Popisná statistika

Klíčové pojmy:

- popisná x inferenční statistika
- absolutní a relativní četnost
- mód, medián, průměr
- rozptyl, směrodatná odhylka

Co je to statistika? Existuje mnoho různých definic. Můžeme říci, že statistika je nástroj, který aplikuje matematicku na poznání informace o nějakých datech. Je důležité si uvědomit, že statistika je nástroj k testování vědeckých hypotéz pomocí dat (evidence). Statistika není dobrá k vytváření hypotéz nebo dokonce vědeckých teorií.

Statistika je užitečná na dva úkoly:

1) Popisování většího množství dat (popisná statistika)

2) Předpovídání nějakého fenomenu (inferenční statistika)

V této přednášce se budeme věnovat popisné statistice. Statistika popisná nám popisuje nějaký soubor čísel. Například, kolik pozorujeme na určitém cvičení Úvodu do statistiky žen a mužů. Statistika inferenční nám pak pomáhá na základě naměřených dat dělat úsudky o celé populaci studentů statistiky, tedy včetně cvičení, které jsme přímo neměřili.

Úkolem popisné statistiky je shrnout informace o našem výběru do pár čísel, které nám pomohou pochopit jaké má náš výběr vlastnosti. Hlavními vlastnostmi, které nás zajímají je:

1) Jakáje typická hodnota měřené proměnné

2) Na kolik se liší hodnoty jendotlivých pozorování (jak jsou rozptýlené)

Výběr popisné statistiky záleží na typu proměnné, kterou měříme. Z předchozí přednášky víme, že existují 3 typy proměnných: 1) Nominální 2) Ordinální 3) Kardinální

Dále víme, že nominální proměnné nemůžeme seřadit, ordinální můžeme seřadit a kardinální můžeme seřadit a zároveň říci o kolik je nějaká hodnota větší než jiná.

Nominální proměnná

Příkladem nominální promněnné je například barva. Vezměme si například situaci, v kterém bychom zjišťovali informace o barvě auta na nějakém konkrétním parkovišti. Na parkovišti parkuje 100 aut. Všechna auta bychom obešli a barvu zaznamenali. Jaký je nejlepší způsob, jak se něco dozvědět o všech 100 autech? Mohli bychom si všechny barvy přečíst a snažit se je zapomatovat tak. To by ale bylo obtížné. A právě proto na to nám slouží popisná statistika. Pomocí jednoho čísla můžeme vystihnout nejpočetnější barvu. Takové popisné statistice říkáme **modus**.

```
barvy <- c("cervena", "stribrna", "zelena", "zluta", "bila")
barvy_aut <- sample(barvy, size = 100, replace = TRUE)
# vytvorime tabulku relativnich cetnosti
tabulka <- table(barvy_aut)
print(tabulka)
```

```
## barvy_aut
##     bila  cervena stribrna   zelena    zluta
##       21       22       22       16       19
```

```
# vratime nazev kategorie, ktera se vyskytuje nejcasteji
names(tabulka)[which.max(tabulka)]
```

```
## [1] "cervena"
```

Protože budeme modus počítat dále, vytvoříme si funkci, která bude výpočet provádět.

```
modus <- function(x) {
  t <- table(x)
  return (names(t)[which.max(t)])
}
```

Ordinální proměnná

U ordinální proměnné, stejně jako u nominální, nemůžeme vypočítat o kolik je nějaká hodnota větší než druhá. Můžeme ale hodnoty seřadit. Toho se využívá k vypočítání popisné statistiky zvané **medián**. Medián nám značí prostřední hodnotu nějaké proměnné. Můžete si pro představit tak, že hodnoty proměnné seřadíte od nejmenší po největší a vyberete hodnotu, která bude přesně uprostřed. No a tato hodnota je medián. Matematicky se medián u proměnné x vypočítá jako $median(x) = x_{(n+1)/2}$. Pokud má naše proměnná sudý počet čísel, vypočítá se medián zpravidla jako průměr dvou prostředních hodnot, tedy $median(x) = \frac{x_{n/2}+x_{n/2+1}}{2}$. Ukážeme si její výpočet na příkladu vzdělání. V našich datech máme 3 stupně vzdělání - zš, sš a vš. Nasimulume příklad, v kterém budeme mít 9 dat.

```
vzdelani <- c("zs", "ss", "vs")
vzdelani_vyber <- sample(vzdelani, size = 9, replace = TRUE)

#udelame z character faktor (abychom mohli mit serazene hodnoty)
vzdelani_vyber <- factor(x = vzdelani_vyber, levels = vzdelani)
vzdelani_vyber <- sort(vzdelani_vyber)
print(paste0("Typ promenne, po prevedeni na factor: ", class(vzdelani_vyber)))
```

```
## [1] "Typ promenne, po prevedeni na factor: factor"
```

```
print(vzdelani_vyber)
```

```
## [1] zs ss ss ss vs vs vs vs vs
## Levels: zs ss vs
```

```
# vypocitame median
n <- 9
indx <- (n + 1) / 2
median_vzdelani <- sort(vzdelani_vyber)[indx]

print(paste0("Median vzdelani je: ", median_vzdelani))
```

```
## [1] "Median vzdelani je: vs"
```

Samozřejmě můžeme u ordinálních proměnných počítat také mód.

```
print(paste0("Modus promenne je: ", modus(vzdelani_vyber)))
```

```
## [1] "Modus promenne je: vs"
```

Jak si asi pamatujete z předchozího semestru, u různých proměnných jsme počítali relativní četnosti (četnost je počet pozorování učité hodnoty. Relativní četnost potom počet pozorování děleno počet případů celkem).

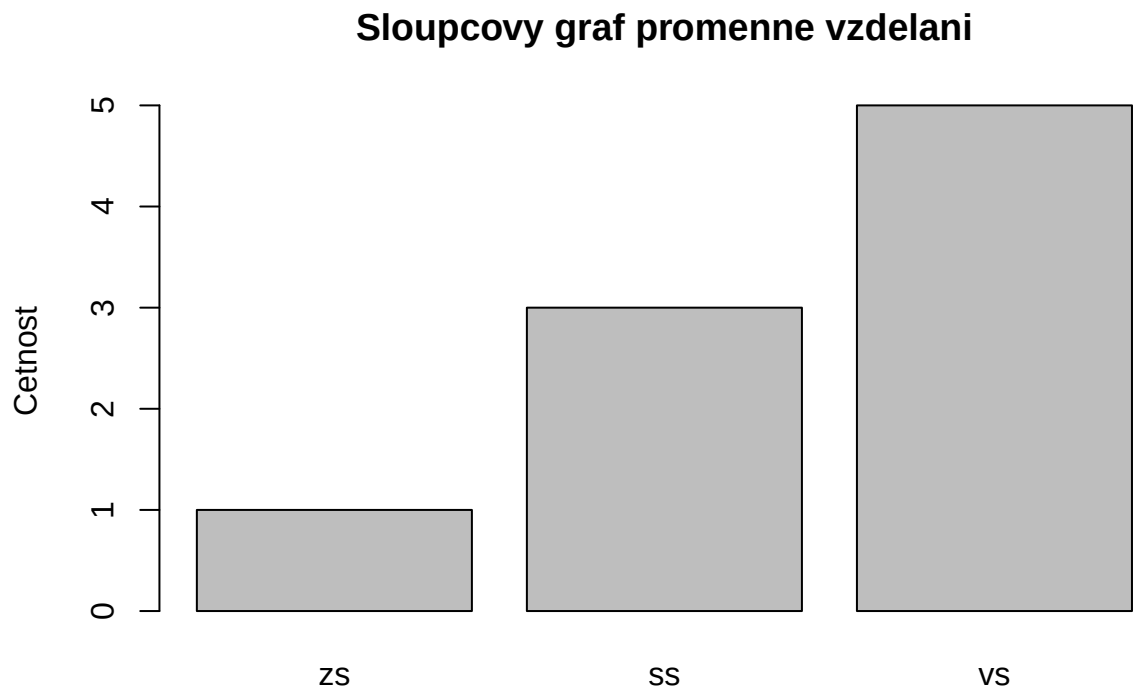
Například jsme počítali relativní četnost výsledků hodů mincí. Výsledky hodu mincí jsou také nominální proměnnou. Další statistikou, kterou můžete u nominálních a ordinálních proměnných spočítat jsou tedy četnosti. Ty nám řeknou více o rozložení hodnot v proměnné. Pojďme si takový příklad ukázat na našem vzorku se vzděláním.

```
#absolutni cetnost
absolutni_cetnost <- table(vzdelani_vyber)
#relativni cetnost
relativni_cetnost <- table(vzdelani_vyber) / length(vzdelani_vyber)
relativni_cetnost
```

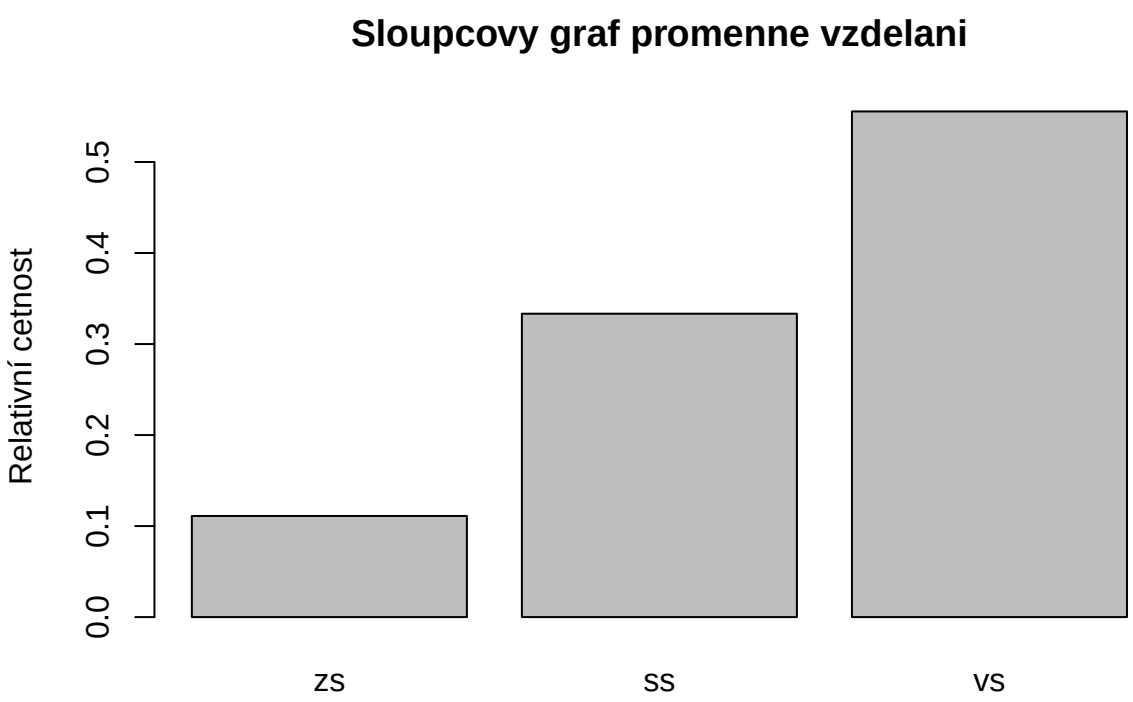
```
## vzdelani_vyber
##        zs        ss        vs
## 0.1111111 0.3333333 0.5555556
```

Nejčastějším způsobem zobrazení četností je sloupcový graf (bar plot).

```
barplot(table(vzdelani_vyber),
        main = "Sloupcovy graf promenne vzdelani",
        ylab = "Cetnost")
```




```
barplot(table(vzdelani_vyber) / length(vzdelani_vyber),
        main = "Sloupcovy graf promenne vzdelani",
        ylab = "Relativní cetnost")
```



Kardinální proměnná

Kardinální proměnná nám umožňují seřadit hodnoty a říci o kolik jsou větší. Kardinální proměnné jsou tedy číselné. Rozlišujeme mezi **diskrétní** a **spojitou**. Diskrétní nabývá celých čísel (1,2,3,4 etc., například počet dětí), tedy $\in Z$. Spojitá proměnná pak teoreticky nebývá nekonečně mnoho hodnot, prakticky je ale omezena tím, jak přesně dokážeme danou metriku měřit. Platí ale, že spojité proměnné nabývají racionálních čísel, tedy $\in R$. Stejně jako u ordinální proměnné, můžeme vypočítat modus a medián. Další popisnou statistikou, která nám prozradí něco o velikosti hodnot vnašich datech je **průměr**. Průměr proměnné x vypočítáme jako $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Pojďme si ukázat, jak průměr vypočítat na datech o počtu dětí.

```
rodiny <- c(2,2,2,2,2,1,1,1,3,3,4,0)
pocet_deti <- sample(rodiny, size = 1000, replace = TRUE)

prumer <- sum(pocet_deti) / length(pocet_deti)
print(paste0("Prumer promenne pocet deti je :", prumer))
```

```
## [1] "Prumer promenne pocet deti je :1.909"
```

Můžeme také požít funckci `mean`

```
print(paste0("Prumer pomoci funkce mean :", mean(pocet_deti)))
```

```
## [1] "Prumer pomoci funkce mean :1.909"
```

Pojďme si ještě vypočítat modus a medián. U číselných proměnných (`numeric`) můžeme k výpočtu mediánu použít funkci `median`.

```
print(paste0("Modus promenne pocet deti je: ", modus(pocet_deti)))
```

```
## [1] "Modus promenne pocet deti je: 2"
```

```
print(paste0("Media promenne pocet deti je: ", median(pocet_deti)))
```

```
## [1] "Media promenne pocet deti je: 2"
```

Někdy nechceme všem pozorováním při výpočtu průměru dát stejnou váhu. V takovém případě vypočítáme **vážený průměr**. Jeho vzorec je$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$. Pokud bychom například měli pouze informace o četnosti počtu dětí, tedy:

```
print(table(pocet_deti))
```

```
## pocet_deti
##   0   1   2   3   4
##  83 246 418 185  68
```

nebylo by průměrný počet dětí v datech možné vypočítat pomocí $\overline{x} = (0+1+2+3+4)/5$, protože v datech nemáme stejný počet rodin s 0 dětmi, 1 dítětem apod. Musíme jednotlivým hodnotám dát jinou váhu w_i podle toho, kolik jich je v našich datech. V R bychom takovýto vážený průměr mohli vypočítat jako

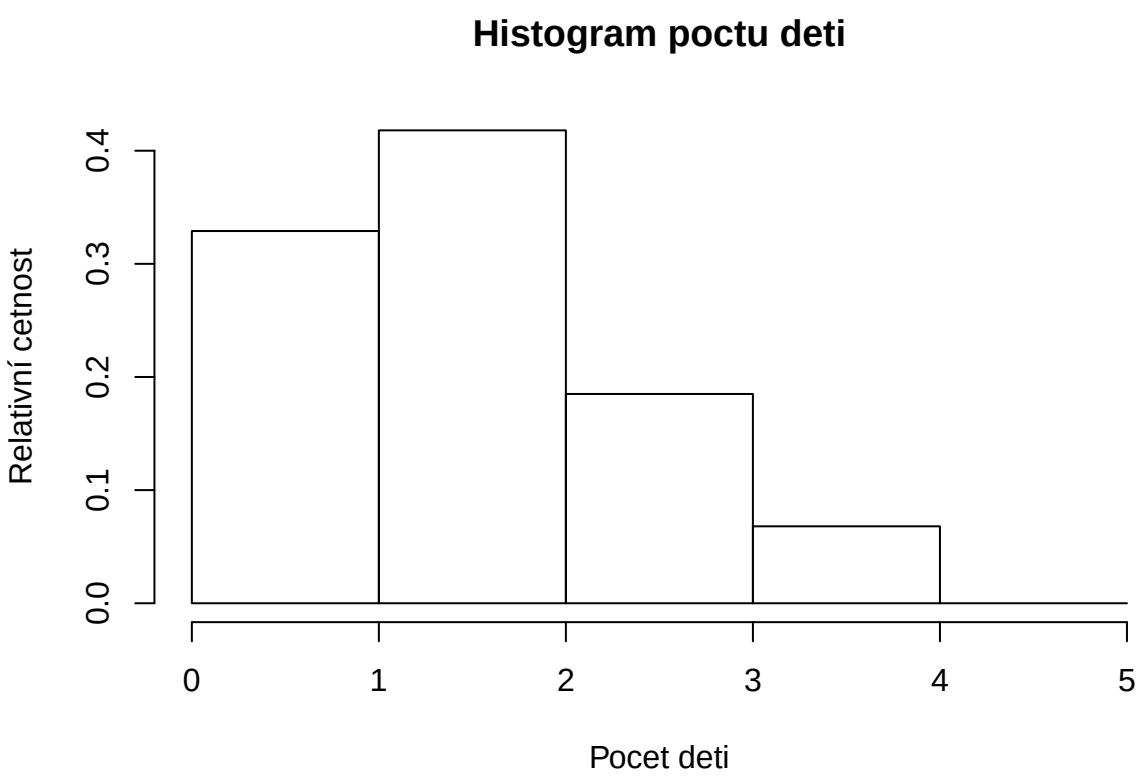
```
w <- table(pocet_deti)
# prevedeme hodnoty 0,1,2,3,4 do numericke promenne,
# abychom mohli nasobit a scitat
x_i <- as.numeric(names(table(pocet_deti)))

vazeny_prumer <- sum(w * x_i) / sum(w)
print(paste0("Vážený průměr je: ", vazeny_prumer))
```

```
## [1] "Vážený průměr je: 1.909"
```

Cože je stejné číslo, jaké jsme dostali, když jsme měli k dispozici všech 1000 pozorování proměnné.

U kardinálních spojitých proměnných používáme k vizuálnímu ověření rozložení hodnot proměnné **histogram**. Histogram je jako sloupcový graf absolutních/relativních četností. Protože u spojitých proměnných neexistuje přirozená hranice pro sloupec, funkce `hist` vytvoří základní hranice pro sloupce. Ty ale můžete upravit pomocí argumentu `breaks`.

```
hist(pocet_deti, breaks = c(0,1,2,3,4,5), probability = TRUE,
     xlab = "Pocet deti",
     ylab = "Relativní cetnost",
     main = "Histogram poctu deti")
```



Jak jsme uvedli na začátku, úkolem popisné statistiky není pouze přiblížit, jak vypadá typická hodnota naší proměnné, ale také přiblížit, jak jsou od sebe odlišné. K tomu nám slouží metrika, kterou nazýváme rozptyl. **Rozptyl** vypočítáme tak, že každou hodnotu odečteme od průměru a umocníme. Tyto hodnoty sečteme a vydělíme počtem pozorování. Matematicky bychom rozptyl σ^2 proměnné x vypočítali jako $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Ve statistice se také používá pro výpočet rozptýlenosti **směrodatná odchylka** σ, které se vypočítá jako $\sigma = \sqrt{\sigma^2}$ Pojďme vypočítat rozptyl a směrodatnou odchylku u naší proměnné počet dětí.

```
vzdalenost_od_prumeru <- (pocet_deti - mean(pocet_deti))^2
rozptyl <- sum(vzdalenost_od_prumeru) / length(pocet_deti)
smerodatna_odchylka <- sqrt(rozptyl)
print(paste0("Rozptyl je:", rozptyl))
```

```
## [1] "Rozptyl je:1.026719"
```

```
print(paste0("Smerodatna odchylka je:", smerodatna_odchylka))
```

```
## [1] "Smerodatna odchylka je:1.01327143451298"
```

V R můžeme roztyl a směrodatnou odchylku vypočítat pomocí funkcí `var` a `sd`. Vzorec, který se v R používá počítá výběrový roztyl a směrodatnou odchylku, ve jmenovateli tedy používá $n-1$. Rozdíl v těchto přístupech teď není důležitý a ukážeme si ho na dalších cvičeních. Uvádíme ho zde jenom, abychom rozuměli proč jsu výsledky jiné než ty, které jsme spočítali my.

```
print(paste0("Rozptyl pomoci funkce var je:", var(pocet_deti)))
```

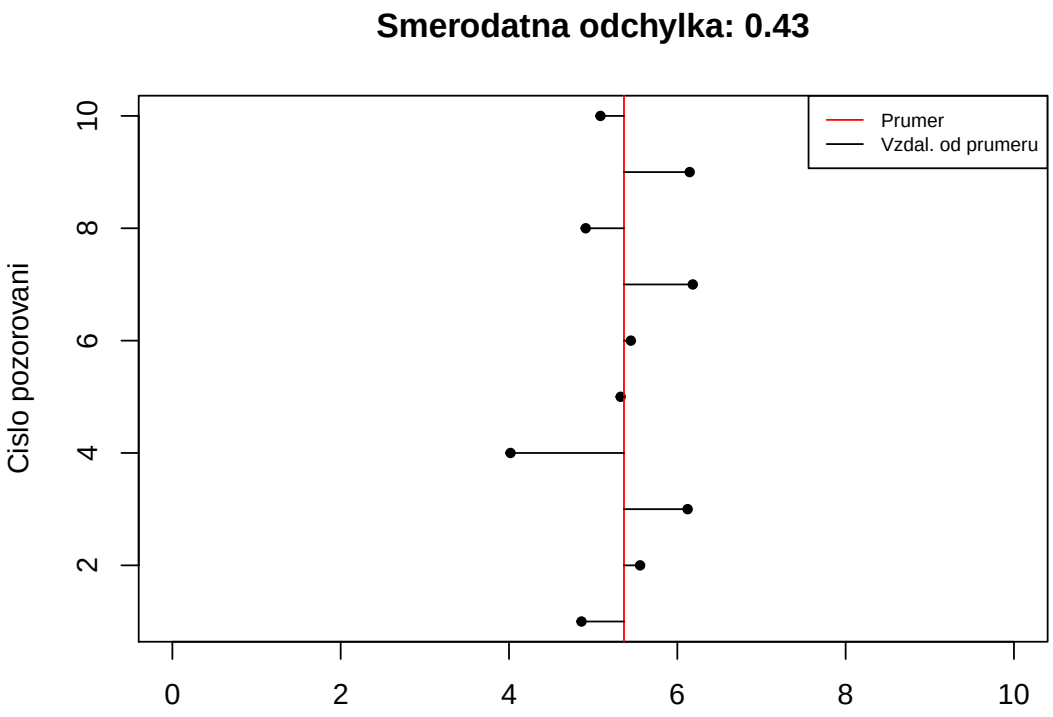
```
## [1] "Rozptyl pomoci funkce var je:1.02774674674675"
```

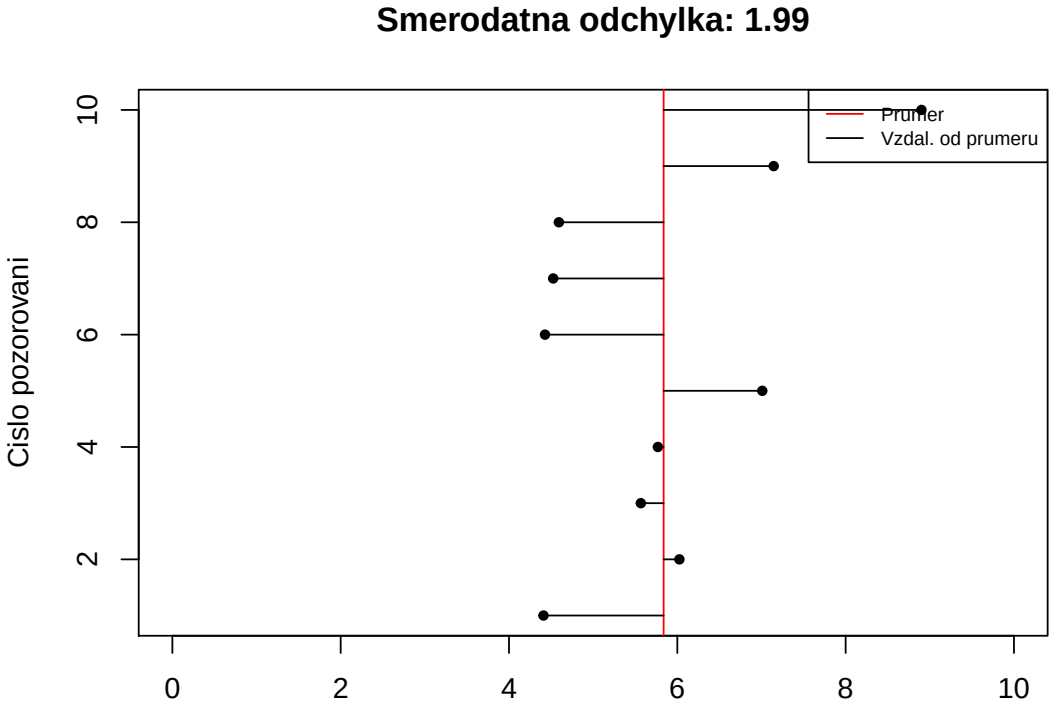
```r
print(paste0("Smerodatna odchylka pomoci funkce sd je:", sd(pocet_deti)))
```

```
## [1] "Smerodatna odchylka pomoci funkce sd je:1.01377845052395"
```

Ukažme si princip rozptylu/směrodatné odchylky na imaginárních datech. Na ukázku si vytvoříme proměnnou, která má 10 pozorování a zobrazíme je do grafu jako body. Červená čára označuje průměr těchto bodů. Horizontální čáry potom označují vzdálenost každého pozorování od průměrné hodnoty. Nejdříve si ukážeme příklad s menším rozptylem hodnot a pod ním příklad rozložení s větším rozptylem hodnot. Protože mají oba příklady stejný počet pozorování (10), můžete si rozdíl v jejich rozptylu představit jako rozdíl jejich

```r
par(mfrow = c(2,1))
x <- rnorm(10, 5, sd = 1)

prumer <- mean(x)
n <- seq(1, length(x))
smerodatna_odchylka <- sum((x - mean(x))^2) / (length(x))

plot(x, n,
     main = paste0("Smerodatna odchylka: ", round(smerodatna_odchylka, 2)),
     xlim = c(0,10),
     xlab = "", ylab = "Cislo pozorovani", pch = 20)
abline(v = prumer, col = "red")
for(i in n) {
    lines(c(prumer, x[i]), c(i,i), col = "black")
}
legend("topright",
        legend = c("Prumer", "Vzdal. od prumeru"),
       col = c("red", "black"),
       lty = c(1,1), cex = 0.7)

x2 <- rnorm(10,5, sd = 2)
prumer <- mean(x2)
n <- seq(1, length(x2))
smerodatna_odchylka <- sum((x2 - mean(x2))^2) / (length(x))

plot(x2, n,
     main = paste0("Smerodatna odchylka: ", round(smerodatna_odchylka, 2)),
     xlim = c(0,10),
     xlab = "", ylab = "Cislo pozorovani", pch = 20)
abline(v = prumer, col = "red")
for(i in n) {
    lines(c(prumer, x2[i]), c(i,i), col = "black")
}
legend("topright",
       legend = c("Prumer", "Vzdal. od prumeru"),
       col = c("red", "black"),
       lty = c(1,1), cex = 0.7)
```


Smerodatna odchylka: 0.43
Prumer
Vzdal. od prumeru
Cislo pozorovani
0
2
4
6
8
10


Smerodatna odchylka: 1.99
Prumer
Vzdal. od prumeru
Cislo pozorovani
0
2
4
6
8
10

Zkuste si zobrazit histrogram obou proměnných, abyste si udělali lepší představu o tom, jak vypadá jejich rozložení.