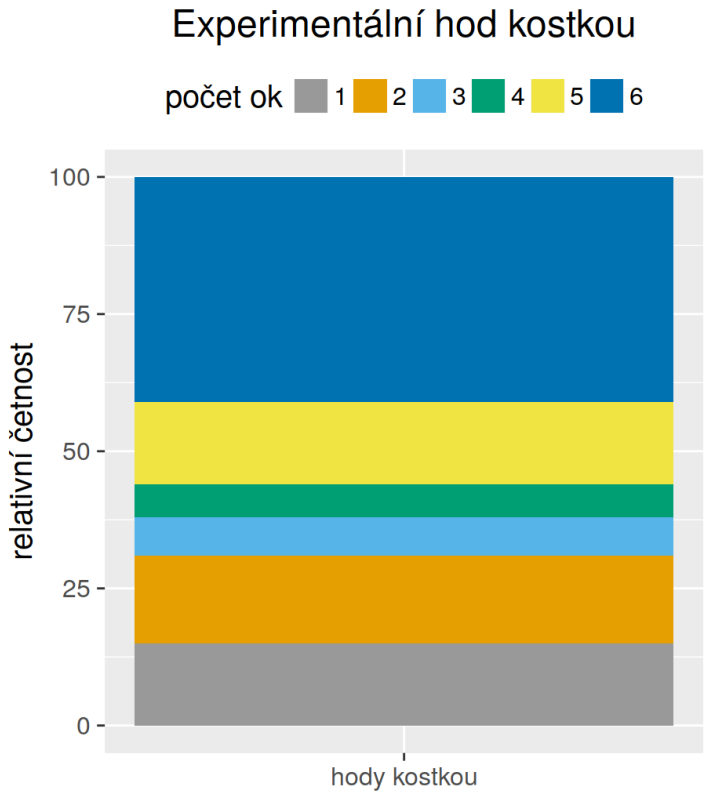
GRAFY – KATEGORIÁLNÍ PROMĚNNÁ





GRAFY – KATEGORIÁLNÍ PROMĚNNÁ





TESTOVÉ STATISTIKY

Dvě proměnné: popis vztahu, hledání (ne)závislosti

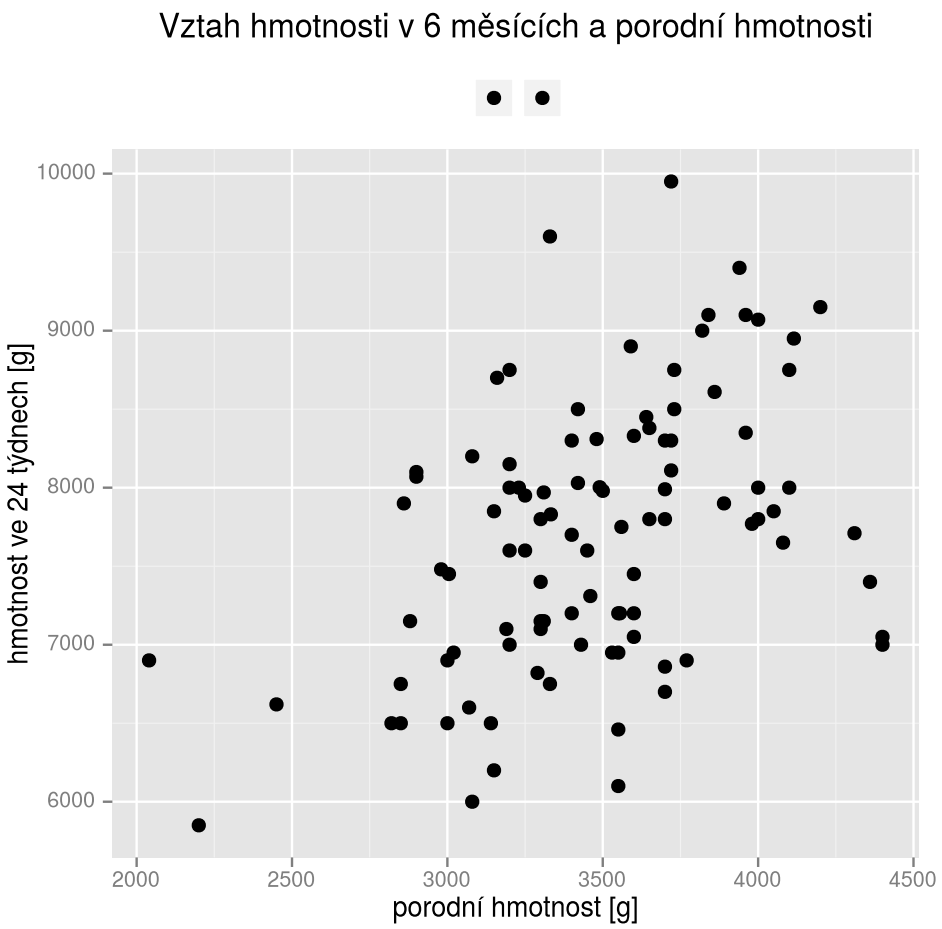
Podle typu proměnných

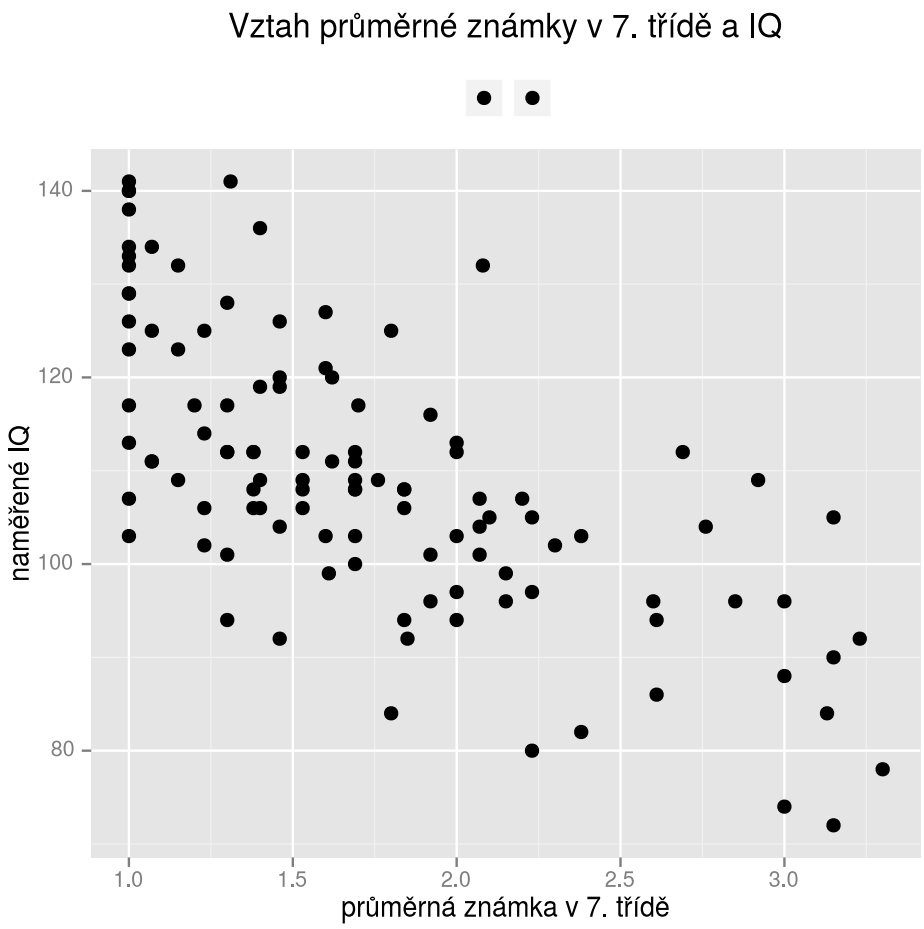
- kvantitativní a kvantitativní
  - bodový graf (scatterplot) → korelační koeficient, regrese
- kvantitativní a kvalitativní
  - krabicový graf (boxplot) → t-test, ANOVA, Wilcoxonův test, ...
- kvalitativní a kvalitativní
  - kontingenční tabulka, sloupcový graf
  - → χ^2-test, Fisherův test

PŘEHLED TESTŮ

rozdělení	**normální**	**spojité**	**alternativní / diskrétní**
populační parametr (o čem je hypotéza)	**populační průměr**	**populační medián (distribuční funkce)**	**pravděpodobnost jevu / (ne)závislost / poměr šancí**
jeden výběr	jednovýběrový t-test	jednovýběrový Wilcoxonův test znaménkový test	test proporcí
výběr dvojic	párový t-test	párový Wilcoxonův test znaménkový test	McNemarův test
dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné)	dvouvýběrový t-test	Mann-Whitney (dvouvýběrový Wilcoxon) Kolmogorov-Smirnov	Fisherův exaktní test χ^2-test
k nezávislých výběrů (závislost na kategoriální proměnné)	analýza rozptylu (jednoduché třídění, F-test)	Kruskal-Wallis	Fisherův exaktní test χ^2-test
závislost na spojité proměnné	lineární regrese	zobecněná lineární regrese (speciální případy) jádrová regrese (kernel smoothing) ...	logistická regrese multinomiální logistická regrese
opakovaná měření	smíšená lineární regrese (mixed models)	smíšená zobecněná lineární regrese	

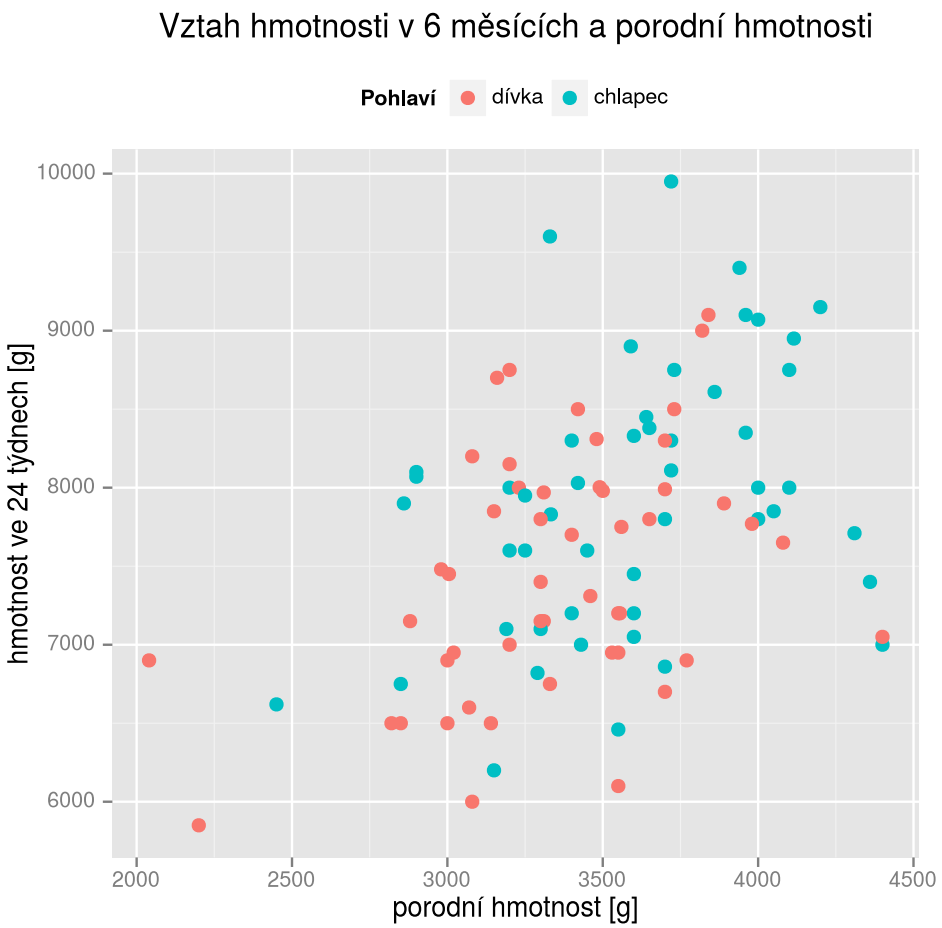
KVANITATIVNÍ A KVANTITATIVNÍ

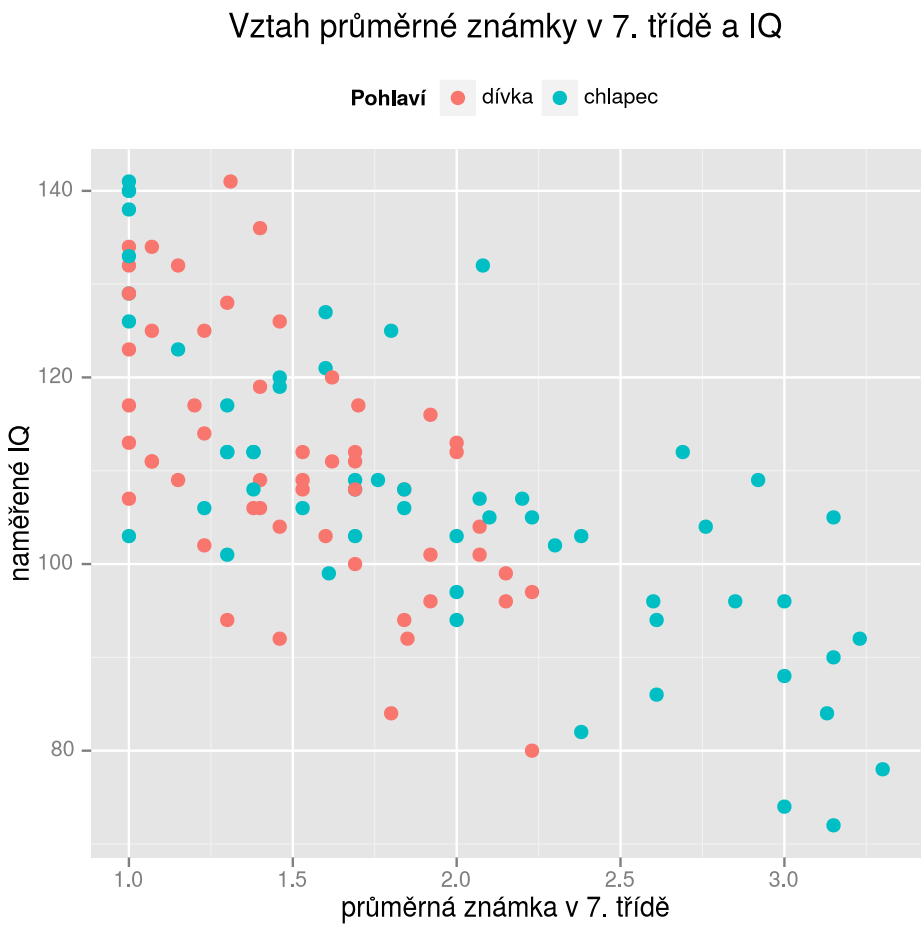




Data Kojeni.csv (vlevo) a data Iq3.csv (vpravo)

KVANITATIVNÍ A KVANTITATIVNÍ

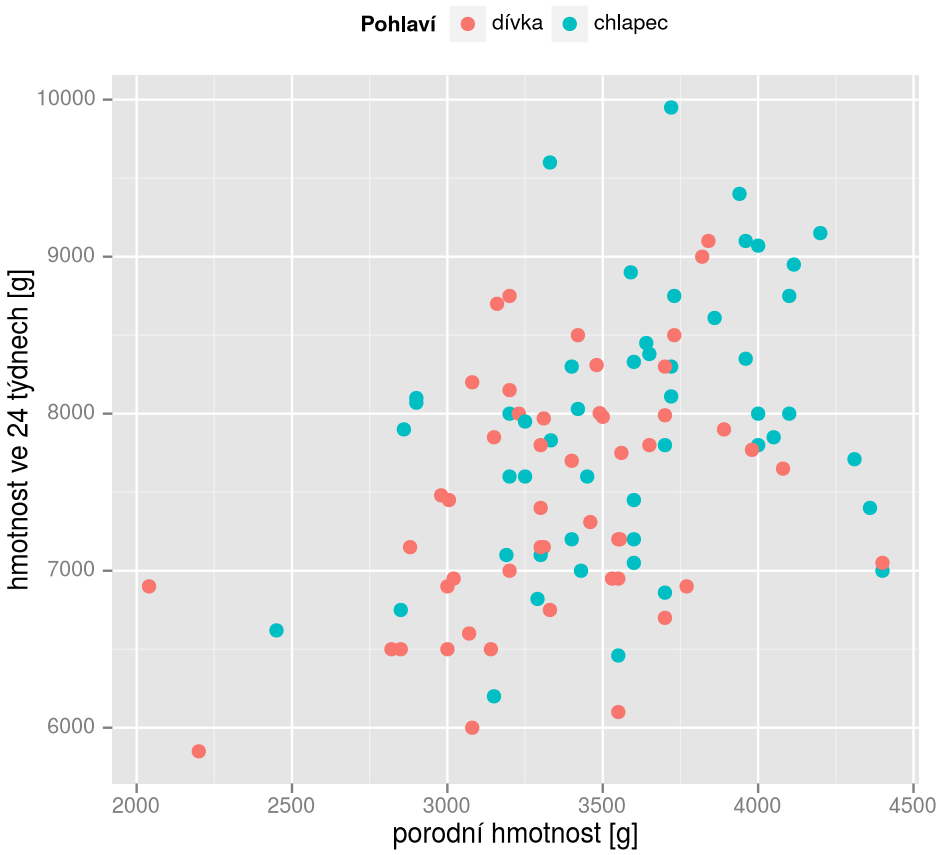




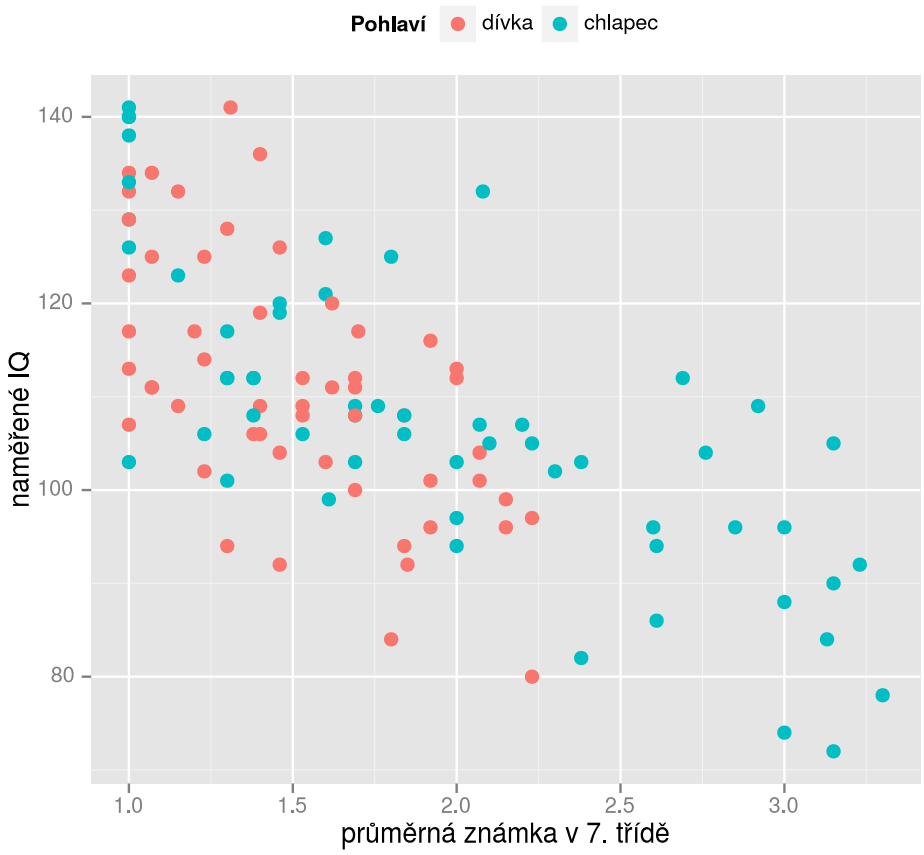
pozn: pokud záleží na směru závislosti, pak vysvětlovanou (závislou) proměnnou dáme na osu y

KVANITATIVNÍ A KVANTITATIVNÍ

Vztah hmotnosti v 6 měsících a porodní hmotnosti

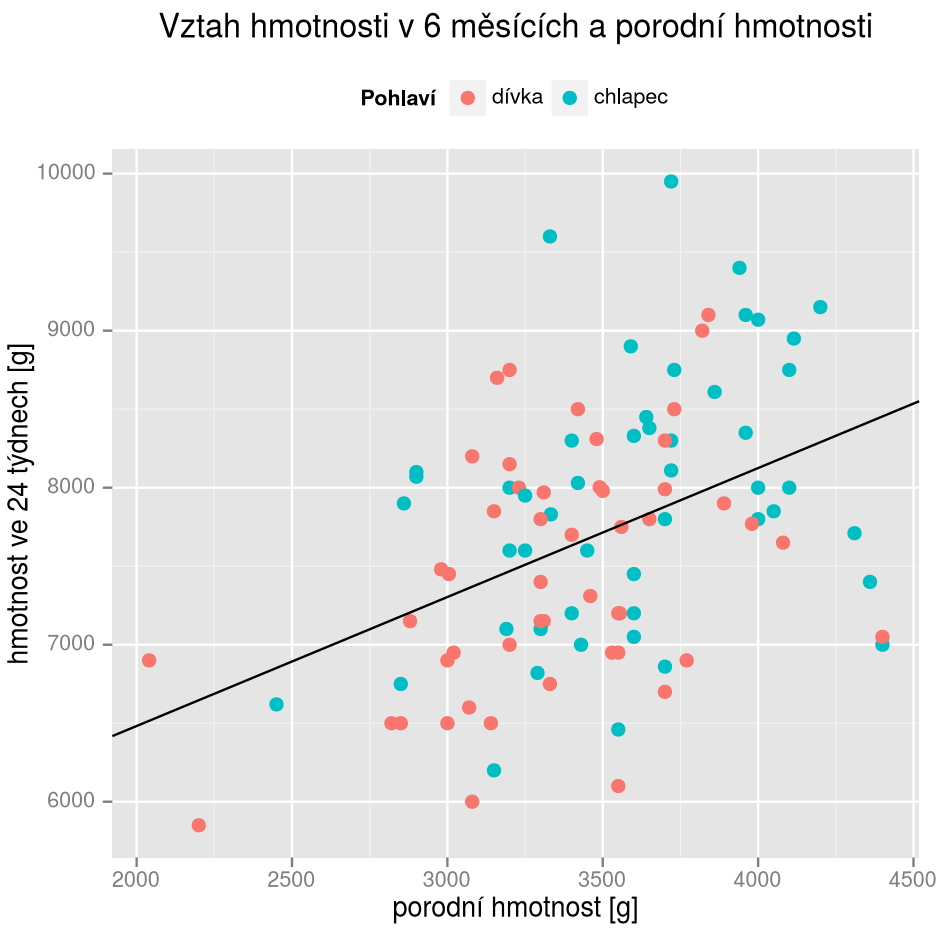


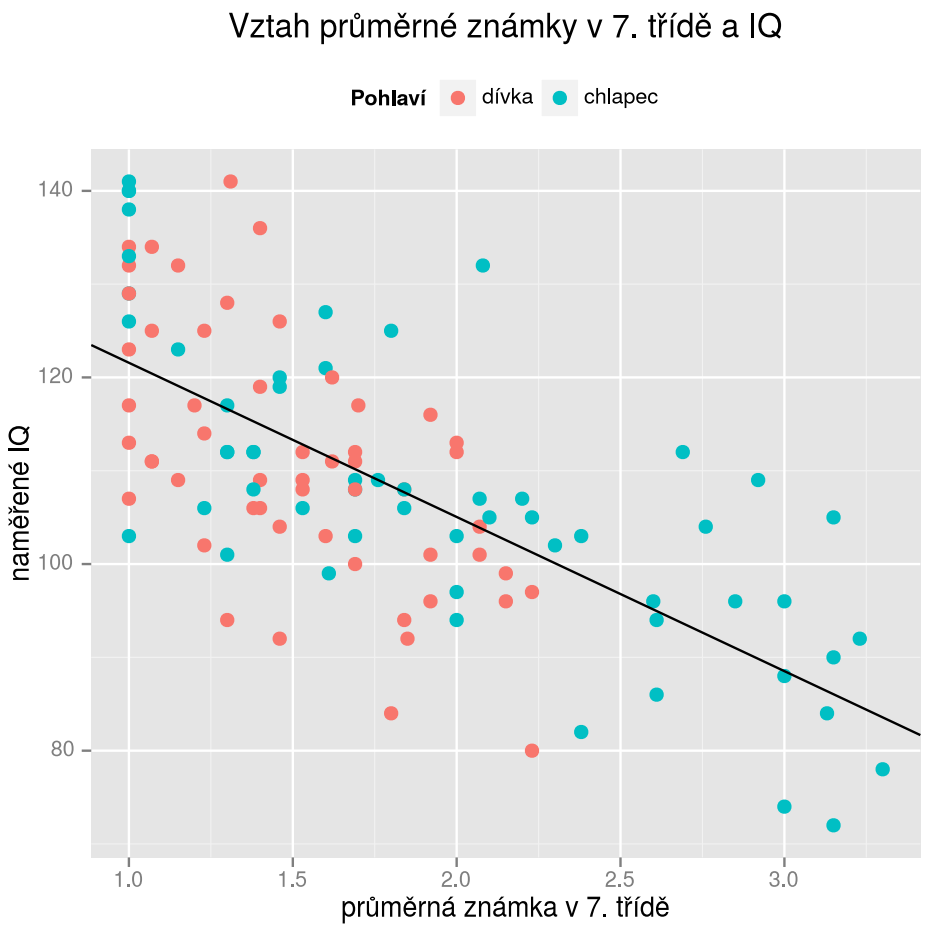
Vztah průměrné známky v 7. třídě a IQ



r = 0.429 výběrový korelační koeficient r = -0.688

KVANITATIVNÍ A KVANTITATIVNÍ



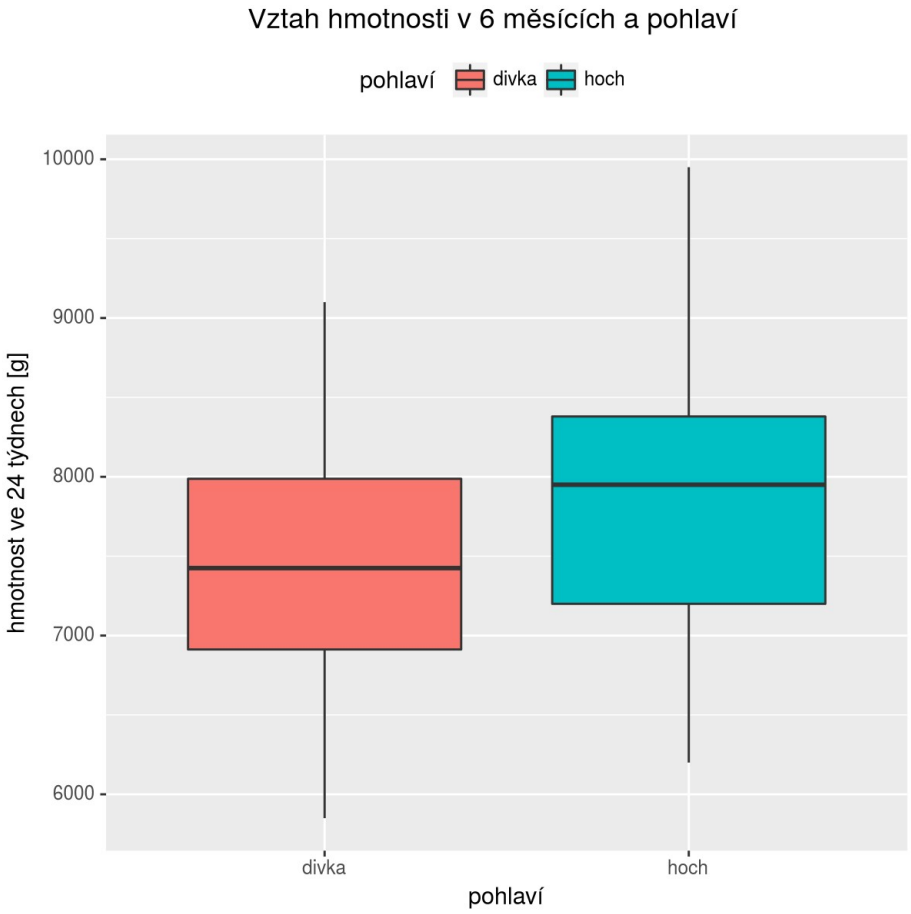


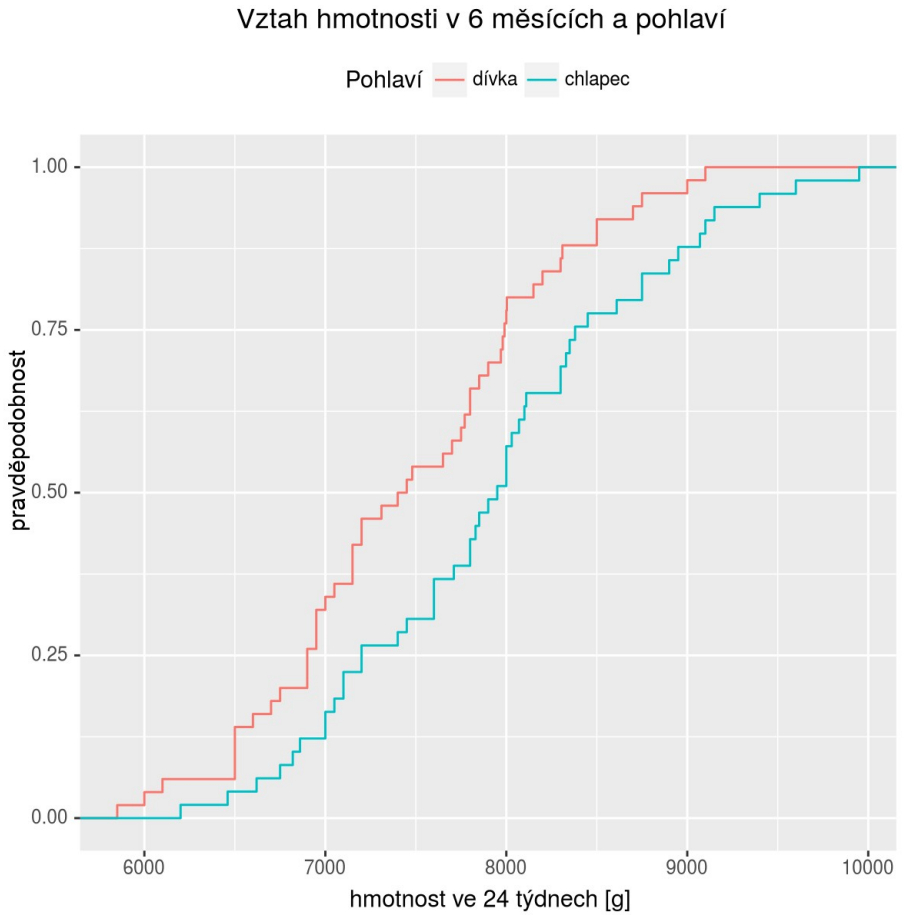
r = 0.429 výběrový korelační koeficient r = -0.688

KVANITATIVNÍ A KVALITATIVNÍ

- máme několik kategorií (skupin) a v nich mají jednotlivci naměřené kvantitativní hodnoty
- srovnáváme soubory dat v těchto skupinách
- lze chápat jako (ne)závislost kvantitativní na kvalitativní
- zobrazení boxplotem nebo empirické distribuční funkce
- nezávislost pokud krabice podobně umístěné

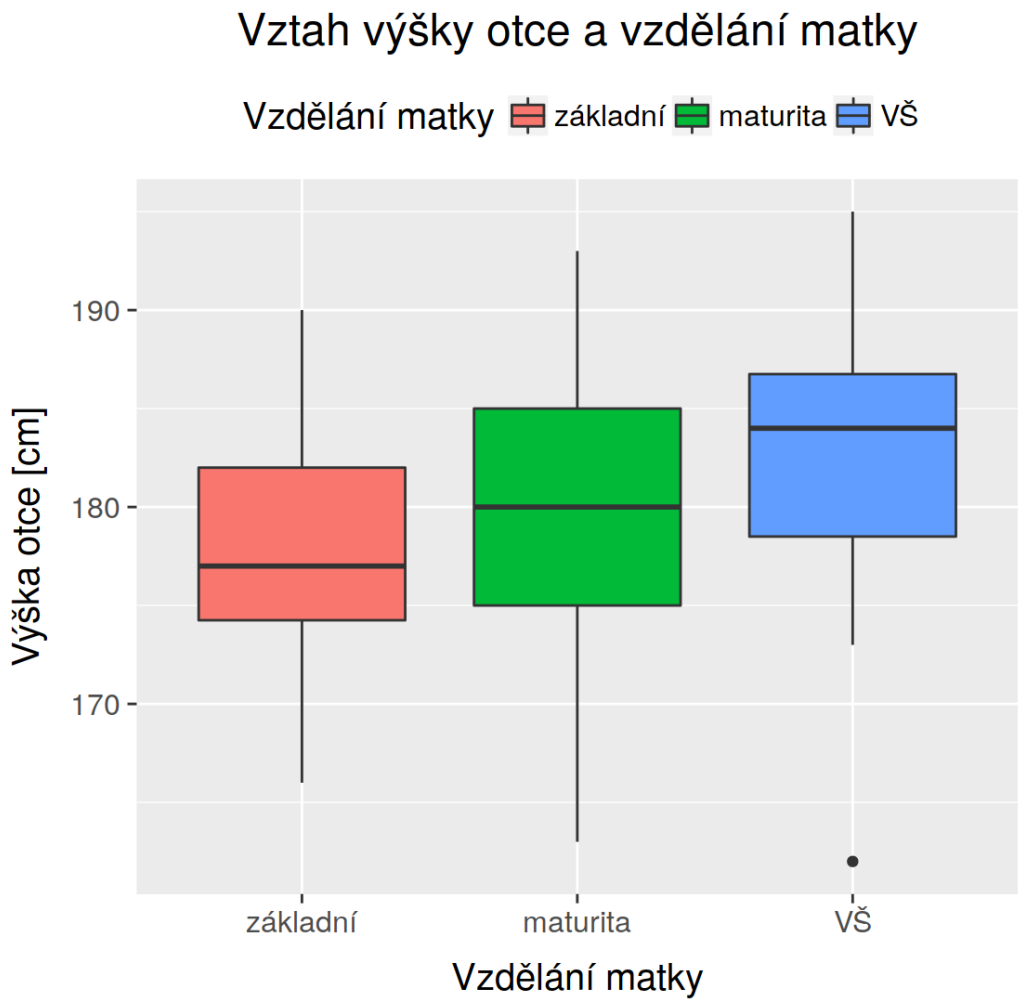
KVANITATIVNÍ A KVALITATIVNÍ





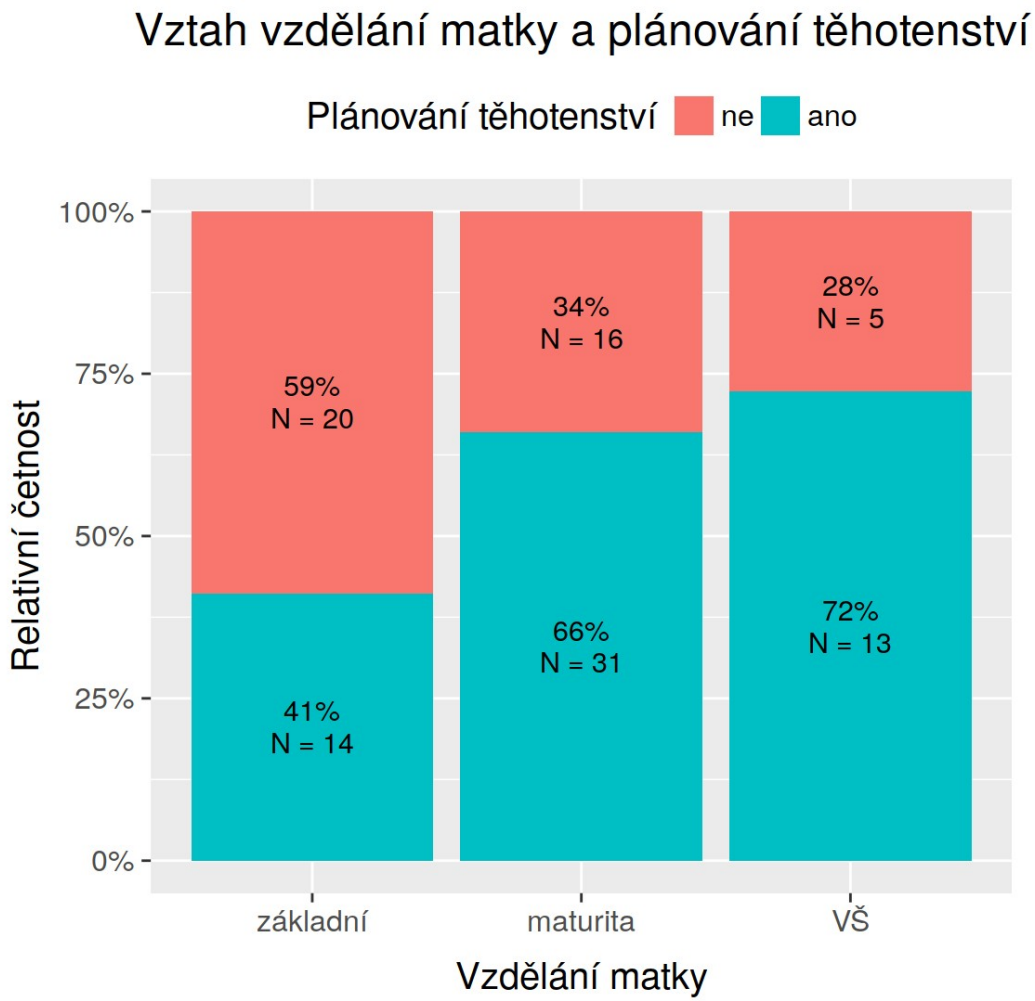
KVANITATIVNÍ A KVALITATIVNÍ

- vidíme dokonce i jistý trend
- lze posuzovat jak různost výšky otců mezi skupinami navzájem, tak možnost, že každý stupeň vzdělání výšku někam posune výš



KVALITATIVNÍ A KVALITATIVNÍ

- zde se ukazuje výhoda sloupců oproti koláčům
- lze posuzovat různost v četnostech
- vidíme dokonce i jistý trend



PRINCIPY STATISTICKÉHO TESTOVÁNÍ

- identifikace závisle proměnné
  - typ (spojitá, kategoriální, předpoklad rozdělení ...)
- identifikace kontextu / hypotézy
  - rozdíly ve skupinách = závislost na příslušnosti ke skupině?
  - závislost na hodnotě jiné proměnné? jakého typu je?
  - závislost na čase? opakovaná měření? ...
- identifikace parametru / principu
  - parametr polohy? rozptylu? kovariance (nezávislost)?
- výběr vhodné techniky

PŘEHLED TESTŮ

rozdělení	**normální**	**spojité**	**alternativní / diskrétní**
populační parametr (o čem je hypotéza)	**populační průměr**	**populační medián (distribuční funkce)**	**pravděpodobnost jevu / (ne)závislost / poměr šancí**
jeden výběr	jednovýběrový t-test	jednovýběrový Wilcoxonův test znaménkový test	test proporcí
výběr dvojic	párový t-test	párový Wilcoxonův test znaménkový test	McNemarův test
dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné)	dvouvýběrový t-test	Mann-Whitney (dvouvýběrový Wilcoxon) Kolmogorov-Smirnov	Fisherův exaktní test χ^2-test
k nezávislých výběrů (závislost na kategoriální proměnné)	analýza rozptylu (jednoduché třídění, F-test)	Kruskal-Wallis	Fisherův exaktní test χ^2-test
závislost na spojité proměnné	lineární regrese	zobecněná lineární regrese (speciální případy) jádrová regrese (kernel smoothing) ...	logistická regrese multinomiální logistická regrese
opakovaná měření	smíšená lineární regrese (mixed models)	smíšená zobecněná lineární regrese	

JEDNOVÝBĚROVÉ TESTY

Modelový příklad: výška desetiletých chlapců v roce 1961

- 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147
- $\mu_{51} = 136.1$

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

- t-test
- znaménkový test (sign test)
  - kolik chlapců v roce 1961 je menších než populační průměr 1951?
  - pokud průměr stejný, mělo by to být půl napůl, pokud vyšší, bude malých méně
  - **130**, 140, **136**, 141, 139, **133**, 149, 151, 139, **136**, 138, 142, **127**, 139, 147
  - totéž jako udělat pro každého rozdíl mezi jeho výškou a μ_{51} a sečíst počet minusů
  - 5 z 15 = 33%, p = 0.20 (jednostranná), nezamítáme
  - oproti t-testu hodně hrubé, ale zase nepotřebuje jiný předpoklad než spojitost
  - pokud je někde 0, vyhodíme měření úplně

JEDNOVÝBĚROVÉ TESTY

- Wilcoxonův pořadový test (Wilcoxon signed rank test)
  - předpoklad spojité a **symetrické**
  - stejně jako u znaménkového: uděláme rozdíl mezi výškou každého chlapce a μ_{51}, ignorujeme nuly, zbude N_r pozorování
  - pamatujeme si znaménko (sgn)
  - seřadíme si hodnoty vzestupně a nahradíme je pořadím
  - tím přestane záležet na rozdělení jako takovém!!
  - testová statistika $$W = \sum_{i=1}^{N_r} [\mathrm{sgn}(x_i) \cdot R_i]$$
  - rozdělení poloha 0, kritická hodnota v tabulkách / počítač
  - pro $N_r \geq 10$ má Z-skór přibližně normální rozdělení, případně Yatesova korekce

$$z = \frac{W}{\sigma_W}, \sigma_W = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}}$$

PÁROVÉ TESTY

- Máme dvojice provázaných pozorování, jednotlivé dvojice na sobě nezávislé
  - výška otce a syna v dospělosti
  - měření jednoho metabolitu dvěma různými způsoby
  - měření před a po terapii
- Testujeme, zda se parametr polohy měření ve dvojici liší.
- Použijeme jednovýběrový test na rozdíl ve dvojici, testujeme, zda je poloha rozdílu 0.
  - t-test (průměr rozdílů je nula, normální nemusí být měření, stačí rozdíly)
  - znaménkový test (polovina rozdílů je záporných)
  - Wilcoxonův párový test (medián rozdílů je 0)

DVOUVÝBĚROVÉ TESTY

- N_X nezávislých pozorování X, N_Y nezávislých pozorování Y,
- výběry jsou také navzájem nezávislé (zajistí způsob pořízení dat)
- chceme
  - vědět, zda X a Y mají stejné rozdělení (ne nutně konkrétní)
  - vědět, zda je jedno rozdělení „větší“ než druhé (tzv. stochasticky dominantní, speciální případ je stejné s vyšším parametrem polohy) nebo stejné
  - má za předpokladu normálního rozdělení se stejným rozptylem různou (nebo stejnou) střední hodnotu

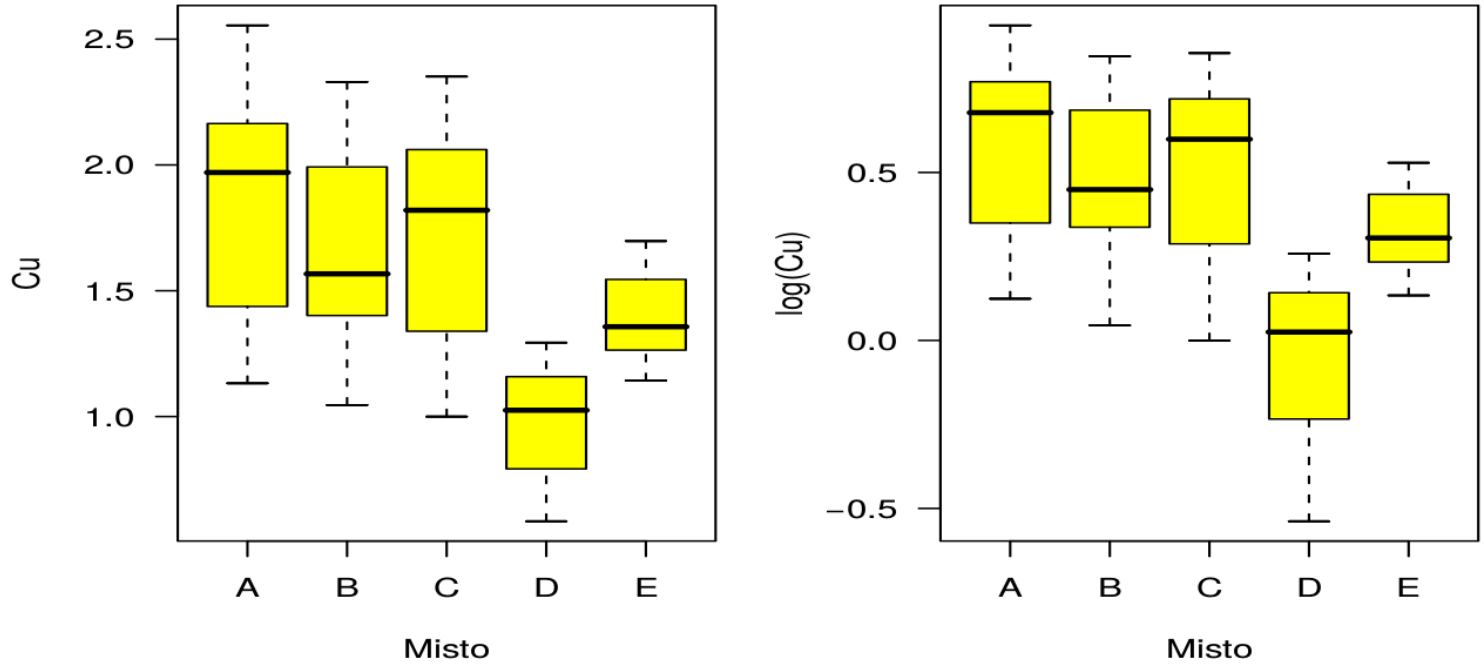
DVOUVÝBĚROVÉ TESTY

- dvouvýběrový t-test
  - předpoklad normálního rozdělení a stejného rozptylu
  - (pro velká N_X a N_Y není normalita tak podstatná)
  - testujeme shodnost $\mu_X = \mu_Y$
- Wilcoxonův dvouvýběrový test (Mann-Whitneyův)
  - libovolné spojité rozdělení, testujeme posun polohy
  - místo skutečných hodnot použijeme jejich pořadí
  - při platnosti H_0 by měly být oba soubory dobře „promíchané“

POROVNÁNÍ VÍCE VÝBĚRŮ

motivační příklad pro analýzu rozptylu (játra):

- pět míst na řece, vždy vyloveno po 7 rybách
- zjišťována koncentrace mědi v játrech
- liší se tato místa svým znečištěním?
- logaritmování na pravé straně stabilizuje rozptyl



POROVNÁNÍ VÍCE VÝBĚRŮ

- ANOVA = ANalysis Of Variance
  - celková variabilita v datech = součet čtverců vzdáleností od celkového průměru
  - variabilita mezi faktory = vážený součet čtverců rozdílů průměrů v rámci faktoru od celkového průměru
  - variabilita uvnitř v rámci jednotlivých faktorů = součet čtvreců vzdáleností od průměru uvnitř faktoru
- když je variabilita mezi faktory dostatečně vysoká, znamená to, že se faktory mezi sebou liší
- parametrická (předpoklad normálního rozdělení)
- neparametrická = Kruskall-Wallisův test

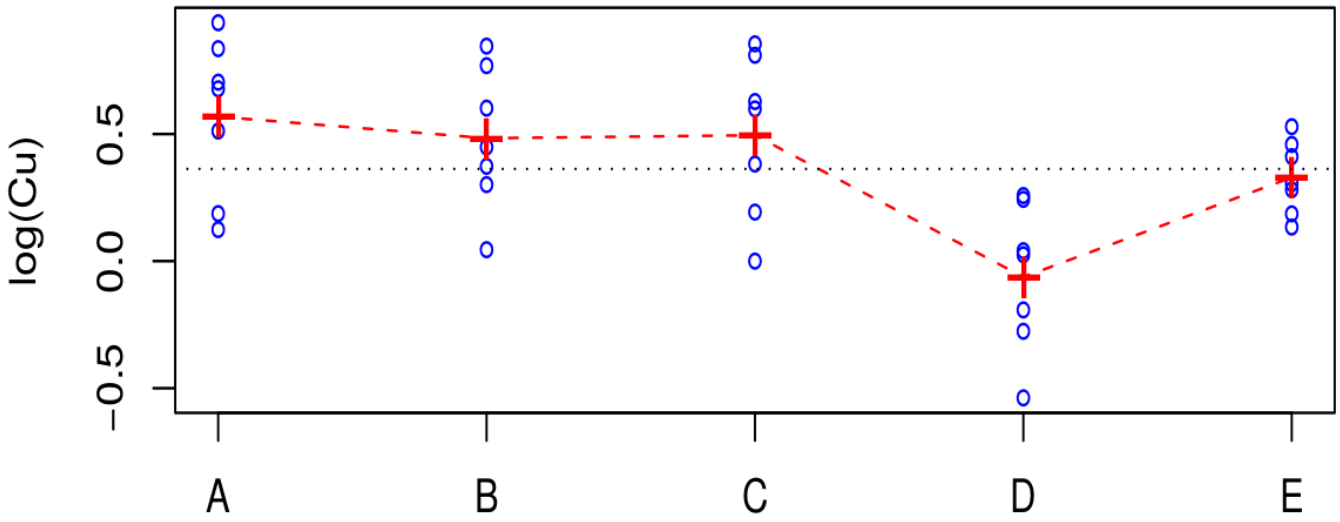
ANALÝZA ROZPTYLU

rozklad součtu čtverců

příklad játra (celkový průměr $\bar{y}_{\bullet\bullet} = 0{,}36$)

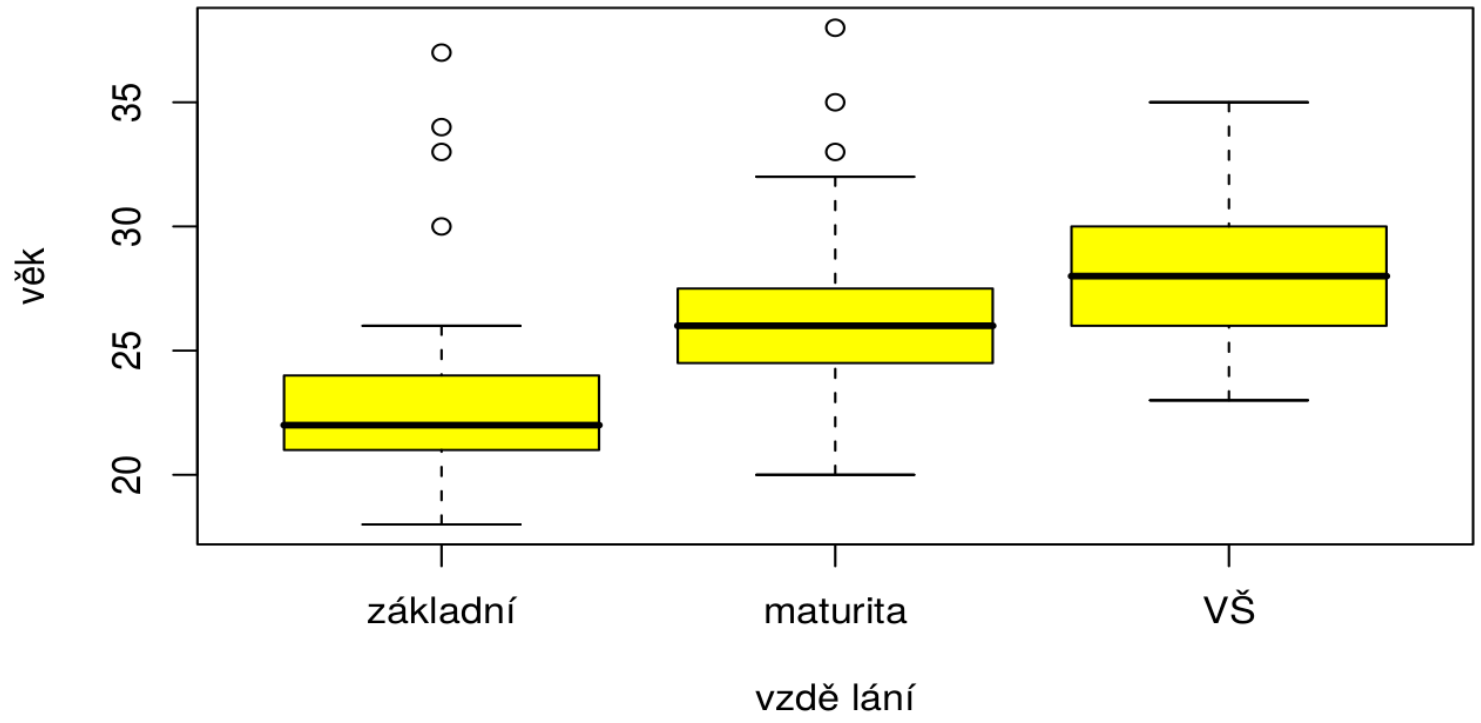
$$(\text{celková variabilita}) = (\text{variabilita mezi}) + (\text{variabilita uvnitř})$$

$$\sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{i\bullet})^2$$



ANALÝZA ROZPTYLU

příklad kojení – věk matek podle vzdělání



je patrná nesymetrie, zejména u základního vzdělání

LINEÁRNÍ REGRESNÍ MODELY

- Vztahy mohou být složitější:
  - variabilita může být způsobena příslušností k nějaké konkrétní skupině (např. tělesná výška a pohlaví)
  - může být ovlivněna i nějakým spojitým faktorem (např. věk)
  - vliv věku na výšku může být jiný pro dívky a chlapce (tzv. interakce

→ konstrukce složitějších lineárních regresních modelů

  - můžeme mít více měření pro jednu osobu

→ smíšený lineární regresní model

HODNOCENÍ KVALITATIVNÍCH ZNAKŮ

- kategoriální proměnné
- znaky v nominálním měřítku (neuspořádané hodnoty)
- existují techniky i pro ordinální měřítko: více struktury = přesněji zacílené testy
- příklady
  - počet osob s krevní skupinou A, B, AB, 0
  - počet dětí narozených v jednotlivých měsících v roce v Praze
  - vzdělání matky novorozence (ZŠ, SŠ, VŠ)

HODNOCENÍ KVALITATIVNÍCH ZNAKŮ

Co nás zajímá?

- A. u jedné proměnné: jsou pravděpodobnosti výskytu jednotlivých kategorií takové, jaké očekáváme?
  - ekvivalent otázky na polohu spojitého rozdělení
  - test typu goodness-of-fit
- B. u dvou (nebo více) populací: jsou pravděpodobnosti výskytu jednotlivých kategorií u obou skupin stejné?
  - ekvivalent otázky na stejný parametr polohy (dvouvýběrový test)
  - test homogenity
- C. dvou proměnných (znaků): jsou na sobě závislé?
  - ekvivalent korelačního koeficientu nebo regrese
  - test nezávislosti

χ^2-TEST

- ve všech třech případech vede na χ^2-test
- porovnání vzdálenosti mezi pozorovanou hodnotou a očekávanou hodnotou

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- **chí-kvadrát test dobré shody** $H_0 : \pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$ (pravděpodobnosti jsou hypotézou dány **jednoznačně**)
- platí-li H_0, očekáváme četnosti blízké hodnotám $E\,N_j = n\pi_j^0$:
- H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(1-\alpha)$,

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}$$

nezapomeň, že očekávané četnosti musí být alespoň 5

- N_j – **empirické** (experimentální) četnosti, $n\pi_j^0$ – **očekávané** (za platnosti H_0, teoretické) četnosti
- statistika X^2 (velké chí-kvadrát) porovnává empirické a očekávané četnosti (měří jejich neshodu, „vzdálenost“)

HODNOCENÍ KVALITATIVNÍCH ZNAKŮ

- Fisherův exaktní test
  - použití podobné jako χ^2-test, i pro nízké počty
  - dívá se na pravděpodobnost existence „naší“ tabulky ze všech možný s pevně danými součty po sloupcích a po řádcích
- McNemarův test pro testování symetrie v tabulce, ekvivalent párového testu (např. stav před ošetřením a po ošetření)
- Cochran-Mantel-Haenszelův test, pokud jsou data ještě s třetí úrovní struktury (např. různá centra)
- RR = risk ratio, OR = odds ratio a jeho konfidenční interval
  - odhad rizika při přítomnosti určitého faktoru
- logistická regrese modelující OR v závislosti na různých faktorech

PŘEHLED TESTŮ

rozdělení	**normální**	**spojité**	**alternativní / diskrétní**
populační parametr (o čem je hypotéza)	**populační průměr**	**populační medián (distribuční funkce)**	**pravděpodobnost jevu / (ne)závislost / poměr šancí**
jeden výběr	jednovýběrový t-test	jednovýběrový Wilcoxonův test znaménkový test	test proporcí
výběr dvojic	párový t-test	párový Wilcoxonův test znaménkový test	McNemarův test
dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné)	dvouvýběrový t-test	Mann-Whitney (dvouvýběrový Wilcoxon) Kolmogorov-Smirnov	Fisherův exaktní test χ^2-test
k nezávislých výběrů (závislost na kategoriální proměnné)	analýza rozptylu (jednoduché třídění, F-test)	Kruskal-Wallis	Fisherův exaktní test χ^2-test
závislost na spojité proměnné	lineární regrese	zobecněná lineární regrese (speciální případy) jádrová regrese (kernel smoothing) ...	logistická regrese multinomiální logistická regrese
opakovaná měření	smíšená lineární regrese (mixed models)	smíšená zobecněná lineární regrese	

PREZENTACE VÝSLEDKŮ

PREZENTACE VÝSLEDKŮ V PRÁCI

Držte se struktury:

- Popis
  - kolik, koho, kdy a kde jsem sebral(a)
  - kolik, koho a proč jsem nesebral(a)
  - jak vypadali obecně
  - jak vypadaly jejich zájmové proměnné
- Analýza
  - test stanovených hypotéz
  - popis doplňkových analýz
- Pouze popisujete, výklady si nechejte do diskuse

PREZENTACE VÝSLEDKŮ V PRÁCI

- Ke každému číslu uvést
  - zda je to průměr, medián apod
  - k parametru polohy (průměr …) i parametr rozptylu (SD, rozsah ..)
  - jednotky!
- Procenta
  - píšeme s (tvrdou) mezerou mezi číslem a % (ctrl + shift + space)
  - zaokrouhlovat na 1 des. místo, případně rovnou na jednotky %
  - uvádět i absolutní počty
- Smysluplná tabulka / graf pro jednotlivé analýzy

PREZENTACE VÝSLEDKŮ V PRÁCI

- Každý obrázek a každá tabulka musí mít popis
  - záhlaví: stručně co je obsahem tabulky / grafu, případně počet (N = 193)
  - zápatí: vysvětlení všech uvedených zkratek a znaků
  - jednotky v tabulce i v grafu, obvykle do hranatých závorek
  - tabulce i grafu by měl čtenář rozumět bez čtení textu
- Grafy
  - nepřehánějte to, nedává smysl ilustrovat podíl mužů a žen graficky
  - něco je na druhou stranu daleko srozumitelnější v grafu než sáhodlouhý popis
- Na každou tabulku a graf musí být v textu odkaz

DĚKUJI ZA POZORNOST

www.biostatisticka.cz