Statistika ve společenskovědním výzkumu 1

přednášející:	Martin Schindler
	KAP, tel. 48 535 2836, budova G
konzul. hodiny:	po dohodě
e-mail:	martin.schindler@tul.cz

naposledy upraveno: 21. února 2018

Požadavek na udělení zápočtu: vypracování a obhájení semestrální práce.

Literatura

- Chráska, M.: Metody pedagogického výzkumu Praha:Grada, 2007.
- HENDL, J.: Přehled statistických metod zpracování dat. Portál: Praha, 2012 (4.vyd.)
- ZVÁRA K., ŠŤĚPÁN J.: Pravděpodobnost a matematická statistika. Praha: Matfyzpress, 2002.

Literatura online

- http://moodle.vsb.cz/vyuka/course/info.php?id=3
- http://www.studopory.vsb.cz/
- http://mathonline.fme.vutbr.cz/
- http://home.zcu.cz/ friesl/hpsb/tit.html
- tato prezentace: http://147.230.193.199/ ms/

Statistika

- **statistika** je jedním z oborů zabývajících se shromažďováním, zpracováním a analyzováním dat vznikajících při studiu tzv. **hromadných jevů**, což jsou jevy vyskytující se teprve u velkého souboru případů, ne jen u případů jednotlivých.
- **statistický soubor** je množina **statistických jednotek** (obyvatelé, obce, firmy,...), na nichž měříme (zjišťujeme) hodnoty **statistických znaků**(věk, počet obyvatel, obrat,...)
- zjištěnou hodnotu znaku vyjadřujeme ve vhodně zvoleném **měřítku** (stupnici).
- na jedné jednotce můžeme měřit několik znaků - to umožní vyšetřovat závislost (existuje souvislost mezi výškou a hmotností osob ve studované populaci?).

Statistika

- **statistika** je jedním z oborů zabývajících se shromažďováním, zpracováním a analyzováním dat vznikajících při studiu tzv. **hromadných jevů**, což jsou jevy vyskytující se teprve u velkého souboru případů, ne jen u případů jednotlivých.
- **statistický soubor** je množina **statistických jednotek** (obyvatelé, obce, firmy,...), na nichž měříme (zjišťujeme) hodnoty **statistických znaků**(věk, počet obyvatel, obrat,...)
- zjištěnou hodnotu znaku vyjadřujeme ve vhodně zvoleném **měřítku** (stupnici).
- na jedné jednotce můžeme měřit několik znaků - to umožní vyšetřovat závislost (existuje souvislost mezi výškou a hmotností osob ve studované populaci?).

Statistika

- **statistika** je jedním z oborů zabývajících se shromažďováním, zpracováním a analyzováním dat vznikajících při studiu tzv. **hromadných jevů**, což jsou jevy vyskytující se teprve u velkého souboru případů, ne jen u případů jednotlivých.
- **statistický soubor** je množina **statistických jednotek** (obyvatelé, obce, firmy,...), na nichž měříme (zjišťujeme) hodnoty **statistických znaků**(věk, počet obyvatel, obrat,...)
- zjištěnou hodnotu znaku vyjadřujeme ve vhodně zvoleném **měřítku** (stupnici).
- na jedné jednotce můžeme měřit několik znaků - to umožní vyšetřovat závislost (existuje souvislost mezi výškou a hmotností osob ve studované populaci?).

Statistika

- **statistika** je jedním z oborů zabývajících se shromažďováním, zpracováním a analyzováním dat vznikajících při studiu tzv. **hromadných jevů**, což jsou jevy vyskytující se teprve u velkého souboru případů, ne jen u případů jednotlivých.
- **statistický soubor** je množina **statistických jednotek** (obyvatelé, obce, firmy,...), na nichž měříme (zjišťujeme) hodnoty **statistických znaků**(věk, počet obyvatel, obrat,...)
- zjištěnou hodnotu znaku vyjadřujeme ve vhodně zvoleném **měřítku** (stupnici).
- na jedné jednotce můžeme měřit několik znaků - to umožní vyšetřovat závislost (existuje souvislost mezi výškou a hmotností osob ve studované populaci?).

Ke studovanému datovému souboru lze přistoupit dvěma způsoby:

1. **Popisná statistika** - ze zjištěných dat chceme činit závěry pouze pro studovaný datový soubor (prošetřili jsme celou populaci, kterou chceme popsat)
2. **Matematická (inferenční) statistika** - Studovaný soubor chápeme jako **výběrový soubor** – množina prvků vybraných náhodně a nezávisle ze **základního souboru**, který je rozsáhlý (z důvodů časových, finančních, organizačních aj. nelze prozkoumat celý). Z hodnot proměnných zjištěných ve výběrovém souboru chceme činit závěry o základním souboru (v druhé půli semestru).

Ke studovanému datovému souboru lze přistoupit dvěma způsoby:

1. **Popisná statistika** - ze zjištěných dat chceme činit závěry pouze pro studovaný datový soubor (prošetřili jsme celou populaci, kterou chceme popsat)
2. **Matematická (inferenční) statistika** - Studovaný soubor chápeme jako **výběrový soubor** – množina prvků vybraných náhodně a nezávisle ze **základního souboru**, který je rozsáhlý (z důvodů časových, finančních, organizačních aj. nelze prozkoumat celý). Z hodnot proměnných zjištěných ve výběrovém souboru chceme činit závěry o základním souboru (v druhé půli semestru).

Typy měřítek

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (rodinný stav, barva očí) - disjunktní kategorie, které nelze uspořádat
- **ordinální** (nejvyšší dosažené vzdělání, míra spokojenosti) - nominální měřítko s uspořádanými kategoriemi
- **intervalové** (teplota v Celsiové stupnici, rok narození) - možné hodnoty jsou číselně označeny, vzdálenost mezi sousedními hodnotami je konstatní
- **poměrové** (hmotnost, výška, počet obyvatel) - hodnoty jsou udávány v násobcích dohodnuté jednotky, nula znamená neexistenci měřené vlastnosti.

  - **Kvalitativní**: nula-jedničkové, nominální, ordinální
  - **Kvantitativní (spojité)**: intervalové, poměrové

Typy měřítek

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (rodinný stav, barva očí) - disjunktní kategorie, které nelze uspořádat
- **ordinální** (nejvyšší dosažené vzdělání, míra spokojenosti) - nominální měřítko s uspořádanými kategoriemi
- **intervalové** (teplota v Celsiové stupnici, rok narození) - možné hodnoty jsou číselně označeny, vzdálenost mezi sousedními hodnotami je konstatní
- **poměrové** (hmotnost, výška, počet obyvatel) - hodnoty jsou udávány v násobcích dohodnuté jednotky, nula znamená neexistenci měřené vlastnosti.

  - **Kvalitativní**: nula-jedničkové, nominální, ordinální
  - **Kvantitativní (spojité)**: intervalové, poměrové

Typy měřítek

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (rodinný stav, barva očí) - disjunktní kategorie, které nelze uspořádat
- **ordinální** (nejvyšší dosažené vzdělání, míra spokojenosti) - nominální měřítko s uspořádanými kategoriemi
- **intervalové** (teplota v Celsiové stupnici, rok narození) - možné hodnoty jsou číselně označeny, vzdálenost mezi sousedními hodnotami je konstatní
- **poměrové** (hmotnost, výška, počet obyvatel) - hodnoty jsou udávány v násobcích dohodnuté jednotky, nula znamená neexistenci měřené vlastnosti.

  - **Kvalitativní**: nula-jedničkové, nominální, ordinální
  - **Kvantitativní (spojité)**: intervalové, poměrové

Typy měřítek

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (rodinný stav, barva očí) - disjunktní kategorie, které nelze uspořádat
- **ordinální** (nejvyšší dosažené vzdělání, míra spokojenosti) - nominální měřítko s uspořádanými kategoriemi
- **intervalové** (teplota v Celsiové stupnici, rok narození) - možné hodnoty jsou číselně označeny, vzdálenost mezi sousedními hodnotami je konstatní
- **poměrové** (hmotnost, výška, počet obyvatel) - hodnoty jsou udávány v násobcích dohodnuté jednotky, nula znamená neexistenci měřené vlastnosti.

- **Kvalitativní**: nula-jedničkové, nominální, ordinální
- **Kvantitativní (spojité)**: intervalové, poměrové

Typy měřítek

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (rodinný stav, barva očí) - disjunktní kategorie, které nelze uspořádat
- **ordinální** (nejvyšší dosažené vzdělání, míra spokojenosti) - nominální měřítko s uspořádanými kategoriemi
- **intervalové** (teplota v Celsiové stupnici, rok narození) - možné hodnoty jsou číselně označeny, vzdálenost mezi sousedními hodnotami je konstatní
- **poměrové** (hmotnost, výška, počet obyvatel) - hodnoty jsou udávány v násobcích dohodnuté jednotky, nula znamená neexistenci měřené vlastnosti.

- **Kvalitativní**: nula-jedničkové, nominální, ordinální
- **Kvantitativní (spojité)**: intervalové, poměrové

Typy měřítek

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (rodinný stav, barva očí) - disjunktní kategorie, které nelze uspořádat
- **ordinální** (nejvyšší dosažené vzdělání, míra spokojenosti) - nominální měřítko s uspořádanými kategoriemi
- **intervalové** (teplota v Celsiové stupnici, rok narození) - možné hodnoty jsou číselně označeny, vzdálenost mezi sousedními hodnotami je konstatní
- **poměrové** (hmotnost, výška, počet obyvatel) - hodnoty jsou udávány v násobcích dohodnuté jednotky, nula znamená neexistenci měřené vlastnosti.
    - **Kvalitativní**: nula-jedničkové, nominální, ordinální
    - **Kvantitativní (spojité)**: intervalové, poměrové

Příklad - jednorozměrný

- jednorozměrná data (zajímá nás pouze jeden znak)

  - zkoumáme IQ 62 žáků 8. tříd v jisté škole
  - jak stručně popsat (zhodnotit), co mají data společného, nebo do jaké míry jsou odlišné?
  - z naměřených hodnot zkoumaného znaku spočítáme charakteristiky (míry) některých jeho hromadných vlastností (charakteristiky polohy, variability, tvaru rozdělení, u vícerozměrných dat to budou i charakteristiky závislosti)
  - charakteristiky (statistiky) jedním číslem vyjádří danou vlastnost

Příklad - jednorozměrný

- jednorozměrná data (zajímá nás pouze jeden znak)

  - zkoumáme IQ 62 žáků 8. tříd v jisté škole
  - jak stručně popsat (zhodnotit), co mají data společného, nebo do jaké míry jsou odlišné?
  - z naměřených hodnot zkoumaného znaku spočítáme charakteristiky (míry) některých jeho hromadných vlastností (charakteristiky polohy, variability, tvaru rozdělení, u vícerozměrných dat to budou i charakteristiky závislosti)
  - charakteristiky (statistiky) jedním číslem vyjádří danou vlastnost

Příklad - jednorozměrný

- jednorozměrná data (zajímá nás pouze jeden znak)

  - zkoumáme IQ 62 žáků 8. tříd v jisté škole
  - jak stručně popsat (zhodnotit), co mají data společného, nebo do jaké míry jsou odlišné?
  - z naměřených hodnot zkoumaného znaku spočítáme charakteristiky (míry) některých jeho hromadných vlastností (charakteristiky polohy, variability, tvaru rozdělení, u vícerozměrných dat to budou i charakteristiky závislosti)
  - charakteristiky (statistiky) jedním číslem vyjádří danou vlastnost

Příklad - naměřená data

naměřená data označme $x_1, x_2 \ldots, x_n$, nyní tedy $n = 62$.

107	141	105	111	112	96	103	140	136	92
92	72	123	140	112	127	120	106	117	92
107	108	117	141	109	109	106	113	112	119
138	109	80	111	86	111	120	96	103	112
104	103	125	101	132	113	108	106	97	121
134	84	108	84	129	116	107	112	128	133
96	94								

uspořádaný soubor označme $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

,

Příklad - naměřená data

naměřená data označme $x_1, x_2 \ldots, x_n$, nyní tedy $n = 62$.

107	141	105	111	112	96	103	140	136	92
92	72	123	140	112	127	120	106	117	92
107	108	117	141	109	109	106	113	112	119
138	109	80	111	86	111	120	96	103	112
104	103	125	101	132	113	108	106	97	121
134	84	108	84	129	116	107	112	128	133
96	94								

uspořádaný soubor označme $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

Třídní rozdělení četností

- Pokud se hodnoty často opakují, tak vytvoříme tzv. **četnostní tabulku**.
- Pokud jde o spojitou veličinu s velkým n (počtem naměřených hodnot), lze pro přehlednost obor hodnot dat rozdělit do M intervalů ohraničených body $a = a_0 < a_1 < a_2 < \ldots < a_{M-1} < a_M = b$.
- všechna pozorování z daného intervalu lze nahradit zástupnou hodnotou (zpravidla středem intervalu) x_i^*, $i = 1, \ldots, k$.
- nechť n_i označuje počet hodnot, které přísluší intervalu $\langle a_{i-1}, a_i)$, $i = 1, \ldots, M$ – tzv. **třídní (absolutní) četnost** (jednotlivé intervaly se nazývají **třídy**).
- **kumulativní četnost** N_i udává počet hodnot v dané (i-té) třídě a třídách předcházejících
- čísla n_i/n označují **relativní četnost**.

Třídní rozdělení četností

- Pokud se hodnoty často opakují, tak vytvoříme tzv. **četnostní tabulku**.
- Pokud jde o spojitou veličinu s velkým n (počtem naměřených hodnot), lze pro přehlednost obor hodnot dat rozdělit do M intervalů ohraničených body $a = a_0 < a_1 < a_2 < \ldots < a_{M-1} < a_M = b$.
- všechna pozorování z daného intervalu lze nahradit zástupnou hodnotou (zpravidla středem intervalu) x_i^*, $i = 1, \ldots, k$.
- nechť n_i označuje počet hodnot, které přísluší intervalu $\langle a_{i-1}, a_i)$, $i = 1, \ldots, M$ – tzv. **třídní (absolutní) četnost** (jednotlivé intervaly se nazývají **třídy**).
- **kumulativní četnost** N_i udává počet hodnot v dané (i-té) třídě a třídách předcházejících
- čísla n_i/n označují **relativní četnost**.

Třídní rozdělení četností

- Pokud se hodnoty často opakují, tak vytvoříme tzv. **četnostní tabulku**.
- Pokud jde o spojitou veličinu s velkým n (počtem naměřených hodnot), lze pro přehlednost obor hodnot dat rozdělit do M intervalů ohraničených body $a = a_0 < a_1 < a_2 < \ldots < a_{M-1} < a_M = b$.
- všechna pozorování z daného intervalu lze nahradit zástupnou hodnotou (zpravidla středem intervalu) x_i^*, $i = 1, \ldots, k$.
- nechť n_i označuje počet hodnot, které přísluší intervalu $\langle a_{i-1}, a_i)$, $i = 1, \ldots, M$ – tzv. **třídní (absolutní) četnost** (jednotlivé intervaly se nazývají **třídy**).
- **kumulativní četnost** N_i udává počet hodnot v dané (i-té) třídě a třídách předcházejících
- čísla n_i/n označují **relativní četnost**.

Třídní rozdělení četností

- Pokud se hodnoty často opakují, tak vytvoříme tzv. **četnostní tabulku**.
- Pokud jde o spojitou veličinu s velkým n (počtem naměřených hodnot), lze pro přehlednost obor hodnot dat rozdělit do M intervalů ohraničených body $a = a_0 < a_1 < a_2 < \ldots < a_{M-1} < a_M = b$.
- všechna pozorování z daného intervalu lze nahradit zástupnou hodnotou (zpravidla středem intervalu) x_i^*, $i = 1, \ldots, k$.
- nechť n_i označuje počet hodnot, které přísluší intervalu $\langle a_{i-1},\ a_i)$, $i = 1, \ldots, M$ – tzv. **třídní (absolutní) četnost** (jednotlivé intervaly se nazývají **třídy**).
- **kumulativní četnost** N_i udává počet hodnot v dané (i-té) třídě a třídách předcházejících
- čísla n_i/n označují **relativní četnost**.

Třídní rozdělení četností

- Pokud se hodnoty často opakují, tak vytvoříme tzv. **četnostní tabulku**.
- Pokud jde o spojitou veličinu s velkým n (počtem naměřených hodnot), lze pro přehlednost obor hodnot dat rozdělit do M intervalů ohraničených body $a = a_0 < a_1 < a_2 < \ldots < a_{M-1} < a_M = b$.
- všechna pozorování z daného intervalu lze nahradit zástupnou hodnotou (zpravidla středem intervalu) x_i^*, $i = 1, \ldots, k$.
- nechť n_i označuje počet hodnot, které přísluší intervalu $\langle a_{i-1},\ a_i)$, $i = 1, \ldots, M$ – tzv. **třídní (absolutní) četnost** (jednotlivé intervaly se nazývají **třídy**).
- **kumulativní četnost** N_i udává počet hodnot v dané (i-té) třídě a třídách předcházejících
- čísla n_i/n označují **relativní četnost**.

Třídní rozdělení četností

- Pokud se hodnoty často opakují, tak vytvoříme tzv. **četnostní tabulku**.
- Pokud jde o spojitou veličinu s velkým n (počtem naměřených hodnot), lze pro přehlednost obor hodnot dat rozdělit do M intervalů ohraničených body $a = a_0 < a_1 < a_2 < \ldots < a_{M-1} < a_M = b$.
- všechna pozorování z daného intervalu lze nahradit zástupnou hodnotou (zpravidla středem intervalu) x_i^*, $i = 1, \ldots, k$.
- nechť n_i označuje počet hodnot, které přísluší intervalu $\langle a_{i-1},\ a_i)$, $i = 1, \ldots, M$ – tzv. **třídní (absolutní) četnost** (jednotlivé intervaly se nazývají **třídy**).
- **kumulativní četnost** N_i udává počet hodnot v dané (i-té) třídě a třídách předcházejících
- čísla n_i/n označují **relativní četnost**.

Příklad - třídní rozdělení četností

Interval	x_i^*	absol. n_i	n_i/n	kumul. N_i	N_i/n
< 80	75	1	0.016	1	0.016
$\langle 80, 90)$	85	4	0.065	5	0.081
$\langle 90, 100)$	95	8	0.129	13	0.210
$\langle 100, 110)$	105	18	0.290	31	0.500
$\langle 110, 120)$	115	14	0.226	45	0.726
$\langle 120, 130)$	125	8	0.129	53	0.855
$\langle 130, 140)$	135	5	0.081	58	0.935
≥ 140	145	4	0.065	62	1.000

Histogram

- grafické znázornění třídních četností
- každému intervalu je přiřazen obdélníček tak, aby jeho plocha byla úměrná četnosti daného intervalu
- nejčastěji mají intervaly stejnou šířku (často vhodně zaokrouhlenou), pak výška obdélníků odpovídá četnostem.
- problém: volba počtu intervalů M
  lze použít např. tzv. Sturgesovo pravidlo:

$$M \approx 1 + 3.3 \log_{10}(n) \doteq 1 + \log_2(n)$$

- u našeho příkladu: $1 + \log_2(62) = 6.95$

Histogram

- grafické znázornění třídních četností
- každému intervalu je přiřazen obdélníček tak, aby jeho plocha byla úměrná četnosti daného intervalu
- nejčastěji mají intervaly stejnou šířku (často vhodně zaokrouhlenou), pak výška obdélníků odpovídá četnostem.
- problém: volba počtu intervalů M
  lze použít např. tzv. Sturgesovo pravidlo:

$$M \approx 1 + 3.3 \log_{10}(n) \doteq 1 + \log_2(n)$$

- u našeho příkladu: $1 + \log_2(62) = 6.95$

Histogram

- grafické znázornění třídních četností
- každému intervalu je přiřazen obdélníček tak, aby jeho plocha byla úměrná četnosti daného intervalu
- nejčastěji mají intervaly stejnou šířku (často vhodně zaokrouhlenou), pak výška obdélníků odpovídá četnostem.
- problém: volba počtu intervalů M
  lze použít např. tzv. Sturgesovo pravidlo:

$$M \approx 1 + 3.3 \log_{10}(n) \doteq 1 + \log_2(n)$$

- u našeho příkladu: $1 + \log_2(62) = 6.95$

Histogram

- grafické znázornění třídních četností
- každému intervalu je přiřazen obdélníček tak, aby jeho plocha byla úměrná četnosti daného intervalu
- nejčastěji mají intervaly stejnou šířku (často vhodně zaokrouhlenou), pak výška obdélníků odpovídá četnostem.
- problém: volba počtu intervalů M
  lze použít např. tzv. Sturgesovo pravidlo:

$$M \approx 1 + 3.3 \log_{10}(n) \doteq 1 + \log_2(n)$$

- u našeho příkladu: $1 + \log_2(62) = 6.95$

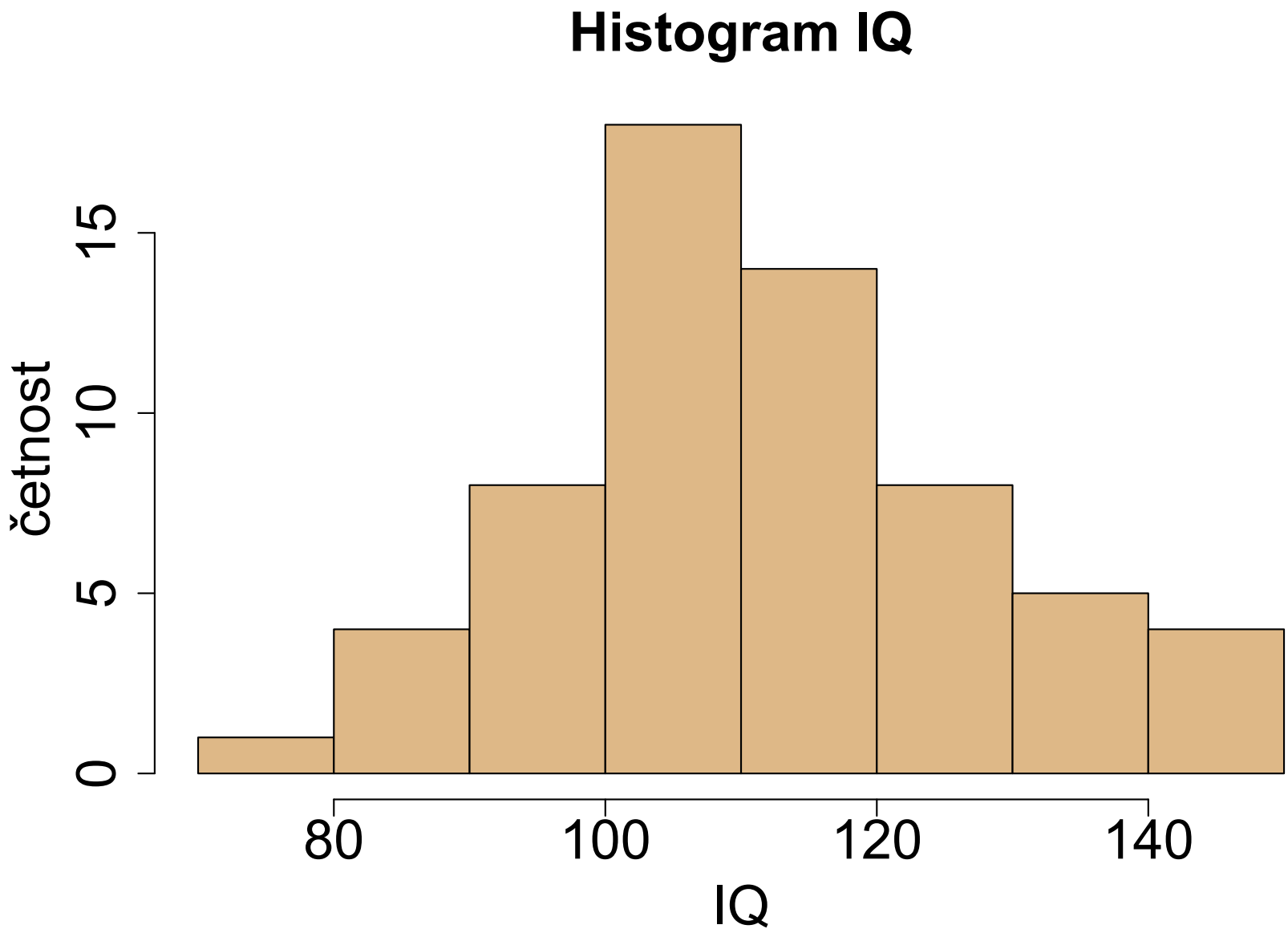
Histogram

- grafické znázornění třídních četností
- každému intervalu je přiřazen obdélníček tak, aby jeho plocha byla úměrná četnosti daného intervalu
- nejčastěji mají intervaly stejnou šířku (často vhodně zaokrouhlenou), pak výška obdélníků odpovídá četnostem.
- problém: volba počtu intervalů M
  lze použít např. tzv. Sturgesovo pravidlo:

$$M \approx 1 + 3.3 \log_{10}(n) \doteq 1 + \log_2(n)$$

- u našeho příkladu: $1 + \log_2(62) = 6.95$

Příklad - histogram



Charakteristiky polohy

- umožní charakterizovat úroveň číselné veličiny jedním číslem - ohodnocení, jak malých či velkých hodnot měření nabývají.
- pro charakteristiku polohy m souboru dat x by mělo platit, že se přirozeně mění se změnou měřítka, tj. že pro libovolné konstanty a, b:

$$m(a \cdot x + b) = a \cdot m(x) + b$$

- přičteme-li ke všem hodnotám konstantu b, tak se výsledná charakteristika zvětší o b
- vynásobíme-li každou hodnotu konstantou a, pak se výsledná charakteristika zvětší a-krát

Charakteristiky polohy

- umožní charakterizovat úroveň číselné veličiny jedním číslem - ohodnocení, jak malých či velkých hodnot měření nabývají.
- pro charakteristiku polohy m souboru dat x by mělo platit, že se přirozeně mění se změnou měřítka, tj. že pro libovolné konstanty a, b:

$$m(a \cdot x + b) = a \cdot m(x) + b$$

- přičteme-li ke všem hodnotám konstantu b, tak se výsledná charakteristika zvětší o b
- vynásobíme-li každou hodnotu konstantou a, pak se výsledná charakteristika zvětší a-krát

Aritmetický průměr

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

- u našeho příkladu: $\overline{x} = \frac{1}{62}(107 + 141 + \ldots + 94) = 111.0645$
- citlivý na hrubé chyby, odlehlá pozorování. Jen pro kvantitativní měřítka.
- z tabulky četností lze spočítat jako tzv. vážený průměr

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{M} n_i x_i^* = \frac{\sum_{i=1}^{M} n_i x_i^*}{\sum_{i=1}^{M} n_i} = \frac{1 \cdot 75 + 4 \cdot 85 + \ldots + 4 \cdot 145}{62} = 111.7742$$

- u nula-jedničkové veličiny: $\frac{\text{počet jedniček}}{\text{počet nul i jedniček}}$ = relativní četnost (procento) jedniček (pozorování s danou vlastností).
- u našeho příkladu $y_i = 0$ (i-tý žák je chlapec) , $y_i = 1$ (i-tý žák je dívka): $\overline{y} = \frac{32}{62} = 0.516$

Aritmetický průměr

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

- u našeho příkladu: $\overline{x} = \frac{1}{62}(107 + 141 + \ldots + 94) = 111.0645$
- citlivý na hrubé chyby, odlehlá pozorování. Jen pro kvantitativní měřítka.
- z tabulky četností lze spočítat jako tzv. vážený průměr

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{M} n_i x_i^* = \frac{\sum_{i=1}^{M} n_i x_i^*}{\sum_{i=1}^{M} n_i} = \frac{1 \cdot 75 + 4 \cdot 85 + \ldots + 4 \cdot 145}{62} = 111.7742$$

- u nula-jedničkové veličiny: $\frac{\text{počet jedniček}}{\text{počet nul i jedniček}}$ = relativní četnost (procento) jedniček (pozorování s danou vlastností).
- u našeho příkladu $y_i = 0$ (i-tý žák je chlapec) , $y_i = 1$ (i-tý žák je dívka): $\overline{y} = \frac{32}{62} = 0.516$

Aritmetický průměr

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

- u našeho příkladu: $\overline{x} = \frac{1}{62}(107 + 141 + \ldots + 94) = 111.0645$
- citlivý na hrubé chyby, odlehlá pozorování. Jen pro kvantitativní měřítka.
- z tabulky četností lze spočítat jako tzv. vážený průměr

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{M} n_i x_i^* = \frac{\sum_{i=1}^{M} n_i x_i^*}{\sum_{i=1}^{M} n_i} = \frac{1 \cdot 75 + 4 \cdot 85 + \ldots + 4 \cdot 145}{62} = 111.7742$$

- u nula-jedničkové veličiny: $\frac{\text{počet jedniček}}{\text{počet nul i jedniček}}$ = relativní četnost (procento) jedniček (pozorování s danou vlastností).
- u našeho příkladu $y_i = 0$ (i-tý žák je chlapec) , $y_i = 1$ (i-tý žák je dívka): $\overline{y} = \frac{32}{62} = 0.516$

Aritmetický průměr

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

- u našeho příkladu: $\overline{x} = \frac{1}{62}(107 + 141 + \ldots + 94) = 111.0645$
- citlivý na hrubé chyby, odlehlá pozorování. Jen pro kvantitativní měřítka.
- z tabulky četností lze spočítat jako tzv. vážený průměr

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{M} n_i x_i^* = \frac{\sum_{i=1}^{M} n_i x_i^*}{\sum_{i=1}^{M} n_i} = \frac{1 \cdot 75 + 4 \cdot 85 + \ldots + 4 \cdot 145}{62} = 111.7742$$

- u nula-jedničkové veličiny: $\frac{\text{počet jedniček}}{\text{počet nul i jedniček}}$ = relativní četnost (procento) jedniček (pozorování s danou vlastností).
- u našeho příkladu $y_i = 0$ (i-tý žák je chlapec) , $y_i = 1$ (i-tý žák je dívka): $\overline{y} = \frac{32}{62} = 0.516$

Aritmetický průměr

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

- u našeho příkladu: $\overline{x} = \frac{1}{62}(107 + 141 + \ldots + 94) = 111.0645$
- citlivý na hrubé chyby, odlehlá pozorování. Jen pro kvantitativní měřítka.
- z tabulky četností lze spočítat jako tzv. vážený průměr

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{M} n_i x_i^* = \frac{\sum_{i=1}^{M} n_i x_i^*}{\sum_{i=1}^{M} n_i} = \frac{1 \cdot 75 + 4 \cdot 85 + \ldots + 4 \cdot 145}{62} = 111.7742$$

- u nula-jedničkové veličiny: $\frac{\text{počet jedniček}}{\text{počet nul i jedniček}}$ = relativní četnost (procento) jedniček (pozorování s danou vlastností).
- u našeho příkladu $y_i = 0$ (i-tý žák je chlapec) , $y_i = 1$ (i-tý žák je dívka): $\overline{y} = \frac{32}{62} = 0.516$

Aritmetický průměr

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

- u našeho příkladu: $\overline{x} = \frac{1}{62}(107 + 141 + \ldots + 94) = 111.0645$
- citlivý na hrubé chyby, odlehlá pozorování. Jen pro kvantitativní měřítka.
- z tabulky četností lze spočítat jako tzv. vážený průměr

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{M} n_i x_i^* = \frac{\sum_{i=1}^{M} n_i x_i^*}{\sum_{i=1}^{M} n_i} = \frac{1 \cdot 75 + 4 \cdot 85 + \ldots + 4 \cdot 145}{62} = 111.7742$$

- u nula-jedničkové veličiny: $\frac{\text{počet jedniček}}{\text{počet nul i jedniček}}$ = relativní četnost (procento) jedniček (pozorování s danou vlastností).
- u našeho příkladu $y_i = 0$ (i-tý žák je chlapec) , $y_i = 1$ (i-tý žák je dívka): $\overline{y} = \frac{32}{62} = 0.516$

Modus

- $\hat{x}$ - nejčastější hodnota
- má smysl určovat i pro nominální a ordinální měřítko
- není vždy jednoznačně určen
- u našeho příkladu:

Modus

- $\hat{x}$ - nejčastější hodnota
- má smysl určovat i pro nominální a ordinální měřítko
- není vždy jednoznačně určen
- u našeho příkladu:

Modus

- $\hat{x}$ - nejčastější hodnota
- má smysl určovat i pro nominální a ordinální měřítko
- není vždy jednoznačně určen
- u našeho příkladu:

Modus

- $\hat{x}$ - nejčastější hodnota
- má smysl určovat i pro nominální a ordinální měřítko
- není vždy jednoznačně určen
- u našeho příkladu:

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

Modus

- $\hat{x}$ - nejčastější hodnota
- má smysl určovat i pro nominální a ordinální měřítko
- není vždy jednoznačně určen
- u našeho příkladu:

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

$$\hat{x} = 112$$

Medián

- $\tilde{x}$ - číslo, které dělí uspořádaný soubor na dvě stejně velké části. V uspořádaném výběru je uprostřed.

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

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

- robustní - není ovlivněn i velkými změnami několika hodnot. Lze často už i pro ordinální měřítko. U našeho příkladu:

Medián

- $\tilde{x}$ - číslo, které dělí uspořádaný soubor na dvě stejně velké části. V uspořádaném výběru je uprostřed.

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

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

- robustní - není ovlivněn i velkými změnami několika hodnot. Lze často už i pro ordinální měřítko. U našeho příkladu:

Medián

- $\tilde{x}$ - číslo, které dělí uspořádaný soubor na dvě stejně velké části. V uspořádaném výběru je uprostřed.

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

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

- robustní - není ovlivněn i velkými změnami několika hodnot. Lze často už i pro ordinální měřítko. U našeho příkladu:

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

Medián

- $\tilde{x}$ - číslo, které dělí uspořádaný soubor na dvě stejně velké části. V uspořádaném výběru je uprostřed.

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

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

- robustní - není ovlivněn i velkými změnami několika hodnot. Lze často už i pro ordinální měřítko. U našeho příkladu:

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

$$\tilde{x} = \frac{1}{2}\left(x_{(31)} + x_{(32)}\right) = 110$$

Kvantily: percentily, decily, kvartily

α-kvantil x_α ($\alpha \in (0,1)$) - dělí uspořádaný soubor na dvě části tak, že právě α-podíl těch nejmenších hodnot je menších než x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  kde $\lceil a \rceil$ označuje a, pokud je to celé číslo, jinak nejbližší vyšší celé číslo.
- speciální případy kvantilů:

  percentily: $\alpha = 0.01, 0.02, \ldots, 0.99$
  decily: $\alpha = 0.1, 0.2, \ldots, 0.9$
  kvartily: $\alpha = 0.25, 0.5, 0.75$

  1. (dolní) kvartil značíme $Q_1 = x_{0.25}$
  3. (horní) kvartil značíme $Q_3 = x_{0.75}$
- medián je vlastně 50%-ní kvantil, 50-tý percentil, 5-tý decil a 2-hý kvartil

Kvantily: percentily, decily, kvartily

α-kvantil x_α ($\alpha \in (0,1)$) - dělí uspořádaný soubor na dvě části tak, že právě α-podíl těch nejmenších hodnot je menších než x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  kde $\lceil a \rceil$ označuje a, pokud je to celé číslo, jinak nejbližší vyšší celé číslo.
- speciální případy kvantilů:

  percentily: $\alpha = 0.01, 0.02, \ldots, 0.99$
  decily: $\alpha = 0.1, 0.2, \ldots, 0.9$
  kvartily: $\alpha = 0.25, 0.5, 0.75$

  1. (dolní) kvartil značíme $Q_1 = x_{0.25}$
  3. (horní) kvartil značíme $Q_3 = x_{0.75}$
- medián je vlastně 50%-ní kvantil, 50-tý percentil, 5-tý decil a 2-hý kvartil

Kvantily: percentily, decily, kvartily

α-kvantil x_α ($\alpha \in (0,1)$) - dělí uspořádaný soubor na dvě části tak, že právě α-podíl těch nejmenších hodnot je menších než x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  kde $\lceil a \rceil$ označuje a, pokud je to celé číslo, jinak nejbližší vyšší celé číslo.
- speciální případy kvantilů:

percentily: $\alpha = 0.01, 0.02, \ldots, 0.99$
decily: $\alpha = 0.1, 0.2, \ldots, 0.9$
kvartily: $\alpha = 0.25, 0.5, 0.75$

1. (dolní) kvartil značíme $Q_1 = x_{0.25}$
3. (horní) kvartil značíme $Q_3 = x_{0.75}$

- medián je vlastně 50%-ní kvantil, 50-tý percentil, 5-tý decil a 2-hý kvartil

Kvantily: percentily, decily, kvartily

α-kvantil x_α ($\alpha \in (0,1)$) - dělí uspořádaný soubor na dvě části tak, že právě α-podíl těch nejmenších hodnot je menších než x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  kde $\lceil a \rceil$ označuje a, pokud je to celé číslo, jinak nejbližší vyšší celé číslo.
- speciální případy kvantilů:

percentily: $\alpha = 0.01, 0.02, \ldots, 0.99$

decily: $\alpha = 0.1, 0.2, \ldots, 0.9$

kvartily: $\alpha = 0.25, 0.5, 0.75$

1. (dolní) kvartil značíme $Q_1 = x_{0.25}$

3. (horní) kvartil značíme $Q_3 = x_{0.75}$

- medián je vlastně 50%-ní kvantil, 50-tý percentil, 5-tý decil a 2-hý kvartil

Kvantily: percentily, decily, kvartily

α-kvantil x_α ($\alpha \in (0,1)$) - dělí uspořádaný soubor na dvě části tak, že právě α-podíl těch nejmenších hodnot je menších než x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  kde $\lceil a \rceil$ označuje a, pokud je to celé číslo, jinak nejbližší vyšší celé číslo.
- speciální případy kvantilů:

percentily: $\alpha = 0.01, 0.02, \ldots, 0.99$

decily: $\alpha = 0.1, 0.2, \ldots, 0.9$

kvartily: $\alpha = 0.25, 0.5, 0.75$

1. (dolní) kvartil značíme $Q_1 = x_{0.25}$

3. (horní) kvartil značíme $Q_3 = x_{0.75}$

- medián je vlastně 50%-ní kvantil, 50-tý percentil, 5-tý decil a 2-hý kvartil

Kvantily: percentily, decily, kvartily

α-kvantil x_α ($\alpha \in (0,1)$) - dělí uspořádaný soubor na dvě části tak, že právě α-podíl těch nejmenších hodnot je menších než x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  kde $\lceil a \rceil$ označuje a, pokud je to celé číslo, jinak nejbližší vyšší celé číslo.
- speciální případy kvantilů:

$$\textbf{percentily}: \alpha = 0.01, 0.02, \ldots, 0.99$$
$$\textbf{decily}: \alpha = 0.1, 0.2, \ldots, 0.9$$
$$\textbf{kvartily}: \alpha = 0.25, 0.5, 0.75$$

1. **(dolní) kvartil** značíme $Q_1 = x_{0.25}$
3. **(horní) kvartil** značíme $Q_3 = x_{0.75}$

- medián je vlastně 50%-ní kvantil, 50-tý percentil, 5-tý decil a 2-hý kvartil

Kvantily: percentily, decily, kvartily

α-kvantil x_α ($\alpha \in (0,1)$) - dělí uspořádaný soubor na dvě části tak, že právě α-podíl těch nejmenších hodnot je menších než x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  kde $\lceil a \rceil$ označuje a, pokud je to celé číslo, jinak nejbližší vyšší celé číslo.
- speciální případy kvantilů:

$$\textbf{percentily}: \alpha = 0.01, 0.02, \ldots, 0.99$$
$$\textbf{decily}: \alpha = 0.1, 0.2, \ldots, 0.9$$
$$\textbf{kvartily}: \alpha = 0.25, 0.5, 0.75$$

1. (dolní) kvartil značíme $Q_1 = x_{0.25}$
3. (horní) kvartil značíme $Q_3 = x_{0.75}$

- medián je vlastně 50%-ní kvantil, 50-tý percentil, 5-tý decil a 2-hý kvartil

Kvantily: percentily, decily, kvartily

α-kvantil x_α ($\alpha \in (0,1)$) - dělí uspořádaný soubor na dvě části tak, že právě α-podíl těch nejmenších hodnot je menších než x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  kde $\lceil a \rceil$ označuje a, pokud je to celé číslo, jinak nejbližší vyšší celé číslo.
- speciální případy kvantilů:

  percentily: $\alpha = 0.01, 0.02, \ldots, 0.99$
  decily: $\alpha = 0.1, 0.2, \ldots, 0.9$
  kvartily: $\alpha = 0.25, 0.5, 0.75$

  1. (dolní) kvartil značíme $Q_1 = x_{0.25}$
  3. (horní) kvartil značíme $Q_3 = x_{0.75}$
- medián je vlastně 50%-ní kvantil, 50-tý percentil, 5-tý decil a 2-hý kvartil

Příklad - kvantily

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

- 1. kvartil $Q_1 = x_{0.25} = x_{(\lceil 0.25 \cdot 62 \rceil)} = x_{(\lceil 15.5 \rceil)} = x_{(16)} = 103$
- 3. kvartil $Q_3 = x_{0.75} = x_{(\lceil 0.75 \cdot 62 \rceil)} = x_{(\lceil 46.5 \rceil)} = x_{(47)} = 120$
- 1. decil (10%-ní kvantil)

$$x_{0.1} = x_{(\lceil 0.1 \cdot 62 \rceil)} = x_{(\lceil 6.2 \rceil)} = x_{(7)} = 92$$

- 9. decil (90%-ní kvantil)

$$x_{0.9} = x_{(\lceil 0.9 \cdot 62 \rceil)} = x_{(\lceil 55.8 \rceil)} = x_{(56)} = 134$$

Příklad - kvantily

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

- 1. kvartil $Q_1 = x_{0.25} = x_{(\lceil 0.25 \cdot 62 \rceil)} = x_{(\lceil 15.5 \rceil)} = x_{(16)} = 103$
- 3. kvartil $Q_3 = x_{0.75} = x_{(\lceil 0.75 \cdot 62 \rceil)} = x_{(\lceil 46.5 \rceil)} = x_{(47)} = 120$
- 1. decil (10%-ní kvantil)

$$x_{0.1} = x_{(\lceil 0.1 \cdot 62 \rceil)} = x_{(\lceil 6.2 \rceil)} = x_{(7)} = 92$$

- 9. decil (90%-ní kvantil)

$$x_{0.9} = x_{(\lceil 0.9 \cdot 62 \rceil)} = x_{(\lceil 55.8 \rceil)} = x_{(56)} = 134$$

Příklad - kvantily

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

- 1. kvartil $Q_1 = x_{0.25} = x_{(\lceil 0.25 \cdot 62 \rceil)} = x_{(\lceil 15.5 \rceil)} = x_{(16)} = 103$
- 3. kvartil $Q_3 = x_{0.75} = x_{(\lceil 0.75 \cdot 62 \rceil)} = x_{(\lceil 46.5 \rceil)} = x_{(47)} = 120$
- 1. decil (10%-ní kvantil)

$$x_{0.1} = x_{(\lceil 0.1 \cdot 62 \rceil)} = x_{(\lceil 6.2 \rceil)} = x_{(7)} = 92$$

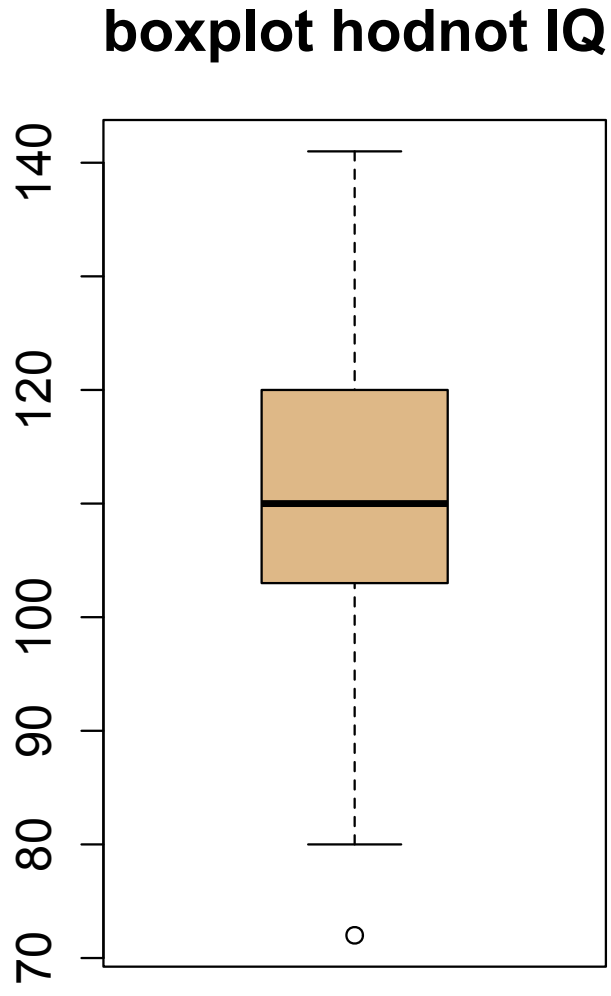
- 9. decil (90%-ní kvantil)

$$x_{0.9} = x_{(\lceil 0.9 \cdot 62 \rceil)} = x_{(\lceil 55.8 \rceil)} = x_{(56)} = 134$$

Boxplot

- česky **krabičkový diagram** - zobrazuje kvartily, medián, minimum, maximum a případně odlehlá pozorování (jsou od bližšího kvartilu dále než $1.5 \cdot (Q_3 - Q_1)$)

- u našeho příkladu: $Q_1 = 103, \tilde{x} = 110,$ $Q_3 = 120$, 72 jako odlehlé pozorování



Charakteristiky variability

- měří rozptýlení, proměnlivost, nestejnost, variabilitu souboru dat.
- pro charakteristiku variability s souboru dat x by mělo platit, že pro libovolnou konstantu b a pro libovolnou kladnou konstantu $a > 0$:

$$s(a \cdot x + b) = a \cdot s(x)$$

- přičteme-li ke všem hodnotám konstantu b, tak se výsledná charakteristika nezmění
- vynásobíme-li každou hodnotu konstantou a, pak se výsledná charakteristika zvětší a-krát

Charakteristiky variability

- měří rozptýlení, proměnlivost, nestejnost, variabilitu souboru dat.
- pro charakteristiku variability s souboru dat x by mělo platit, že pro libovolnou konstantu b a pro libovolnou kladnou konstantu $a > 0$:

$$s(a \cdot x + b) = a \cdot s(x)$$

- přičteme-li ke všem hodnotám konstantu b, tak se výsledná charakteristika nezmění
- vynásobíme-li každou hodnotu konstantou a, pak se výsledná charakteristika zvětší a-krát

Rozptyl (variance)

(populační) **rozptyl** $s_x^2 = var(x)$ - střední kvadratická odchylka od průměru

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

- u našeho příkladu:

$$s_x^2 = \frac{1}{62}\left[(107 - 111.0645)^2 + \ldots + (94 - 111.0645)^2\right] = 246.4797$$

- z tabulky četností:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{M} n_i(x_i^* - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{M} n_i x_i^{*2}\right) - \overline{x}^2$$

- pro rozptyl platí $s^2_{a\cdot x+b} = a^2 s_x^2$

Rozptyl (variance)

(populační) **rozptyl** $s_x^2 = var(x)$ - střední kvadratická odchylka od průměru

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

- u našeho příkladu:

$$s_x^2 = \frac{1}{62}\left[(107 - 111.0645)^2 + \ldots + (94 - 111.0645)^2\right] = 246.4797$$

- z tabulky četností:
  $s_x^2 = \frac{1}{n}\sum_{i=1}^{M} n_i(x_i^* - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{M} n_i x_i^{*2}\right) - \overline{x}^2$
- pro rozptyl platí $s_{a\cdot x+b}^2 = a^2 s_x^2$

Rozptyl (variance)

(populační) **rozptyl** $s_x^2 = var(x)$ - střední kvadratická odchylka od průměru

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

- u našeho příkladu:

$$s_x^2 = \frac{1}{62}\left[(107 - 111.0645)^2 + \ldots + (94 - 111.0645)^2\right] = 246.4797$$

- z tabulky četností:
  $s_x^2 = \frac{1}{n}\sum_{i=1}^{M} n_i(x_i^* - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{M} n_i x_i^{*2}\right) - \overline{x}^2$
- pro rozptyl platí $s_{a\cdot x+b}^2 = a^2 s_x^2$

Rozptyl (variance)

(populační) **rozptyl** $s_x^2 = var(x)$ - střední kvadratická odchylka od průměru

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

- u našeho příkladu:

$$s_x^2 = \frac{1}{62}\left[(107 - 111.0645)^2 + \ldots + (94 - 111.0645)^2\right] = 246.4797$$

- z naší tabulky četností:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{M} n_i(x_i^* - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{M} n_i x_i^{*2}\right) - \overline{x}^2$$
$$= (1 \cdot 75^2 + \ldots + 4 \cdot 145^2) - 111.7742^2 = 257.3361$$

- pro rozptyl platí $s^2_{a\cdot x+b} = a^2 s_x^2$

Rozptyl (variance)

(populační) **rozptyl** $s_x^2 = var(x)$ - střední kvadratická odchylka od průměru

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

- u našeho příkladu:

$$s_x^2 = \frac{1}{62}\left[(107 - 111.0645)^2 + \ldots + (94 - 111.0645)^2\right] = 246.4797$$

- z naší tabulky četností:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{M} n_i(x_i^* - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{M} n_i x_i^{*2}\right) - \overline{x}^2$$
$$= (1 \cdot 75^2 + \ldots + 4 \cdot 145^2) - 111.7742^2 = 257.3361$$

- pro rozptyl platí $s_{a\cdot x+b}^2 = a^2 s_x^2$

Směrodatná odchylka, variační koeficient

(nevýběrová) **směrodatná odchylka**: odmocnina z rozptylu

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

- stejný fyzikální rozměr jako původní data

variační koeficient:

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

- definován pouze pro kladné hodnoty $x_1, \ldots, x_n > 0$
- nezávisí na volbě měřítka, lze použít na porovnání různých souborů

u našich dat: $s_x = \sqrt{246.4797} = 15.70$

$$v = \frac{15.70}{111.0645} = 0.1414$$

Směrodatná odchylka, variační koeficient

(nevýběrová) **směrodatná odchylka**: odmocnina z rozptylu

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

- stejný fyzikální rozměr jako původní data

variační koeficient:

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

- definován pouze pro kladné hodnoty $x_1, \ldots, x_n > 0$
- nezávisí na volbě měřítka, lze použít na porovnání různých souborů

u našich dat: $s_x = \sqrt{246.4797} = 15.70$

$$v = \frac{15.70}{111.0645} = 0.1414$$

Směrodatná odchylka, variační koeficient

(nevýběrová) **směrodatná odchylka**: odmocnina z rozptylu

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

- stejný fyzikální rozměr jako původní data

variační koeficient:

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

- definován pouze pro kladné hodnoty $x_1, \ldots, x_n > 0$
- nezávisí na volbě měřítka, lze použít na porovnání různých souborů

u našich dat: $s_x = \sqrt{246.4797} = 15.70$

$$v = \frac{15.70}{111.0645} = 0.1414$$

Směrodatná odchylka, variační koeficient

(nevýběrová) **směrodatná odchylka**: odmocnina z rozptylu

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

- stejný fyzikální rozměr jako původní data

variační koeficient:

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

- definován pouze pro kladné hodnoty $x_1, \ldots, x_n > 0$
- nezávisí na volbě měřítka, lze použít na porovnání různých souborů

u našich dat: $s_x = \sqrt{246.4797} = 15.70$

$$v = \frac{15.70}{111.0645} = 0.1414$$

Směrodatná odchylka, variační koeficient

(nevýběrová) **směrodatná odchylka**: odmocnina z rozptylu

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

- stejný fyzikální rozměr jako původní data

variační koeficient:

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

- definován pouze pro kladné hodnoty $x_1, \ldots, x_n > 0$
- nezávisí na volbě měřítka, lze použít na porovnání různých souborů

u našich dat: $s_x = \sqrt{246.4797} = 15.70$

$$v = \frac{15.70}{111.0645} = 0.1414$$

rozpětí: rozdíl maxima a minima souboru

$$R = x_{(n)} - x_{(1)}$$

mezikvartilové rozpětí: rozdíl třetího a prvního kvartilu

$$R_M = Q_3 - Q_1 = x_{0.75} - x_{0.25}$$

střední odchylka: průměr absolutních odchylek od mediánu

(nebo průměru)

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

u našich dat: $R = 141 - 72 = 69$ $R_M = 120 - 103 = 17$

$$d = \frac{1}{62}(|107 - 110| + \ldots + |94 - 110|) = 12.03$$

rozpětí: rozdíl maxima a minima souboru

$$R = x_{(n)} - x_{(1)}$$

mezikvartilové rozpětí: rozdíl třetího a prvního kvartilu

$$R_M = Q_3 - Q_1 = x_{0.75} - x_{0.25}$$

střední odchylka: průměr absolutních odchylek od mediánu

(nebo průměru)

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

u našich dat: $R = 141 - 72 = 69$ $R_M = 120 - 103 = 17$

$$d = \frac{1}{62}(|107 - 110| + \ldots + |94 - 110|) = 12.03$$

rozpětí: rozdíl maxima a minima souboru

$$R = x_{(n)} - x_{(1)}$$

mezikvartilové rozpětí: rozdíl třetího a prvního kvartilu

$$R_M = Q_3 - Q_1 = x_{0.75} - x_{0.25}$$

střední odchylka: průměr absolutních odchylek od mediánu

(nebo průměru)

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

u našich dat: $R = 141 - 72 = 69$ $R_M = 120 - 103 = 17$

$$d = \frac{1}{62}(|107 - 110| + \ldots + |94 - 110|) = 12.03$$

rozpětí: rozdíl maxima a minima souboru

$$R = x_{(n)} - x_{(1)}$$

mezikvartilové rozpětí: rozdíl třetího a prvního kvartilu

$$R_M = Q_3 - Q_1 = x_{0.75} - x_{0.25}$$

střední odchylka: průměr absolutních odchylek od mediánu

(nebo průměru)

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

u našich dat: $R = 141 - 72 = 69$ $R_M = 120 - 103 = 17$

$$d = \frac{1}{62}(|107 - 110| + \ldots + |94 - 110|) = 12.03$$

Míra variability pro kategoriální znak

Entropie

$$h = -\sum_{i=1}^{r} \frac{n_i}{n} \log\left(\frac{n_i}{n}\right)$$

Příklad - vícerozměrný

- vícerozměrná data (zajímá nás více znaků)

  - zjištěno IQ, pohlaví, průměrná známka v pololetí v 7. a 8. třídě 62 žáků
  - jak zhodnotit vztah (závislost) mezi jednotlivými znaky?
  - vypočtením vhodné statistiky (čísla) nebo grafickým zobrazením

Příklad - vícerozměrný

- vícerozměrná data (zajímá nás více znaků)

  - zjištěno IQ, pohlaví, průměrná známka v pololetí v 7. a 8. třídě 62 žáků
  - jak zhodnotit vztah (závislost) mezi jednotlivými znaky?
  - vypočtením vhodné statistiky (čísla) nebo grafickým zobrazením

Příklad - vícerozměrný

- vícerozměrná data (zajímá nás více znaků)

- zjištěno IQ, pohlaví, průměrná známka v pololetí v 7. a 8. třídě 62 žáků
- jak zhodnotit vztah (závislost) mezi jednotlivými znaky?
- vypočtením vhodné statistiky (čísla) nebo grafickým zobrazením

Příklad - naměřená vícerozměrná data

Dívka	1	0	0	1	0	1	0	0	1	1
Zn7	1	1	3.15	1.62	2.69	1.92	2.38	1	1.4	1.46
Zn8	1	1	3	1.73	2.09	2.09	2.55	1	1.9	1.45
IQ	107	141	105	111	112	96	103	140	136	92

Dívka	1	0	0	0	1	0	1	1	1	0
Zn7	1.85	3.15	1.15	1	1.69	1.6	1.62	1.38	1.7	3.23
Zn8	1.45	3.18	1.18	1	1.91	1.72	1.63	1.36	1.9	3.36
IQ	92	72	123	140	112	127	120	106	117	92

Dívka	0	0	1	1	1	1	0	1	0	1
Zn7	2.07	1.84	1.2	1.31	1.4	1.53	1.84	1	1.3	1.4
Zn8	2.45	1.9	1.36	1.45	1.73	1.6	1.54	1	1.45	1.82
IQ	107	108	117	141	109	109	106	113	112	119

Dívka	0	0	1	1	0	1	0	1	0	0
Zn7	1	2.92	2.23	1.69	2.61	1.07	1.46	2.15	1.69	1.38
Zn8	1	2.82	2.45	1.54	2.54	1	1.36	1.9	1.82	1.18
IQ	138	109	80	111	86	111	120	96	103	112

,

vícerozměrná data - pokračování

Dívka	1	1	1	0	0	1	0	1	1	0
Zn7	1.46	1.6	1.07	1.3	2.08	2	1.69	1.4	2.23	1.6
Zn8	1.54	1.63	1	1.27	1.54	2.09	1.91	1.45	2	1.81
IQ	104	103	125	101	132	113	108	106	97	121

Dívka	1	0	1	1	0	1	0	1	1	0
Zn7	1.07	3.13	1.84	1.8	1	1.92	2.2	1.53	1.3	1
Zn8	1.27	3.27	1.82	1.63	1	1.9	2.25	1.54	1.45	1.18
IQ	134	84	108	84	129	116	107	112	128	133

Dívka	0	0
Zn7	2.85	2.61
Zn8	2.91	2.81
IQ	96	94

,

Grafické znároznění závislosti

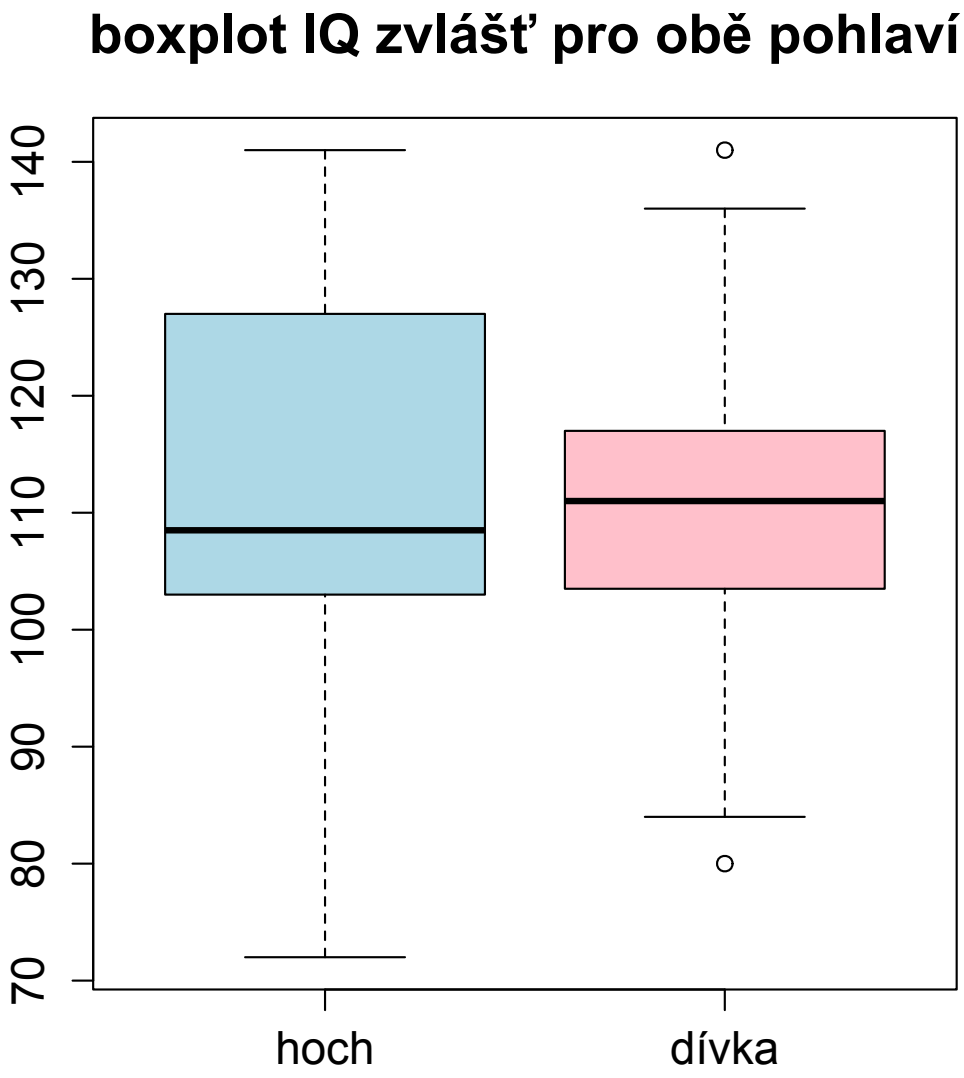
- Záleží na typu měřítka
- pro závislost kvantit. znaku na kvalitativním lze nakreslit boxplot/histogram pro každou kategorii kvalit. znaku
- zobrazení závislosti IQ na pohlaví
- $\overline{x}_{hoch} = 112.0$
  $\overline{x}_{divka} = 110.2$

Grafické znázornění závislosti

- Záleží na typu měřítka
- pro závislost kvantit. znaku na kvalitativním lze nakreslit boxplot/histogram pro každou kategorii kvalit. znaku
- zobrazení závislosti IQ na pohlaví
- $\overline{x}_{hoch} = 112.0$
  $\overline{x}_{divka} = 110.2$

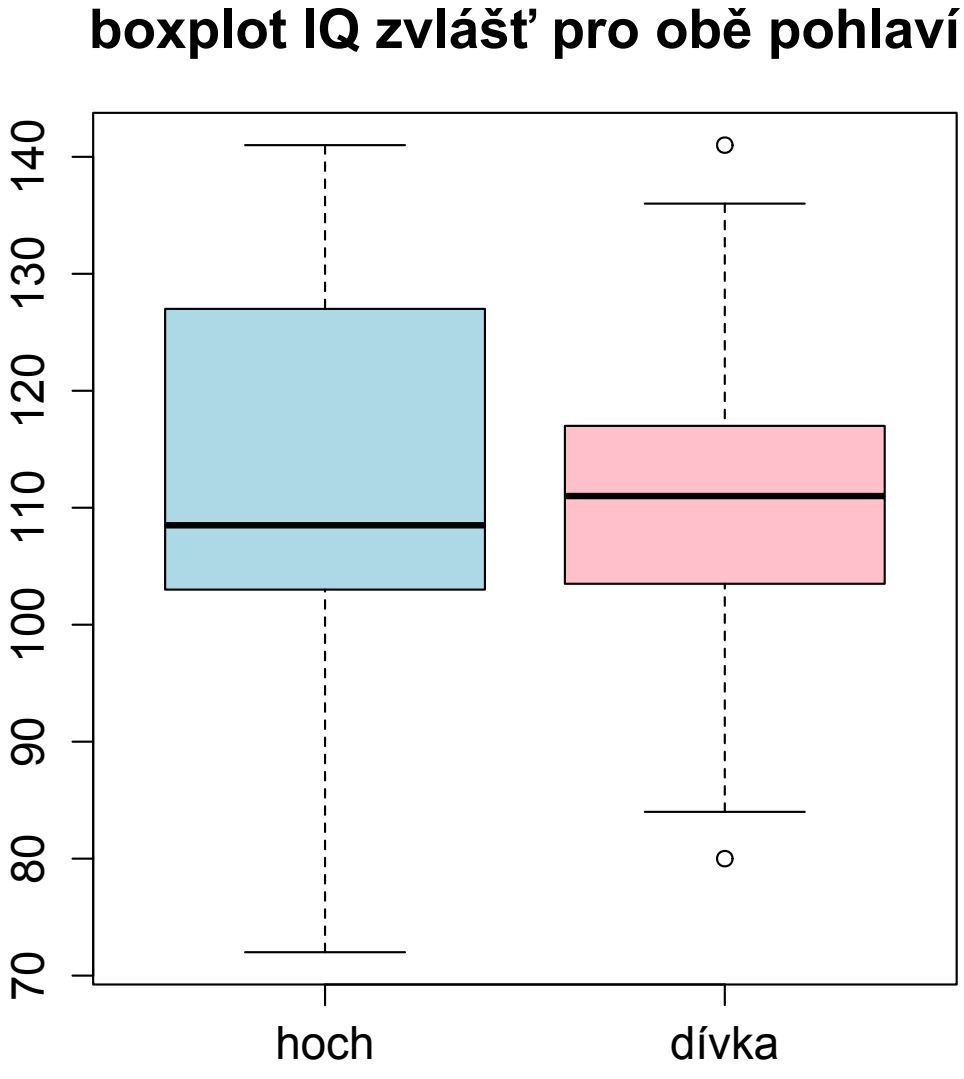
Grafické znázornění závislosti

- Záleží na typu měřítka
- pro závislost kvantit. znaku na kvalitativním lze nakreslit boxplot/histogram pro každou kategorii kvalit. znaku
- zobrazení závislosti IQ na pohlaví
- $\overline{x}_{hoch} = 112.0$
  $\overline{x}_{divka} = 110.2$



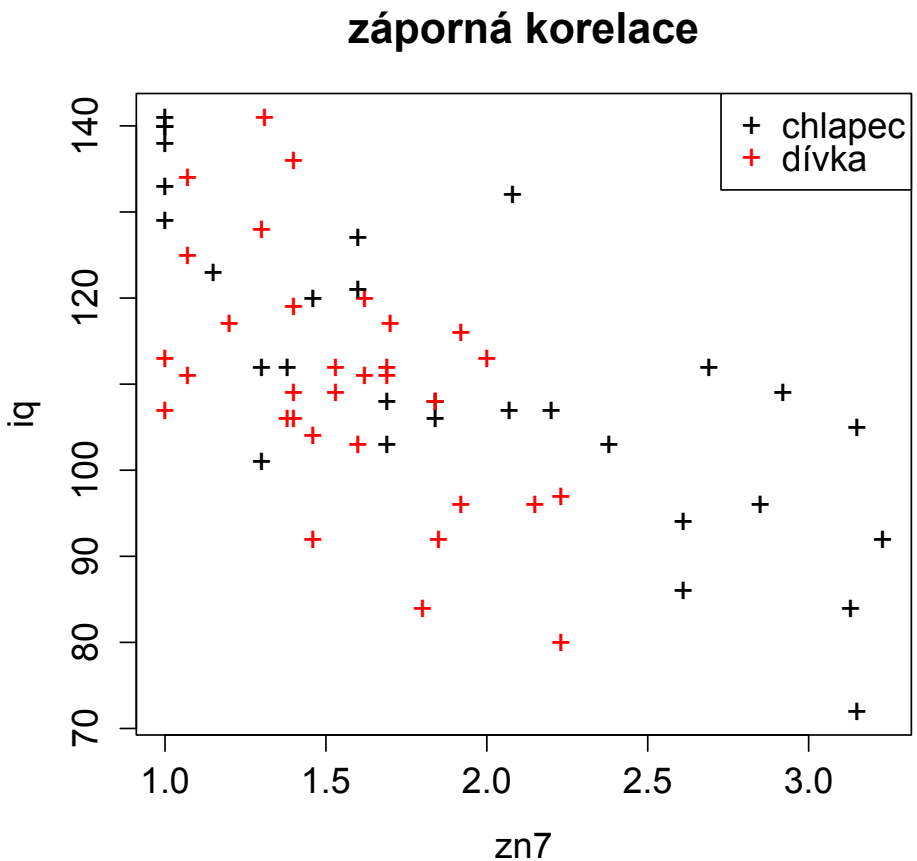
Grafické znázornění závislosti

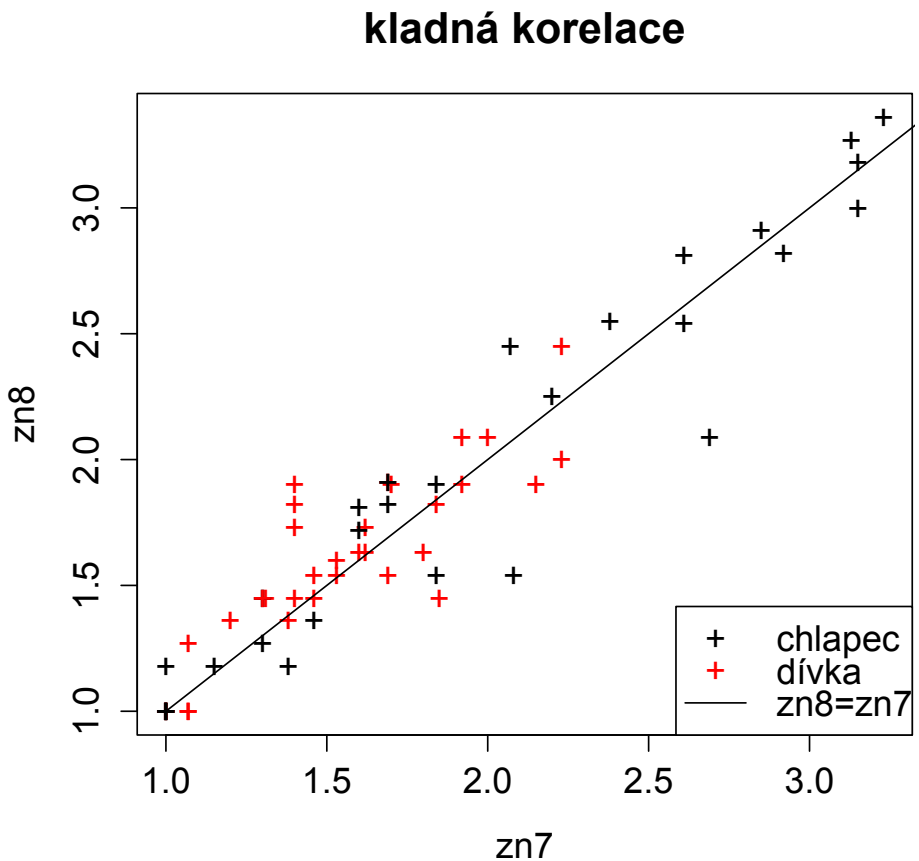
- Záleží na typu měřítka
- pro závislost kvantit. znaku na kvalitativním lze nakreslit boxplot/histogram pro každou kategorii kvalit. znaku
- zobrazení závislosti IQ na pohlaví
- $\overline{x}_{hoch} = 112.0$
  $\overline{x}_{divka} = 110.2$



Grafické znároznění závislosti - 2

Rozptylový diagram: závislost dvou kvantitativních znaků





,

Charakteristiky závislosti

dva znaky na každé jednotce, tj. máme $(x_1, y_1), \ldots, (x_n, y_n)$

kovariance: měří směr závislosti, ovlivněna změnou měřítka

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\cdot\overline{y},$$

- Platí $s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s_x^2, \quad s_{yy} = s_y^2$

(Pearsonův) korelační koeficient: normovaná kovariance, měří směr i velikost závislosti

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

- u našich dat pro znaky IQ a zn7:

$$r_{IQ,zn7} = \frac{-6.2876}{15.6997 \cdot 0.6106} = -0.6559$$

Charakteristiky závislosti

dva znaky na každé jednotce, tj. máme $(x_1, y_1), \ldots, (x_n, y_n)$

kovariance: měří směr závislosti, ovlivněna změnou měřítka

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\cdot\overline{y},$$

- Platí $s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s_x^2, \quad s_{yy} = s_y^2$

(Pearsonův) korelační koeficient: normovaná kovariance, měří směr i velikost závislosti

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

- u našich dat pro znaky IQ a zn7:

$$r_{IQ,zn7} = \frac{-6.2876}{15.6997 \cdot 0.6106} = -0.6559$$

Charakteristiky závislosti

dva znaky na každé jednotce, tj. máme $(x_1, y_1), \ldots, (x_n, y_n)$

kovariance: měří směr závislosti, ovlivněna změnou měřítka

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\cdot\overline{y},$$

- Platí $s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s_x^2, \quad s_{yy} = s_y^2$

(Pearsonův) korelační koeficient: normovaná kovariance, měří směr i velikost závislosti

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

- u našich dat pro znaky IQ a zn7:

$$r_{IQ,zn7} = \frac{-6.2876}{15.6997 \cdot 0.6106} = -0.6559$$

Charakteristiky závislosti

dva znaky na každé jednotce, tj. máme $(x_1, y_1), \ldots, (x_n, y_n)$

kovariance: měří směr závislosti, ovlivněna změnou měřítka

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\cdot\overline{y},$$

- Platí $s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s_x^2, \quad s_{yy} = s_y^2$

(Pearsonův) korelační koeficient: normovaná kovariance, měří směr i velikost závislosti

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

- u našich dat pro znaky IQ a zn7:

$$r_{IQ,zn7} = \frac{-6.2876}{15.6997 \cdot 0.6106} = -0.6559$$

Korelační koeficient

- měří směr a míru lineární závislosti
- nabývá jen hodnot z intervalu $\langle -1, 1 \rangle$
- $r_{x,y} \approx 0$ (znaky x a y vzájemně nezávislé)
- $r_{x,y}$ blízko 1 (kladná závislost: s rostoucím x znak y v průměru roste)
- $r_{x,y}$ blízko -1 (záporná závislost: s rostoucím x znak y v průměru klesá)

U našich dat lze spočítat pro každou dvojici znaků dívka, iq, zn7, zn8: tzv. **korelační matice**

	dívka	iq	zn7	zn8
dívka	1.0000	-0.0597	-0.3054	-0.2661
iq	-0.0597	1.0000	-0.6559	-0.6236
zn7	-0.3054	-0.6559	1.0000	0.9481
zn8	-0.2661	-0.6236	0.9481	1.0000

Korelační koeficient

- měří směr a míru lineární závislosti
- nabývá jen hodnot z intervalu $\langle -1, 1 \rangle$
- $r_{x,y} \approx 0$ (znaky x a y vzájemně nezávislé)
- $r_{x,y}$ blízko 1 (kladná závislost: s rostoucím x znak y v průměru roste)
- $r_{x,y}$ blízko -1 (záporná závislost: s rostoucím x znak y v průměru klesá)

U našich dat lze spočítat pro každou dvojici znaků dívka, iq, zn7, zn8: tzv. **korelační matice**

	dívka	iq	zn7	zn8
dívka	1.0000	-0.0597	-0.3054	-0.2661
iq	-0.0597	1.0000	-0.6559	-0.6236
zn7	-0.3054	-0.6559	1.0000	0.9481
zn8	-0.2661	-0.6236	0.9481	1.0000

Korelační koeficient

- měří směr a míru lineární závislosti
- nabývá jen hodnot z intervalu $\langle -1, 1 \rangle$
- $r_{x,y} \approx 0$ (znaky x a y vzájemně nezávislé)
- $r_{x,y}$ blízko 1 (kladná závislost: s rostoucím x znak y v průměru roste)
- $r_{x,y}$ blízko -1 (záporná závislost: s rostoucím x znak y v průměru klesá)

U našich dat lze spočítat pro každou dvojici znaků dívka, iq, zn7, zn8: tzv. **korelační matice**

	dívka	iq	zn7	zn8
dívka	1.0000	-0.0597	-0.3054	-0.2661
iq	-0.0597	1.0000	-0.6559	-0.6236
zn7	-0.3054	-0.6559	1.0000	0.9481
zn8	-0.2661	-0.6236	0.9481	1.0000

Korelační koeficient

- měří směr a míru lineární závislosti
- nabývá jen hodnot z intervalu $\langle -1, 1 \rangle$
- $r_{x,y} \approx 0$ (znaky x a y vzájemně nezávislé)
- $r_{x,y}$ blízko 1 (kladná závislost: s rostoucím x znak y v průměru roste)
- $r_{x,y}$ blízko -1 (záporná závislost: s rostoucím x znak y v průměru klesá)

U našich dat lze spočítat pro každou dvojici znaků dívka, iq, zn7, zn8: tzv. **korelační matice**

	dívka	iq	zn7	zn8
dívka	1.0000	-0.0597	-0.3054	-0.2661
iq	-0.0597	1.0000	-0.6559	-0.6236
zn7	-0.3054	-0.6559	1.0000	0.9481
zn8	-0.2661	-0.6236	0.9481	1.0000

Korelační koeficient

- měří směr a míru lineární závislosti
- nabývá jen hodnot z intervalu $\langle -1, 1 \rangle$
- $r_{x,y} \approx 0$ (znaky x a y vzájemně nezávislé)
- $r_{x,y}$ blízko 1 (kladná závislost: s rostoucím x znak y v průměru roste)
- $r_{x,y}$ blízko -1 (záporná závislost: s rostoucím x znak y v průměru klesá)

U našich dat lze spočítat pro každou dvojici znaků dívka, iq, zn7, zn8: tzv. **korelační matice**

	dívka	iq	zn7	zn8
dívka	1.0000	-0.0597	-0.3054	-0.2661
iq	-0.0597	1.0000	-0.6559	-0.6236
zn7	-0.3054	-0.6559	1.0000	0.9481
zn8	-0.2661	-0.6236	0.9481	1.0000

Korelační koeficient

- měří směr a míru lineární závislosti
- nabývá jen hodnot z intervalu $\langle -1, 1 \rangle$
- $r_{x,y} \approx 0$ (znaky x a y vzájemně nezávislé)
- $r_{x,y}$ blízko 1 (kladná závislost: s rostoucím x znak y v průměru roste)
- $r_{x,y}$ blízko -1 (záporná závislost: s rostoucím x znak y v průměru klesá)

U našich dat lze spočítat pro každou dvojici znaků dívka, iq, zn7, zn8: tzv. **korelační matice**

	dívka	iq	zn7	zn8
dívka	1.0000	-0.0597	-0.3054	-0.2661
iq	-0.0597	1.0000	-0.6559	-0.6236
zn7	-0.3054	-0.6559	1.0000	0.9481
zn8	-0.2661	-0.6236	0.9481	1.0000

Korelační koeficient

- měří směr a míru lineární závislosti
- nabývá jen hodnot z intervalu $\langle -1, 1 \rangle$
- $r_{x,y} \approx 0$ (znaky x a y vzájemně nezávislé)
- $r_{x,y}$ blízko 1 (kladná závislost: s rostoucím x znak y v průměru roste)
- $r_{x,y}$ blízko -1 (záporná závislost: s rostoucím x znak y v průměru klesá)

U našich dat lze spočítat pro každou dvojici znaků dívka, iq, zn7, zn8: tzv. **korelační matice**

	dívka	iq	zn7	zn8
dívka	1.0000	-0.0597	-0.3054	-0.2661
iq	-0.0597	1.0000	-0.6559	-0.6236
zn7	-0.3054	-0.6559	1.0000	0.9481
zn8	-0.2661	-0.6236	0.9481	1.0000

Regresní přímka - metoda nejmenších čtverců

- Máme sadu dvojic $(x_i, y_i), i = 1, \ldots, n$. Chceme z daných hodnot znaku x odhadnout hodnoty znaku y. Předpokládáme lineární závislost y na x, tj. že přibližně platí

$$y \doteq a + b \cdot x$$

- Parametry a a b regresní přímky se odhadnou **metodou nejmenších čtverců**, tj. hledáme hodnoty, pro které je výraz $\sum_{i=1}^{n}(y_i - (a + b \cdot x_i))^2$ minimální. Řešením jsou:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i \cdot y_i) - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} = \frac{S_{xy}}{S_x^2} \qquad \hat{a} = \overline{y} - \hat{b} \cdot \overline{x}$$

Regresní přímka - metoda nejmenších čtverců

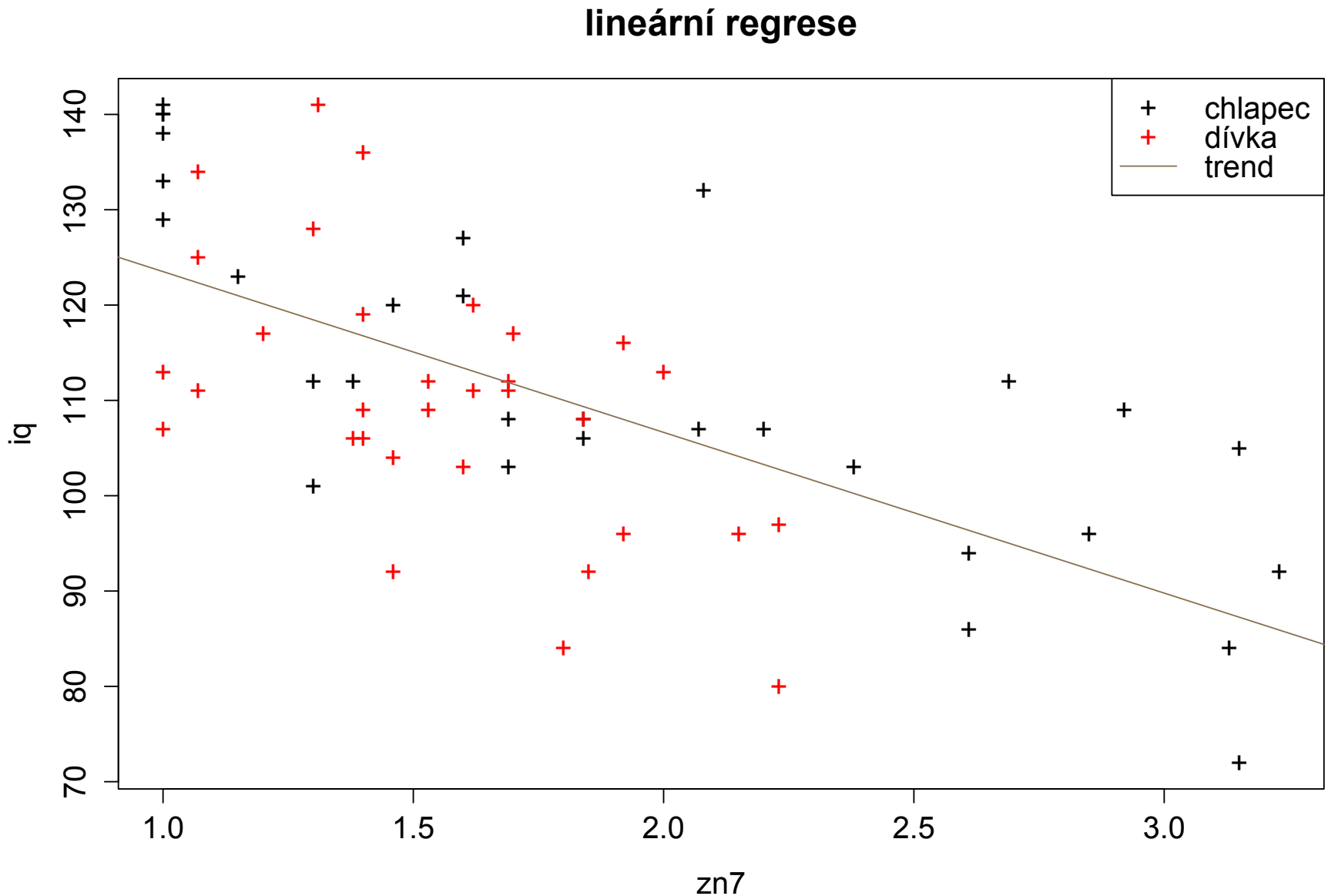
- Máme sadu dvojic $(x_i, y_i), i = 1, \ldots, n$. Chceme z daných hodnot znaku x odhadnout hodnoty znaku y. Předpokládáme lineární závislost y na x, tj. že přibližně platí

$$y \doteq a + b \cdot x$$

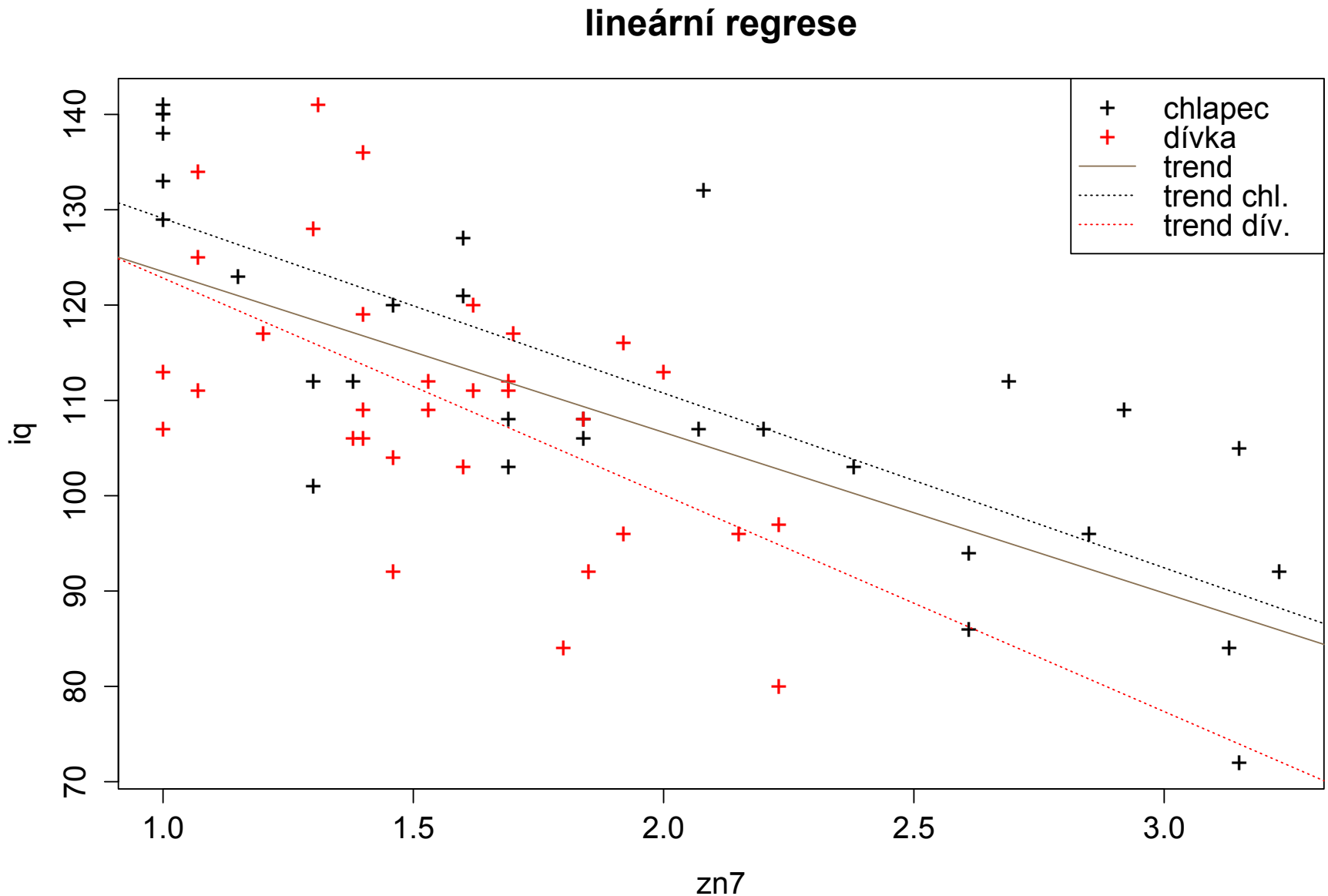
- Parametry a a b regresní přímky se odhadnou **metodou nejmenších čtverců**, tj. hledáme hodnoty, pro které je výraz $\sum_{i=1}^{n}(y_i - (a + b \cdot x_i))^2$ minimální. Řešením jsou:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i \cdot y_i) - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} = \frac{S_{xy}}{S_x^2} \qquad \hat{a} = \overline{y} - \hat{b} \cdot \overline{x}$$

Regresní přímka: znázornění lineární závislosti dvou kvantitativních znaků



Regresní přímka: znázornění lineární závislosti dvou kvantitativních znaků



Míra závislosti v kontingenční tabulce

Někdy máme k dispozici data v kontingenční tabulce, např. proto, že měříme současně dva znaky v nominálním měřítku na n nezávislých objektech. Cílem je opět zjistit míru závislost mezi těmito dvěma znaky.

Př.: Za účelem zjištění, zda existuje vztah mezi pohlavím a úrovní strachu z matematiky bylo náhodně vybráno 100 středoškolských studentů, kteří byli podrobeni psychologickému testu, kterým byla zjištěna úrovně strachu (nízká, střední, vysoká), který v nich vyvolává matematika. Výsledky byly následující:

	strach z matematiky			
pohlaví	nízký	střední	vysoký	součet
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

lze použít míru založenou na Pearsonově χ^2 statistice: porovnává napozorované četnosti s očekávanými za nezávislosti znaků

Míra závislosti v kontingenční tabulce

Někdy máme k dispozici data v kontingenční tabulce, např. proto, že měříme současně dva znaky v nominálním měřítku na n nezávislých objektech. Cílem je opět zjistit míru závislost mezi těmito dvěma znaky.

Př.: Za účelem zjištění, zda existuje vztah mezi pohlavím a úrovní strachu z matematiky bylo náhodně vybráno 100 středoškolských studentů, kteří byli podrobeni psychologickému testu, kterým byla zjištěna úrovně strachu (nízká, střední, vysoká), který v nich vyvolává matematika. Výsledky byly následující:

	strach z matematiky			
pohlaví	nízký	střední	vysoký	součet
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

Lze použít míru založenou na Pearsonově χ^2 statistice: porovnává napozorované četnosti s očekávanými za nezávislosti znaků

,

Míra závislosti v kontingenční tabulce

Někdy máme k dispozici data v kontingenční tabulce, např. proto, že měříme současně dva znaky v nominálním měřítku na n nezávislých objektech. Cílem je opět zjistit míru závislost mezi těmito dvěma znaky.

Př.: Za účelem zjištění, zda existuje vztah mezi pohlavím a úrovní strachu z matematiky bylo náhodně vybráno 100 středoškolských studentů, kteří byli podrobeni psychologickému testu, kterým byla zjištěna úrovně strachu (nízká, střední, vysoká), který v nich vyvolává matematika. Výsledky byly následující:

	strach z matematiky			
pohlaví	nízký	střední	vysoký	součet
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

Ize použít míru založenou na Pearsonově χ^2 statistice: porovnává napozorované četnosti s očekávanými za nezávislosti znaků

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	18%	46%	36%	100%
žena	9%	23%	68%	100%
celkem	14%	36%	50%	100%

- existuje vztah mezi pohlavím a strachem z matematiky?
- pokud jsou tyto dva znaky nezávislé, rozdělení procent pro obě pohlaví by mělo být podobné
- odhad pravděp., že pohl. studenta je ženské
  $P(\text{pohl.} = \text{ž}) = 44/100$
- odhad pravd., že strach studenta je vys. $P(\text{strach} = \text{v}) = 50/100$
- tedy odhad pravděp. (za nezávislosti), že studentem je žena s vysokým strachem
  $P(\text{pohl.} = \text{ž} \cap \text{strach} = \text{v}) = (44/100) \cdot (50/100) = 0{,}22$
- tedy mezi 100 studenty bych takových studentů očekával
  $100 \cdot (44/100) \cdot (50/100) = 22$
- podobně: očekávané četnosti pro 5 zbývajících buněk.

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	18%	46%	36%	100%
žena	9%	23%	68%	100%
celkem	14%	36%	50%	100%

- existuje vztah mezi pohlavím a strachem z matematiky?
- pokud jsou tyto dva znaky nezávislé, rozdělení procent pro obě pohlaví by mělo být podobné
- odhad pravděp., že pohl. studenta je ženské
  $P(\text{pohl.} = \text{ž}) = 44/100$
- odhad pravd., že strach studenta je vys. $P(\text{strach} = \text{v}) = 50/100$
- tedy odhad pravděp. (za nezávislosti), že studentem je žena s vysokým strachem
  $P(\text{pohl.} = \text{ž} \cap \text{strach} = \text{v}) = (44/100) \cdot (50/100) = 0{,}22$
- tedy mezi 100 studenty bych takových studentů očekával
  $100 \cdot (44/100) \cdot (50/100) = 22$
- podobně: očekávané četnosti pro 5 zbývajících buněk.

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	18%	46%	36%	100%
žena	9%	23%	68%	100%
celkem	14%	36%	50%	100%

- existuje vztah mezi pohlavím a strachem z matematiky?
- pokud jsou tyto dva znaky nezávislé, rozdělení procent pro obě pohlaví by mělo být podobné
- odhad pravděp., že pohl. studenta je ženské
  $P(\text{pohl.} = \text{ž}) = 44/100$
- odhad pravd., že strach studenta je vys. $P(\text{strach} = \text{v}) = 50/100$
- tedy odhad pravděp. (za nezávislosti), že studentem je žena s vysokým strachem
  $P(\text{pohl.} = \text{ž} \cap \text{strach} = \text{v}) = (44/100) \cdot (50/100) = 0{,}22$
- tedy mezi 100 studenty bych takových studentů očekával
  $100 \cdot (44/100) \cdot (50/100) = 22$
- podobně: očekávané četnosti pro 5 zbývajících buněk.

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	18%	46%	36%	100%
žena	9%	23%	68%	100%
celkem	14%	36%	50%	100%

- existuje vztah mezi pohlavím a strachem z matematiky?
- pokud jsou tyto dva znaky nezávislé, rozdělení procent pro obě pohlaví by mělo být podobné
- odhad pravděp., že pohl. studenta je ženské
  $P(\text{pohl.} = \text{ž}) = 44/100$
- odhad pravd., že strach studenta je vys. $P(\text{strach} = \text{v}) = 50/100$
- tedy odhad pravděp. (za nezávislosti), že studentem je žena s vysokým strachem
  $P(\text{pohl.} = \text{ž} \cap \text{strach} = \text{v}) = (44/100) \cdot (50/100) = 0{,}22$
- tedy mezi 100 studenty bych takových studentů očekával
  $100 \cdot (44/100) \cdot (50/100) = 22$
- podobně: očekávané četnosti pro 5 zbývajících buněk.

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	18%	46%	36%	100%
žena	9%	23%	68%	100%
celkem	14%	36%	50%	100%

- existuje vztah mezi pohlavím a strachem z matematiky?
- pokud jsou tyto dva znaky nezávislé, rozdělení procent pro obě pohlaví by mělo být podobné
- odhad pravděp., že pohl. studenta je ženské
  $P(\text{pohl.} = \text{ž}) = 44/100$
- odhad pravd., že strach studenta je vys. $P(\text{strach} = \text{v}) = 50/100$
- tedy odhad pravděp. (za nezávislosti), že studentem je žena s vysokým strachem
  $P(\text{pohl.} = \text{ž} \cap \text{strach} = \text{v}) = (44/100) \cdot (50/100) = 0{,}22$
- tedy mezi 100 studenty bych takových studentů očekával $100 \cdot (44/100) \cdot (50/100) = 22$
- podobně: očekávané četnosti pro 5 zbývajících buněk.

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	18%	46%	36%	100%
žena	9%	23%	68%	100%
celkem	14%	36%	50%	100%

- existuje vztah mezi pohlavím a strachem z matematiky?
- pokud jsou tyto dva znaky nezávislé, rozdělení procent pro obě pohlaví by mělo být podobné
- odhad pravděp., že pohl. studenta je ženské
  $P(\text{pohl.} = \text{ž}) = 44/100$
- odhad pravd., že strach studenta je vys. $P(\text{strach} = \text{v}) = 50/100$
- tedy odhad pravděp. (za nezávislosti), že studentem je žena s vysokým strachem
  $P(\text{pohl.} = \text{ž} \cap \text{strach} = \text{v}) = (44/100) \cdot (50/100) = 0{,}22$
- tedy mezi 100 studenty bych takových studentů očekával
  $100 \cdot (44/100) \cdot (50/100) = 22$
- podobně: očekávané četnosti pro 5 zbývajících buněk.

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	18%	46%	36%	100%
žena	9%	23%	68%	100%
celkem	14%	36%	50%	100%

- existuje vztah mezi pohlavím a strachem z matematiky?
- pokud jsou tyto dva znaky nezávislé, rozdělení procent pro obě pohlaví by mělo být podobné
- odhad pravděp., že pohl. studenta je ženské
  $P(\text{pohl.} = \text{ž}) = 44/100$
- odhad pravd., že strach studenta je vys. $P(\text{strach} = \text{v}) = 50/100$
- tedy odhad pravděp. (za nezávislosti), že studentem je žena s vysokým strachem
  $P(\text{pohl.} = \text{ž} \cap \text{strach} = \text{v}) = (44/100) \cdot (50/100) = 0{,}22$
- tedy mezi 100 studenty bych takových studentů očekával
  $100 \cdot (44/100) \cdot (50/100) = 22$
- podobně: očekávané četnosti pro 5 zbývajících buněk.

Míra závislosti v kontingenční tabulce

- označme n_{ij} četnost v i-tém řádku a j-tém sloupci tabulky (celkem I řádků a J sloupců)
- označme n_{i+} (resp. n_{+j}) součet četností v i-tém řádku (resp. j-tém sloupci)
- očekávaná četnost v i-tém řádku a j-tém sloupci za hypotézy nezávislosti je

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

Pearsonova statistika je mírou shody mezi n_{ij} a o_{ij}:

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

Míra závislosti v kontingenční tabulce

- označme n_{ij} četnost v i-tém řádku a j-tém sloupci tabulky (celkem I řádků a J sloupců)
- označme n_{i+} (resp. n_{+j}) součet četností v i-tém řádku (resp. j-tém sloupci)
- očekávaná četnost v i-tém řádku a j-tém sloupci za hypotézy nezávislosti je

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

Pearsonova statistika je mírou shody mezi n_{ij} a o_{ij}:

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

Míra závislosti v kontingenční tabulce

- označme n_{ij} četnost v i-tém řádku a j-tém sloupci tabulky (celkem I řádků a J sloupců)
- označme n_{i+} (resp. n_{+j}) součet četností v i-tém řádku (resp. j-tém sloupci)
- očekávaná četnost v i-tém řádku a j-tém sloupci za hypotézy nezávislosti je

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

Pearsonova statistika je mírou shody mezi n_{ij} a o_{ij}:

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

Míra závislosti v kontingenční tabulce

- označme n_{ij} četnost v i-tém řádku a j-tém sloupci tabulky (celkem I řádků a J sloupců)
- označme n_{i+} (resp. n_{+j}) součet četností v i-tém řádku (resp. j-tém sloupci)
- očekávaná četnost v i-tém řádku a j-tém sloupci za hypotézy nezávislosti je

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

Pearsonova statistika je mírou shody mezi n_{ij} a o_{ij}:

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

z ▸ příkladu:

Napozorované (resp. očekávané) četnosti jsou:

pohlaví	**strach z matematiky** nízký	střední	vysoký	součet
muž	10 (7,84)	26 (20,16)	20 (28)	56
žena	4 (6,16)	10 (15,84)	30 (22)	44
součet	14	36	50	100

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-o_{ij})^2}{o_{ij}} = \frac{(10-7{,}84)^2}{7{,}84} + \frac{(26-20{,}16)^2}{20{,}16} +$$

$$+ \frac{(20-28)^2}{28} + \frac{(4-6{,}16)^2}{6{,}16} + \frac{(10-15{,}84)^2}{15{,}84} + \frac{(30-22)^2}{22} = 10{,}39$$

,

z příkladu:

Napozorované (resp. očekávané) četnosti jsou:

pohlaví	**strach z matematiky** nízký	střední	vysoký	součet
muž	10 (7,84)	26 (20,16)	20 (28)	56
žena	4 (6,16)	10 (15,84)	30 (22)	44
součet	14	36	50	100

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-o_{ij})^2}{o_{ij}} = \frac{(10-7{,}84)^2}{7{,}84} + \frac{(26-20{,}16)^2}{20{,}16} +$$

$$+ \frac{(20-28)^2}{28} + \frac{(4-6{,}16)^2}{6{,}16} + \frac{(10-15{,}84)^2}{15{,}84} + \frac{(30-22)^2}{22} = 10{,}39$$

Míra závislosti v kontingenční tabulce

Pearsonova statistika χ^2 neomezeně roste s počtem pozorování n.

Proto jako míry závislosti: tzv. koeficienty kontingence:

Cramérův koeficient V

$$V = \sqrt{\chi^2/(nh)}$$

$$h = \min(r-1, s-1)$$

Pearsonův koeficient kontingence

$$C_p = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

$$max(C_p) = \sqrt{\frac{h}{h+1}}.$$

$$C_p^\star = \frac{C_p}{\max(C_p)}.$$

Míra závislosti v kontingenční tabulce

Pearsonova statistika χ^2 neomezeně roste s počtem pozorování n.

Proto jako míry závislosti: tzv. koeficienty kontingence:

Cramérův koeficient V

$$V = \sqrt{\chi^2/(nh)}$$

$$h = \min(r-1, s-1)$$

Pearsonův koeficient kontingence

$$C_p = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

$$max(C_p) = \sqrt{\frac{h}{h+1}}.$$

$$C_p^\star = \frac{C_p}{\max(C_p)}.$$

,

Míra závislosti v kontingenční tabulce

Pearsonova statistika χ^2 neomezeně roste s počtem pozorování n.

Proto jako míry závislosti: tzv. koeficienty kontingence:

Cramérův koeficient V

$$V = \sqrt{\chi^2/(nh)} = \sqrt{10{,}39/(100 \cdot 2)} = 0{,}228$$

$$h = \min(r-1, s-1) = \min(2-1, 3-1) = 2$$

Pearsonův koeficient kontingence

$$C_p = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{10{,}39}{10{,}39 + 100}} = 0{,}307,$$

$$max(C_p) = \sqrt{\frac{h}{h+1}} = \sqrt{\frac{2}{2+1}} = 0{,}816.$$

$$C_p^\star = \frac{C_p}{\max(C_p)} = \frac{0{,}307}{0{,}816} = 0{,}376.$$