Cvičení z matematické statistiky 06

Poslední úprava dokumentu: 25. března 2025.

Normální rozdělení: centrální limitní věta, normální diagram (QQ plot), interval spolehlivosti pro střední hodnotu normálně rozděleného výběru

---

1 Centrální limitní věta

Teorii k centrální limitní větě si můžete prostudovat v textu *CLV.pdf*, který najdete na mých stránkách.

Dále si stáhněte skript „ukazkaCLT.R“ a uložte si ho do svého pracovního adresáře. Skript si načtěte do R pomocí příkazu:

```
source("ukazkaCLT.R")
```

Nejprve si prohlédněte, jak vypadají hustoty některých spojitých rozdělení, o kterých jste slyšeli na přednášce.

```
ukazkaCLT(n=0)
```

Jako ilustraci centrální limitní věty si prohlédněte, jak se chová rozdělení výběrových průměrů z výše uvedených rozdělení, měníme-li rozsah výběru n.

```
ukazkaCLT(n=1)
```

- ✧ Volte postupně n: 1, 2, 5, 20, 100
- ✧ Povšimněte si, že měřítka na y-ové ose pro jednotlivá rozdělení jsou různá.
- ✧ Důvodem je fakt, že limitní (normální) rozdělení mají různé parametry pro jednotlivá rozdělení.

Galtonova deska

`https://www.youtube.com/watch?v=EvHiee7gs9Y`

Příklad se zaokrouhlováním čísel

Bylo sečteno 300 čísel zaokrouhlených na jedno desetinné místo. Vyšetříme chybu tohoto součtu vzniklou zaokrouhlováním sčítanců. Předpokládejme, že zaokrouhlovací chyby (označme si je $X_1, X_2, X_3, \ldots, X_{300}$) jednotlivých sčítanců jsou nezávislé náhodné veličiny s rovnoměrným rozdělením na intervalu (-0.05, 0.05).

1) Označme si jako X_k chybu k-tého sčítance. Víme, že $X_k \sim R(-0.05, 0.05)$. Pak nutně platí, že $EX_k = 0$ a $var(X_k) = 1/1200$ (připomeňte si vzoreček pro střední hdonotu a rozptyl rovnoměrného rozdělení).

2) Chyba součtu je náhodná veličina $Y = \sum_{k=1}^{300} X_k$.

3) Normovaná chyba je náhodná veličina

$$Z = \frac{Y}{\sqrt{300/1200}} = 2Y$$

a dle Centrální limitní věty má tato veličina rozdělení $N(0,1)$.

4) Pravděpodobnost, že chyba součtu nepřekročí v absolutní hodnotě dané číslo $\varepsilon > 0$ je

$$P(|Y| \leq \varepsilon) = P(|Z| \leq 2\varepsilon) = P(-2\varepsilon \leq Z \leq 2\varepsilon) = \Phi(2\varepsilon) - (1 - \Phi(-2\varepsilon)) = 2\Phi(2\varepsilon) - 1,$$

kde Φ značí distribuční funkci rozdělení $N(0,1)$. Např. pro $\varepsilon = 1$ tato pravděpodobnost rovna 0.9545.

2 Dataset `Deti`

Datový soubor Deti.RData obsahuje údaje o 99 matkách a jejich dětech (data získala Mgr. P. Hajná při přípravě své diplomové práce). Najdeme zde následující proměnné:

`pocet.deti`	o kolikáté dítě matky jde;
`Vzdelani`	vzdělání matky (*základní/maturita/VŠ*);
`por.hmotnost`	porodní hmotnost dítěte (g);
`por.delka`	porodní délka dítětete (cm);
`hmotnost`	hmotnost dítěte ve 24. týdnu po porodu (g);
`delka`	délka dítěte ve 24. týdnu po porodu (cm);
`vyska.m`	výška matky (cm);
`vyska.o`	výška otce (cm);
`vek.m`	věk matky (roky);
`vek.o`	věk otce (roky);
`Dudlik`	dítě mělo dudlík (0=*ne*, 1=*ano*);
`Plan`	dítě bylo podle matky plánované (0=*ne*, 1=*ano*);
`Porodnice`	rozlišení dvou porodnic, v nichž byla pořizována data (1=*Praha*, 2=*okresní město*);
`hoch`	dítě je hoch (0=*ne*, 1=*ano*);
`Hoch`	pohlaví dítěte (*dívka/hoch*).

Načtěte si data Deti a zajistěte si přímý přístup k jednotlivým proměnným tohoto datového souboru

```
load("Deti.RData")
attach(Deti)
```

3 Zkoumání normality dat - QQ plot

Shodu rozdělení dat s předpokládaným rozdělením lze graficky nejlépe posoudit pomocí kvantil-kvantilového grafu (angl. *QQ plot*, což je zkratka pro *quantile-quantile plot*). Chceme-li tedy posoudit, zda by naše data mohla pocházet z normálního rozdělení, je nejlepším obrázek právě QQ plot, v případě normálního rozdělení nazývaný *normální diagram*.

V normálním diagramu jsou na ose x teoretické kvantily normálního rozdělení $N(0,1)$ (ty jsou známy) a na ose y jsou výběrové kvantily vypočítané z příslušných dat (výběrové kvantily jsme počítali na předchozích cvičeních pomocí příkazu `quantile`). Počet vykreslených kvantilů je roven počtu našich dat. Pocházejí-li naše data z normálního rozdělení (ne nutně jen $N(0,1)$), měly by si teoretické a výběrové kvantily odpovídat a body v normálním diagramu by měly ležet na přímce.

Pokud data z normálního rozdělení nepocházejí, budou se body v normálním diagramu kolem přímky různě vlnit, popř. utíkat na koncích.

Nenechte se zmást tím, že výběrové kvantily v normálním diagramu porovnáváme s teoretickými kvantily zrovna $N(0,1)$ rozdělení. Pokud by naše data pocházela z normálního rozdělení s nějakou jinou střední hodnotou a jiným rozptylem (tj. obecně $N(\mu, \sigma^2)$), bude mít přímka z bodů akorát jiný sklon, ale její tvar zůstane zachován.

Posouzení normality na základě normálního diagramu je samozřejmě zcela subjektivní. Později budeme normální rozdělení dat rigorózně testovat pomocí **Shapiro-Wilkova testu**.

Nejprve budeme analyzovat hmotnost dítěte ve 24. týdnu po narození (proměnná hmotnost).

5) Nakreslete **normální diagram** pro hmotnost dítěte ve 24. týdnu po narození (proměnná hmotnost).

```
qqnorm(hmotnost)
qqline(hmotnost)
```

✧ Ukazuje obrázek na nenormalitu rozdělení hmotnosti dětí?

6) Obrázek samozřejmě můžete dále zkrášlovat:

```
qqnorm(hmotnost, pch=16, col="red", xlab="Teoreticke kvantily N(0, 1)",
ylab="Kvantily hmotnosti dětí", main="Normalni QQ graf (hmotnost)")
qqline(hmotnost, lty=6, col="darkblue")
```

4 Interval spolehlivosti pro střední hodnotu normálně rozděleného výběru

Statistika se snaží na základě několika vybraných jedinců usoudit něco o celé populaci. Formuluje různá teoretická tvrzení (hypotézy) o celé populaci a snaží se je zodpovědět pomocí výběrových charakteristik. Od tohoto cvičení už tedy bude velmi důležité důsledně rozlišovat populační charakteristiky (populační průměr/střední hodnota, populační rozptyl, atd.) od jejich výběrových protějšků (výběrový průměr, výběrový rozptyl, atd.).

Přečtěte si, prosím, text *Konfidenční intervaly*, který najdete na mých stránkách, nebo si příslušnou teorii připomeňte z přednášky.

Spusťte si RStudio a nastavte si pracovní adresář. Dále si otevřte nový skript a uložte si jej.

4.1 Interpretace intervalu spolehlivosti

1) Hodnota IQ v populaci se řídí rozdělením $\mathcal{N}(100,\, 15^2)$. Nalezněte nejkratší interval, ve kterém leží IQ 50 % populace.

```
curve(dnorm(x,100,15),from=60,to=140,lwd=2)   # graf hustoty
abline(h=0,col="grey") # vykresleni prislusne plochy
dolni_kv <- 0.25
horni_kv <- 0.75
lines(qnorm(c(dolni_kv,horni_kv),100,15),c(0,0),lwd=3,col="green")
xx <- seq(qnorm(dolni_kv,100,15),qnorm(horni_kv,100,15),length.out=101)
polygon(c(qnorm(dolni_kv,100,15),xx,qnorm(horni_kv,100,15)),
        c(0,dnorm(xx,100,15),0),col="lightgreen")
```

2) Nalezněte nejkratší interval, ve kterém leží IQ 95 % populace.

3) V praxi hodnotu $\mu = 100$ samozřejmě neznáme a chceme ji odhadnout z pozorovaných dat.

✧ Nagenerujte si náhodný výběr o rozsahu 50.

```
vyber <- rnorm(50, 100, 15)
```

✧ Jaký je bodový odhad μ založený na našem výběru?

✧ Jaké rozdělení má průměrné IQ spočítané z výběru o rozsahu n?
Výběrový průměr $\bar{X}$ spočítaný z výběru z $\mathcal{N}(\mu, \sigma^2)$ o rozsahu n má rozdělení $\mathcal{N}(\mu, \sigma^2/n)$.

✧ V jakém intervalu tato průměrná hodnota leží s pravděpodobností 0.95?
Výběrový průměr $\bar{X}$ leží s pravděpodobností 0.95 s intervalu $(q_{0.025}, q_{0.975})$, kde $q_{0.025}$ je 2.5% kvantil rozdělení $\mathcal{N}(\mu, \sigma^2/n)$.

```
n <- 50
shp <- 100                  #  střední hodnota průměru
sop <- 15/sqrt(n)          # směrodatná odchylka průměru
curve(dnorm(x,shp,sop),from=60,to=140,lwd=2)   # graf hustoty
abline(h=0,col="grey") # vykresleni prislusne plochy
dolni_kv <- 0.025
horni_kv <- 0.975
lines(qnorm(c(dolni_kv,horni_kv),shp,sop),c(0,0),lwd=3,col="green")
xx <- seq(qnorm(dolni_kv,shp,sop),qnorm(horni_kv,shp,sop),length.out=101)
polygon(c(qnorm(dolni_kv,shp,sop),xx,qnorm(horni_kv,shp,sop)),
        c(0,dnorm(xx,shp,sop),0),col="lightgreen")
```

✧ Výraz pro tento interval upravte tak, abyste dostali interval spolehlivosti pro μ.
Z předchozího bodu víme, že $P(\bar{X} \in (q_{0.025}, q_{0.975})) = 0.95$. Poznamenejme, že kvantil $q_{0.975}$ (což je 97.5% kvantil $\mathcal{N}(\mu, \sigma^2/n)$) lze zapsat jako $\mu + \frac{\sigma}{\sqrt{n}} u_{0.975}$, kde $u_{0.975}$ je 975% kvantil rozdělení $\mathcal{N}(0, 1)$. Analogicky $q_{0.025}$ lze zapsat jako $\mu + \frac{\sigma}{\sqrt{n}} u_{0.025}$. Jelikož rozdělení $\mathcal{N}(0, 1)$ je symetrické kolem nuly, tak $u_{0.025} = -u_{0.975}$. Dostáváme tedy:

$$P\left(\mu - \frac{\sigma}{\sqrt{n}} u_{0.975} \leq \bar{X} \leq \mu + \frac{\sigma}{\sqrt{n}} u_{0.975}\right) = 0.95.$$

Nyní nerovnosti uvnitř upravíme tak, aby bylo uprostřed μ:

$$\begin{aligned}
\mu - \frac{\sigma}{\sqrt{n}} u_{0.975} \leq \bar{X} \leq \mu + \frac{\sigma}{\sqrt{n}} u_{0.975} \quad & / - \mu \\
-\frac{\sigma}{\sqrt{n}} u_{0.975} \leq \bar{X} - \mu \leq \frac{\sigma}{\sqrt{n}} u_{0.975} \quad & / - \bar{X} \\
-\bar{X} - \frac{\sigma}{\sqrt{n}} u_{0.975} \leq -\mu \leq -\bar{X} + \frac{\sigma}{\sqrt{n}} u_{0.975} \quad & / * (-1) \\
\bar{X} + \frac{\sigma}{\sqrt{n}} u_{0.975} \geq \mu \geq \bar{X} - \frac{\sigma}{\sqrt{n}} u_{0.975} &
\end{aligned}$$

Dostáváme tedy, že

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} u_{0.975} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} u_{0.975}\right) = 0.95,$$

což znamená, že interval

$$(\bar{X} - \frac{\sigma}{\sqrt{n}} u_{0.975}, \bar{X} + \frac{\sigma}{\sqrt{n}} u_{0.975}) \tag{1}$$

pokrývá μ s pravděpodobností 0.95, a proto ho nazýváme 95% interval spolehlivosti pro μ.

4) Z mých stránek či z mé složky na disku `V:` si stáhněte funkci „ukazkaCI.R“ a uložte si ji do svého pracovního adresáře. Dále funkci načtěte do R:

```
source("ukazkaCI.R")
```

5) Podívejme se nyní, co by mohlo nastat, kdyby M výzkumníků provádělo postupně měření IQ na n různých lidech. Každý výzkumník by každý počítal (svůj) interval spolehlivosti.

```
ukazkaCI(mean.sim=100, sd=15, n=10, reps=50, conf.level=0.95, method="z")
```

Význam jednotlivých parametrů:

mean.sim = populační průměr	100
sd = populační směrodatná odchylka	15
n = rozsah výběru (naše n)	volte postupně 10, 25, 50, 100
reps = počet výběrů (naše M)	volte 50 nebo 100
conf.level = spolehlivost	0.95
method = použitá metoda výpočtu	"z" odpovídá známé populační sd "t" odpovídá neznámé populační sd "both" uvidíme porovnání obou metod

Povšimněte si následujícího:

- ✧ Ne každý spočtený interval se trefí do skutečné hodnoty $\mu = 100$. Přibližně 5 % výzkumníků obrželo interval spolehlivosti, který neobsahuje skutečné $\mu = 100$.
- ✧ Zvětšujete-li n (rozsah výběru), intervaly se zkracují (jejich délka je přímo úměrná $1/\sqrt{n}$).
- ✧ Při neznámém σ jsou intervaly (při jinak shodném n) o něco širší než při známém σ.
- ✧ Čím vyšší spolehlivost interval má, tím je širší.

4.2 Interval spolehlivosti pro neznámé σ (hmotnost ve 24. týdnu)

1) Znázorněte graficky hlavní charakteristiky rozdělení hmotnosti. Nakreslete krabičkový graf, histogram a normální QQ graf pro hmotnost.

```
boxplot(hmotnost)     # krabickovy graf
hist(hmotnost)        # histogram
qqnorm(hmotnost)      # normalni diagram
qqline(hmotnost)
```

2) Domníváte se, že je smysluplné předpokládat, že rozdělení hmotnosti ve 24. týdnu je normální (s neznámou střední hodnotou μ a neznámým rozptylem σ^2)?

- ✧ Ano. Histogram má docela symetrický tvar připomínající Gaussovu křivku (hustotu normálního rozdělení). Body na normálním diagramu celkem obstojně kopírují přímku. Lze tedy předpokládat, že data hmotností pocházejí z normálního rozdělení $N(\mu, \sigma^2)$, kde parametry μ a σ^2 bohužel neznáme.
- ✧ Střední hodnota μ představuje populační průměr hmotností, nebo-li střední hmotnost. Je to neznámá konstanta, kterou se pokusíme odhadnout.
- ✧ Později budeme normalitu dat testovat rigorózně pomocí Shapiro-Wilkova testu.

3) Spočtěte základní popisné statistiky pro hmotnost.

```
summary(hmotnost)     # charakteristiky polohy
sd(hmotnost)          # smerodatna odchylka
```

4) Odhadněme bodově i intervalově (s 95% spolehlivostí) střední hmotnost (tj. populační průměr hmotností) dětí ve 24. týdnu při předpokladu normálního rozdělení.

✧ Bodový odhad již máme - je to výběrový průměr (v dalším značeno jako $\bar{x}$).

```
mean(hmotnost)
```

✧ Jelikož jsme zjistili, že data hmotností by mohla pocházet z normálního rozdělení, můžeme pro intervalový odhad μ použít obdobu vzorce (1). Jen je třeba v něm nahradit populační směrodatnou odchylku σ za výběrovou směrodatnou odchylku $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ a dále nahradit kvantil normovaného normálního rozdělení $u_{0.975}$ za kvantil t-rozdělení (s $n-1$ stupni volnosti) $qt_{n-1}(0.975)$:

$$\left(\bar{x} - \frac{s_x}{\sqrt{n}} \cdot qt_{n-1}(0.975)\ ,\ \bar{x} + \frac{s_x}{\sqrt{n}} \cdot qt_{n-1}(0.975)\right)$$

kde

$$\begin{aligned} \bar{x} &= \text{výběrový průměr} = \text{bodový odhad pro } \mu, \\ s_x &= \text{výběrová směrodatná odchylka}, \\ n &= \text{rozsah výběru (tj. počet pozorování)}, \\ \frac{s_x}{\sqrt{n}} &= \text{odhad chyby odhadu}, \\ t_{n-1}(1-\alpha/2) &= \text{kvantil Studentova t-rozdělení}. \end{aligned}$$

✧ Zvolili jsme tzv. oboustrannou verzi intervalu spolehlivosti, která se používá nejčastěji.

✧ Nejprve zkusíme spočítat interval spolehlivosti ručně, podle vzorečku výše.

```
prum <- mean(hmotnost)      # prumer
std.dev <- sd(hmotnost)     # smer. odchylka hmotnosti
n <- length(hmotnost)       # pocet pozorovani
se.prum <- std.dev/sqrt(n) # odhad chyby odhadu
qq <- qt(0.975, df=n-1)     # kvantil t-rozdeleni
puldelka <- se.prum * qq    # polovina delky int. spol.
CI.dolni <- prum - puldelka
CI.horni <- prum + puldelka
CI.rucne <- c(CI.dolni, CI.horni)
CI.rucne
```

5) Intervalové odhady (intervaly spolehlivosti) souvisejí úzce s testováním hypotéz. V R proto zpravidla nenajdeme zvláštní procedury pro výpočet intervalu spolehlivosti. Nicméně procedury určené primárně k testování hypotéz produkují též související interval spolehlivosti. S odhadem μ v náhodném výběru z $\mathcal{N}(\mu, \sigma^2)$ souvisí jednovýběrový t-test. Odsud tedy postup výpočtu intervalu spolehlivosti pro μ v R:

```
t.test(hmotnost)
```

případně:

```
t.test(hmotnost, conf.level=0.95)
```

kde v argumentu `conf.level` lze nastavit i jinou spolehlivost.

✧ Ve výstupu si v tuto chvíli všímejte pouze posledních pěti řádků, zbytek ignorujte.

✧ 95% interval spolehlivosti pro populační hmotnost tedy je: (7521.0 , 7858.1)

✧ Vyšlo to samé jako při ručním výpočtu výše?

6) Znovu si uvědomme, že 95% interval spolehlivosti je interval, ve kterém s danou spolehlivostí (95%) najdeme skutečnou (nám skrytou) hodnotu střední hodnoty (**populačního** průměru), kterou zde značíme jako μ.

```
        One Sample t-test

data:  hmotnost
t = 90.534, df = 98, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 7520.974 7858.077
sample estimates:
mean of x
 7689.525
```

95% interval spolehlivosti

průměr (bodový odhad)

5 Konec práce

Než zavřete všechna okna, nezapomeňte si uložit poslední změny ve skriptovém souboru:
File ➥ **Save**
nebo klávesovou skratkou `Ctrl-s`.