

ZÁKLADNÍ POPISNÉ STATISTIKY A GRAFY

23.11.2017

ÚVODNÍ NASTAVENÍ.

- Ve svém domovském adresáři si založte speciální adresář `nmsa331` na toto cvičení.
- Z internetové stránky `www.karlin.mff.cuni.cz/~hudecova/education/` si stáhněte datový soubor `Hosi.txt` a uložte si jej do adresáře `nmsa331`. Můžete si také stáhnout zdrojový kód k dnešnímu cvičení `cviceni8.R`.
- Otevřete si program `R Studio`.
- Změňte si pracovní adresář pomocí `Session` → `Set working directory` → `Choose directory` na Váš právě založený adresář `nmsa331` nebo napište přímo

```
setwd("H:/nmsa331")
```
- Pomocí `File` → `Open..` si otevřete soubor `cviceni8.R`. Během dnešního cvičení můžete buď jen postupně spouštět řádky tohoto souboru nebo pracovat samostatněji a psát si vlastní zdrojový kód na základě tohoto pdf.
- Nechejte si vypsat seznam objektů, které jsou aktivní:
`ls()`
a případně proveďte před další prací vyčištění
`rm(list=ls())`

ZÁKLADNÍ OPERACE V PROGRAMU R

1. R používáme buď tak, že píšeme příkazy přímo do okna `Console` nebo (což je preferováno) si příkazy píšeme do zvláštního souboru a odtud je spouštíme.
2. Použijte R jako kalkulačku a spočítejte následující výrazy:

$$1 + 1, \quad \frac{2}{3}, \quad 3^4, \quad \sqrt{3}, \quad \log(10), \quad \exp(10), \quad \sin\left(\frac{\pi}{2}\right).$$

Nechte si vypsat nápovědu k funkci `log` tak, že zadáte `?log`. Stejným způsobem si lze zavolat nápovědu ke každé funkci v R.

3. Do vektoru nazvaného `x` si uložíme počty bodů ze zápočtové písemky u chlapců a do `y` počty bodů u dívek.

```
x=c(30,57,67,70,72,80,80,89,91,93,98)
```

```
y=c(50,58,61,62,63,77,81,83,85)
```

Pomocí funkcí `min`, `max`, `mean` si spočítejte minimální, maximální a průměrné počty bodů. Jaké jsou mediány počtu bodů?

PRÁCE S DATY – POPISNÉ STATISTIKY

4. Načtete si data `Hosi.txt`. Bud' naklikáním pomocí `Import data set` nebo pomocí příkazu

```
Hosi=read.table("Hosi.txt",header=TRUE)
```

5. Základní prohlídka dat: Na data můžeme nahlédnout kliknutím na jejich název v seznamu proměnných vpravo nahoře. Užitečné příkazy jsou

```
head(Hosi)
dim(Hosi)
summary(Hosi)
```

Můžeme vidět, že v datech máme následující proměnné: porodní hmotnost v g, porodní délku v cm, věk matky, věk otce, hmotnost dítěte v 1 roce v g, délku v 1 roce a pořadové číslo dítěte.

6. Dále nás bude zajímat pouze porodní hmotnost. Můžeme předpokládat, že data odpovídají realizaci náhodného výběru. Pro jednoduchost si je uložíme do vektoru `hmot` a do `n` si uložíme rozsah výběru.

```
hmot= Hosi$por.hmot
(n=length(hmot))
```

7. Spočítejte si základní charakteristiky polohy: minimum, maximum, průměr.
8. Spočtete si medián pomocí funkce `median`. Pozor ale, R počítá výběrový medián trochu jinak než bylo zavedeno na přednášce (viz dále u kvantilů).
9. Spočítejte si základní charakteristiky variability pomocí funkcí `var`, `sd`. Uvědomte si, v jakých jednotkách jsou tyto kvantily a která z nich je tedy vhodnější pro prezentaci v případném výstupu.
Variabilitu můžeme ještě charakterizovat rozpětím, které získáme jako rozdíl minima a maxima.
10. Budeme počítat výběrové kvantily porodní hmotnosti. Jak je známo z přednášky, existuje více definic výběrových kvantilů. Zjistíme si, jak je počítá R

```
?quantile
```

Odtud vyčteme, že defaultně je nastaven jiný postup, než jaký byl na přednášce. Pro naše případy tedy budeme používat nastavení `type=1`. Spočteme si tedy několik výběrových kvantilů

```
(kvant <- quantile(hmot, prob = c(0.1, 0.25, 0.5, 0.75, 0.9), type=1))
```

Připomeňme si, že na přednášce byly výběrové kvantily definované tak, že pro $\alpha \in (0, 1)$ je $\hat{u}_n(\alpha) = X_{(k_\alpha)}$, kde $k_\alpha = \alpha n$, pokud αn je celé číslo, a $k_\alpha = \lfloor n\alpha \rfloor + 1$ pokud αn není celé číslo. Pro $\alpha = 0.25$ ověříme, že to opravdu odpovídá

```
quantile(hmot, prob=0.25, type=1)
sort(hmot)[floor(n*0.25)+1]
```

Pro $\alpha = 0.99$ vyzkoušejte zadat různé typy ve funkci `quantile` (1 až 9) a porovnejte výsledky.

11. Můžeme si nechat vykreslit obrázek empirické distribuční funkce a její souvislost s výběrovými kvantily:

```
plot.stepfun(ecdf(hmot), verticals=TRUE, do.points=FALSE, ylab=expression(F[n](x)),
  main="Empiricka distribucni funkce")
abline(h=0.1, col="blue")
abline(h=0.25, col="blue")
abline(h=0.5, col="blue")
abline(h=0.75, col="blue")
abline(h=0.9, col="blue")
lines(rep(kvant[1], 2), c(-1,0.1), col="red")
lines(rep(kvant[2], 2), c(-1,0.25), col="red")
lines(rep(kvant[3], 2), c(-1,0.5), col="red")
lines(rep(kvant[4], 2), c(-1,0.75), col="red")
lines(rep(kvant[5], 2), c(-1,0.9), col="red")
text(rep(2000,5), c(0.1,0.25,0.5,0.75,0.9)+0.02, labels=c(0.1,0.25,0.5,0.75,0.9), col="blue")
```

Funkce `ecdf` počítá empirickou distribuční funkci. Tu si pak můžeme vykreslit (funkce `plot` nebo `plot.stepfun` – porovnejte výsledky) nebo můžeme chtít její hodnotu v nějakém bodě. Např. pro 4000 zjistíme výsledek `ecdf(hmot)(4000)`. Připomeňte si, co touto kvantitou odhadujeme.

12. Další charakteristikou variability, založenou na kvantilech, je mezikvartilové rozpětí. Můžeme si ho spočítat pomocí funkce `IQR`. Toto číslo je pro popis dat často užitečnější než směrodatná odchylka. Uložte si ho do proměnné `iqr`.
13. Vyzkoušejte si výpočet decilů a percentilů

```
quantile(hmot, prob = seq(0, 1, by=0.1)) # decily
quantile(hmot, prob = seq(0, 1, by=0.01)) # percentily
```

Ujistěte se, že rozumíte uvedenému zápisu.

POPISNÉ GRAFY

14. Tzv. krabicový graf nám graficky znázorňuje některé popisné statistiky a také nám dává určitou představu o tvaru zkoumaného rozdělení

```
boxplot(hmot, ylab="Porodni hmotnost [g]")
```

Pomocí porovnání s popisnými statistika hmotnosti `summary(hmot)` zkuste přijít na to, co je na grafu znázorněno.

Příkaz `abline(h=3140)` nám např. vykreslí horizontální čáru s y souřadnicí 3140. Ověřte tímto způsobem, že v tomto našem případě horní a dolní „fousy“ odpovídají hodnotám $Q_3 + 1.5 IQR$ a $Q_1 - 1.5 IQR$.

Proč právě 1.5 násobek? V případě výběru z $N(0, 1)$ by mezi fousy měla ležet taková proporce dat:

```
pnorm(qnorm(0.75)+1.5*(qnorm(0.75)-qnorm(0.25))) -
  pnorm(qnorm(0.25)-1.5*(qnorm(0.75)-qnorm(0.25)))
```

V případě jiných rozdělení už to však neplatí!

15. Kdybychom chtěli vědět, které hodnoty leží mimo fousy a jaká je jejich relativní četnost:

```
bobj = boxplot(hmot)
sort(bobj$out)
length(bobj$out)/n
```

Proměnná `bobj` je typu `list`. Pomocí funkce `names(bobj)` se můžeme nechat vypsat její složky. Vyzkoušejte a nechte si vypsat

```
bobj$stat
```

16. Samostatně si vykreslete boxploty počtů bodů z písemky. Co jsou zde fousy boxplotu? Jak tedy obecně popíšeme vzhled boxplotu?
17. Nyní se budeme zabývat histogramem, který nám slouží jako odhad hustoty rozdělení.

```
hist(hmot)
hist(hmot, prob = TRUE, xlab = "Porodni hmotnost [g]", main = "Histogram")
```

18. U histogramu je poměrně zásadní počet uvažovaných intervalů. Porovnejte:

```
par(mfrow=c(1,2));
hist(hmot, prob = TRUE, breaks=seq(1750, 5100, by=10), main = "Delka intervalu 10");
hist(hmot, prob = TRUE, breaks=seq(1500, 5500, by=1000), main = "Delka intervalu 1000");
par(mfrow=c(1,1));
```

19. Funkce `hist` používá následující výpočet intervalů histogramu: Nejprve se použije tzv. Sturgesovo pravidlo, které říká, že optimální počet intervalů je roven horní celé části z $\log_2(n)$ plus jedna. Pak se použije funkce, která vytvoří „hezké intervaly“

```
(k <- ceiling(1+log2(length(hmot))))
pretty(hmot, k)
```

Srovnáme tento výsledek s tím, co dělá funkce `hist`. Opět si uložíme celý objekt (typu `list`) do proměnné `hobj` a podíváme se na jeho složky.

```
hobj <- hist(hmot, prob = TRUE)
hobj$breaks
hobj$counts
```

Spočtete si pomocí funkce `length` kolik tedy máme v histogramu sloupců.

20. Porovnáme histogram našich dat s hustotou normálního rozdělení, které by mělo střední hodnotu rovnou průměru našich dat a směrodatnou odchylku rovnou výběrové směrodatné odchylce.

```
xbar <- mean(hmot)
smoch <- sd(hmot)
xgrid <- seq(xbar - 3.5*smoch, xbar + 3.5*smoch, length = 500)
fxgrid <- dnorm(xgrid, mean = xbar, sd = smoch)
hist(hmot, prob = TRUE, xlab = "Porodni hmotnost [g]", main = "")
lines(xgrid, fxgrid, col = "red", lwd = 2)
```

21. V budoucnu nás bude často zajímat, zda můžeme data považovat za náhodný výběr z normálního rozdělení. Z grafických metod můžeme použít výše uvedené srovnání histogramu s hustotou, ale z něho někdy nevidíme velmi dobře, jak nám normální rozdělení sedí „v krajích“. Proto je vhodnější se podívat se na tzv. Q-Q graf.

```
qqnorm(hmot, cex=0.2)
qqline(hmot)
```

Q-Q graf obecně srovnává výběrové kvantily spočtené z dat (osa y) s teoretickými kvantily nějakého rozdělení. Zde v případě normálního Q-Q grafu odpovídají y souřadnice přímo uspořádaným datům a x -ové souřadnice kvantilům standardizovaného normálního rozdělení. V případě, že data pocházejí z normálního rozdělení, tak by body měly ležet na přímce. Umíte to teoreticky zdůvodnit?

22. Vykreslíme si Q-Q graf počtu bodů z písemky (bez rozlišení) a necháme si vypsat souřadnice bodů

```
z=c(x,y)
qqnorm(z)
qqline(z)
```

Pro $n > 10$ jsou kvantily na ose x počítané na hladině $\frac{i-1/2}{n}$ pro $i = 1, \dots, n$. Viz také funkce `ppoints`.

SAMOSTATNÁ PRÁCE

1. Spočítejte si základní popisné statistiky (charakteristiky polohy i variability) pro počet bodů z písemky (bez rozlišení, zda jde o chlapce nebo dívku).
2. Spočítejte výběrový 95% kvantil podle definice z přednášky. Kolika studentů má více bodů než je hodnota tohoto kvantilu?
3. Spočítejte si výběrový 90% kvantil pomocí funkce `quantile` i přímo podle definice z přednášky. Máte shodný výsledek?
4. Vykreslete si empirickou distribuční funkci počtu bodů. Vyznačte v ní 90% kvantil.
5. Nakreslete si histogram počtu bodů. Přidejte do obrázku hustotu normálního rozdělení s vhodnými parametry.
6. Vykreslete si vedle sebe do jednoho obrázku boxplot počtu bodů chlapců a dívek.
7. Vytvořte „ručně“ Q-Q graf počtu bodů z písemky, tj. bez použití funkcí `qqnorm` a `qqline`.
8. Projděte si samostatně cvičení v souboru `dc-cviceni8.R`.