Cvičení z biostatistiky 02

Poslední úprava dokumentu: 28. února 2024.

Popisná statistika pro jednu proměnnou

Rozcvička

Vytvořte si libovolný vektor o pěti hodnotách a spočtěte průměr a směrodatnou odchylku.

1 Úvod

1) Spusťte RStudio z nabídky.

2) Nastavte si pracovní adresář.

- ✧ Buď v Menu nahoře:
  Session ➥ **Set Working Directory** ➥ **Choose Directory . . .**
  a vyberte složku `biostat`, kterou jste si vytvořili minule.
- ✧ Nebo příkazem:

```
setwd("popis_cesty/biostat")
```

3) Otevřte si nový skriptový soubor

- ✧ Buď v Menu nahoře:
  File ➥ **New File** ➥ **R Script**
- ✧ nebo kliknutím na ikonku papíru se zeleným plus vlevo nahoře a zvolením `R Script`,
- ✧ nebo klávesovou zkratkou `Ctrl+Shift+N`.

Nebo použijte skript z minula:

- ✧ V Menu nahoře:
  File ➥ **Open File**
  a vyberte příslušný soubor.

4) Pokud jste se rozhodli pro vytvoření nového skriptového souboru, nezapomeňte si ho uložit:

- ✧ **File** ➥ **Save as. . .**
- ✧ nebo kliknutím na symbol diskety,
- ✧ nebo klávesovou zkratkou `Ctrl+S`.

5) Nemáte-li z minula uložen datový soubor Deti23 v R formátu (soubor `Deti23.RData`), vytvořte si ho dle návodu v předchozím pracovním listu "Seznámení s R". Máte-li data již načtena, můžete tento i další krok přeskočit a přejít rovnou k bodu 7).

6) Načtěte data Deti23 do RStudia:

✧ Buď v Menu nahoře:

File ➥ **Open file...**

a potvrdit soubor `Deti23.RData` ze složky `biostat\data`.
Do you want to load the R data file into the global environment?

Yes

✧ Nebo příkazem:

```
load("data/Deti23.RData")
```

7) Prohlédněte si data a ujistěte se, že jsou správně načtena. Data si můžete zobrazit buď poklepáním na jejich název v pravém horním okně, nebo příkazem

```
View(Deti23)
```

8) Zajistěte si přímý přístup k jednotlivým proměnným datového souboru Deti23. Ze skriptového okna spust'te příkaz:

```
attach(Deti23)
```

2 Jedna kvantitativní proměnná

2.1 Numerické hodnoty popisných statistik

1) Připomeňte si z přednášky pojmy: průměr, směrodatná odchylka, medián, dolní a horní kvartil, kvantily. Uměli byste je pro konkrétní data vypočítat ručně pomocí kalkulačky?

2) Spočtěte výše uvedené popisné statistiky pro věk matek.

```
summary(vekMatky)          # prumer a nektere "dulezite" kvantily

mean(vekMatky)             # vyberovy prumer
sd(vekMatky)               # vyberova smerodatna odchylka
var(vekMatky)              # vyberovy rozptyl
median(vekMatky)           # median
min(vekMatky)              # minimum
max(vekMatky)              # maximum
quantile(vekMatky, prob=c(0, 0.25, 0.50, 0.75, 1.00))
                           # vybrane kvantily
length(vekMatky)           # delka vektoru (pocet pozorovani)
```

3) Zkusme zcela ručně (dle „vzorečku“ z přednášky) spočítat medián pro věk otce.

✧ Bude potřeba setřídit pozorované věky otců. Toho lze docílit příkazem:

```
sort(vekOtce)
```

✧ Zde máme 23 pozorování. Kolikáté pozorování v setříděné posloupnosti věků je tedy mediánem?

✧ Jak by se určil medián, kdybychom měli sudý počet pozorování?

4) Základní popisné statistiky pro všechny proměnné datového souboru lze získat následujícím příkazem:

```
summary(Deti23)
```

✧ Povšimněte si jiného tvaru výstupu u proměnné Pohlavi. Jedná se o kvalitativní proměnnou, kterými se budeme zabývat v některém z příštích cvičení.

✧ Doporučuji po každém načtení dat spočítat vždy tyto statistiky a alespoň rychlým pohledem si je prohlédnout. Odhalíte tak hrubé chyby vzniklé při přepisu dat, nebo případné problémy s načtením dat.

2.2 Grafické znázornění popisných statistik

1) Mnohé popisné statistiky kvantitativní proměnné lze přehledně znázornit pomocí krabičkového grafu. Nakresleme krabičkový graf pro věk matky.

```
boxplot(vekMatky)
```

Příkaz lze modifikovat a zkrášlovat obrázek.

```
boxplot(vekMatky, ylab="Vek matky", col="steelblue")
boxplot(vekMatky, xlab="Vek matky", col="sandybrown", horizontal=TRUE)
```

2) Samostatně nakreslete krabičkový graf pro věk otce. V grafu se objeví jedno odlehlé pozorování

3) Samostatně spočtěte popisné statistiky pro váhu a délku dítěte.

4) Samostatně nakreslete krabičkové grafy pro váhu a délku dítěte.

5) Připomeňme, že

$$\text{dobaPlnolet} = \text{vekMatky} - 18$$
$$\text{delka.m} = \text{delka}/100$$

✧ **Vliv posunutí:** Jaký je vztah mezi popisnými statistikami pro proměnné vekMatky a dobaPlnolet?

✧ **Vliv změny měřítka:** Jaký je vztah mezi popisnými statistikami pro proměnné delka a delka.m?

3 Jedna kvalitativní proměnná

3.1 Tabulka absolutních a relativních četností

1) Spočítejte základní popisné statistiky pro proměnnou hcd (počet onemocnění horních cest dýchacích během prvního roku života).

```
summary(Deti23$hcd)
```

2) Veličina hcd je sice kvantitativní, ale nabývá pouze celočíselných hodnot 0–9. Podívejme se, kolikrát se která hodnota vyskytuje. Ze skriptového okna spusťte příkaz

```
table(Deti23$hcd)
```

3.2 Tvorba `factorů`

3) Z pohledu R je veličina hcd kvantitativní/číselná (těm se v R říká `numeric` nebo `integer`). Zkuste

```
class(Deti23$hcd)
```

Někdy se ale může hodit z ní udělat veličinu kategoriální (těm se v R říká `factor`), jejíž kategorie budou dány počtem onemocnění.

✧ Buď

```
Deti23 <- transform(Deti23, fhcd = factor(hcd))
```

✧ nebo

```
Deti23$fhcd <- as.factor(Deti23$hcd)
```

✧ Že se skutečně jedná o `factor`ovou proměnnou si můžete ověřit pomocí:

```
class(Deti23$fhcd)
```

4) Řekněme, že nás ale takto jemné dělení nemocnosti nezajímá a stačí nám rozlišovat děti pouze na *nikdy* nemocné (hcd $= 0$), *jednou* nemocné (hcd $= 1$) a *opakovaně* nemocné (hcd > 1). Jinými slovy, chceme z proměnné hcd vytvořit kvalitativní proměnnou o třech úrovních *nikdy/jednou/opakovaně*.

✧ Vytvořme nyní proměnnou fhcd3, jež bude `factor`em vytvořeným z hcd výše popsaným způsobem.

```
pom.hcd3 <- 0*(Deti23$hcd == 0) + 1*(Deti23$hcd == 1) + 2*(Deti23$hcd > 1)
Deti23 <- transform(Deti23, fhcd3 = factor(pom.hcd3,
labels=c("nikdy", "jednou", "opakovane")))
```

5) Spočtěme nyní absolutní a relativní četnosti jednotlivých úrovní fhcd3.

```
table(Deti23$fhcd3)
prop.table(table(Deti23$fhcd3))
round(prop.table(table(Deti23$fhcd3)) * 100, 2)
```

3.3 Grafické zobrazení absolutních a relativních četností

1) Zobrazme absolutní četnosti pomocí sloupcového grafu.

```
barplot(table(Deti23$fhcd3), xlab="Nemocnost", ylab="Cetnost",
col=terrain.colors(3))
```

2) Zobrazme relativní četnosti pomocí sloupcového grafu.

```
barplot(prop.table(table(Deti23$fhcd3)),
xlab="Nemocnost", ylab="Relativni cetnost", col=topo.colors(3))
```

3) Zobrazme relativní četnosti pomocí koláčového grafu.

```
pie(table(Deti23$fhcd3), main="Nemocnost", col=heat.colors(3))
```

4 Samostatná práce

1) Vytvořte novou proměnnou nazvanou fporadi2, jež bude odlišovat prvorozence od ostatních dětí. Jednotlivé kategorie si vhodně označte.

2) Spočtěte absolutní a relativní četnosti prvorozenců a dětí, jež mají sourozence.

3) Graficky znázorněte absolutní a relativní četnosti prvorozenců a dětí, jež mají sourozence.

5 Konec práce

Rozšířený datový soubor Deti23 (s novými proměnnými fhcd3 a fporadi2) uložte na síťovém disku ve svém adresáři v R formátu jako soubor Deti23.RData (tj. přepište dřívější Deti23.RData novou verzí) do podsložky `data`.

```
save(Deti23, file = "data/Deti23.RData")
```