Cvičení z biostatistiky **05**

Poslední úprava dokumentu: 21. března 2024.

Interval spolehlivosti pro střední hodnotu normálně rozděleného výběru, úvod do testování hypotéz (jednovýběrový t-test)

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.

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.

2 Hmotnost ve 24. týdnu (interval spolehlivosti pro neznámé σ)

Načtěte si dříve uložená data Kojeni, se kterými jsme se seznámili v pracovním listu 03 a zajistěte si přímý přístup k jednotlivým proměnným tohoto datového souboru

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

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

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.

```
        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)

✧ 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.

3 Hmotnost chlapců ve 24. týdnu

3.1 Interval spolehlivosti

V situaci, kdy by pohlaví mělo vliv na hmotnost ve 24. týdnu, vypovídá výše spočtený interval o populačním průměru pouze tehdy, když poměr chlapců a dívek v datech odpovídá poměru chlapců a dívek v populaci. Toto by mělo platit za předpokladu, že děti do studie byly vybírány skutečně náhodně. V každém případě bude zajímavé zjistit, jakých hodnot nabývá populační průměr (střední hodnota) hmotnosti u chlapců, resp. dívek. Odborně řečeno budeme odhadovat podmíněnou (pohlavím) střední hodnotu hmotnosti.

1) Pro počítání intervalů spolehlivosti bude potřeba vytvořit podmnožiny dat obsahující pouze dívky, resp. chlapce. V dalším budeme střídavě pracovat se všemi dětmi, pouze s chlapci, či pouze s dívkami. Bude proto vhodnější odpojit přímý přítup ke všem proměnným, aby nedošlo k nedorozumění. Spusťte ze skriptového okna:

```
detach(Kojeni)
```

2) Nyní přistupme k vytvoření podmnožiny chlapců, kterou uložíme do datové tabulky KojeniH.

```
KojeniH <- subset(Kojeni, Hoch == "hoch")
```

3) Nejprve spočtěme 95% interval spolehlivosti pro střední hmotnost hochů.

```
t.test(KojeniH$hmotnost)
```

4) Ještě bychom se měli podívat, zda pro rozdělení hmotnosti u chlapců i dívek lze předpokládat normální rozdělení. Nakreslete histogram a normální diagram (QQ graf) pro hmotnost chlapců.

```
hist(KojeniH$hmotnost)
qqnorm(KojeniH$hmotnost)
qqline(KojeniH$hmotnost)
```

✧Body na normálním diagramu se od přímky příliš neodchylují. Lze tedy předpokládat, že hmotnosti chlapců by mohly pocházet z normálního rozdělení.

5) Užitečné srovnání průměrů dvou populací poskytuje graf průměrů (Plot of means). Buď můžete použít funkci `plot_of_means.R`, kterou si můžete stáhnout ze složky `V:/turcicova` nebo z mé webové stránky, nebo můžete využít knihovnu `RcmdrMisc`.

```
install.packages("RcmdrMisc")
library(RcmdrMisc)                                # otevre knihovnu
plotMeans(Kojeni$hmotnost, Kojeni$Hoch, error.bars = "conf.int", level=0.95)
```

Body v grafu představují bodové odhady populační hmotnosti (= odhady střední hmotnosti) v obou skupinách, to jest výběrové průměry. „Fousy“ představují příslušné 95% intervaly spolehlivosti. Každý z těchto intervalů tedy s pravděpodobností 0.95 pokrývá skutečnou hodnotu populační hmotnosti pro dané pohlaví.

3.2 Jednovýběrový t-test (oboustranná alternativa)

6) Zajistěte si přímý přístup k chlapeckým hodnotám.

```
attach(KojeniH)
```

7) Naším úkolem bude rozhodnout, zda data podporují hypotézu, že střední (to jest populační) hmotnost chlapců ve 24. týdnu je 8 200 g. Přitom požadujeme, aby k neoprávněnému zamítnutí této hypotézy došlo s pravděpodobností nejvýše 0.05, tj. zvolili jsme si *hladinu testu* $\alpha = 0.05$, neboli 5 %.

✧ Abychom mohli tuto hypotézu (statisticky) otestovat, je potřeba si toto slovní zadání přepsat do řeči matematiky...

8) Za náhodnout veličinu X si označíme hmotnost náhodně vybraného chlapce. Z QQ diagramu, který jsme si vykreslili dříve, víme, že lze předpokládat, že $X \sim \mathcal{N}(\mu, \sigma^2)$, tj. že rozdělení hmotnosti chlapců je normální s nějakou neznámou střední hodnotou μ a neznámým rozptylem σ^2. Ze zadání úlohy je zřejmé, že chceme testovat hypotézu $\mathrm{H}_0 : \mu = 8\,200$ g. Jako alternativní hypotézu si zvolíme $\mathrm{H}_1 : \mu \neq 8\,200$ g.

✧ Za předpokladu normálního rozdělení (který je ale vždy potřeba ověřit) lze k testu této hypotézy využít *jednovýběrový t-test*, jehož princip si prosím připomeňte z přednášky, nebo z textu *Úvod do testování hypotéz* (dále jako *ÚTH*), který najdete na mých stránkách k 5. cvičení.

9) Jednovýběrový t-test (angl. one-sample t-test) se v R provede jednoduchým příkazem

```
t.test(hmotnost, mu=8200)
```

kde do argumentu „**mu=**“ se udává testovaná hodnota z nulové hypotézy, v našem případě tedy 8 200. Řecké μ se totiž v angličtině nazývá „mu“ [mju:]. Nyní se podívejme na jednotlivé údaje ve výstupu:

```
                 One Sample t-test
                                         počet stupňů volnosti (degrees of freedom) příslušného t-rozdělení
data:  hmotnost
hodnota testové statistiky T — t = -2.235, df = 48, p-value = 0.03011 — p-hodnota
alternative hypothesis: true mean is not equal to 8200 — tvrzení alternativní hypotézy
95 percent confidence interval:
 7684.391 8172.752 — 95% interval spolehlivosti pro skutečnou střední hodnotu
sample estimates:
mean of x
 7928.571 — výběrový průměr
```

✧ Jaký je závěr (zamítáme/nezamítáme H_0 na požadované hladině)?

Na základě výstupu z R můžeme o platnosti nulové hypotézy H_0 rozhodnout třemi způsoby (viz text *ÚTH*, str. 6, úplně dole):

(a) Podíváme se na hodnotu testové statistiky $T = -2.235$ a porovnáme ji s příslušným kvantilem[1] t-rozdělení $qt_{n-1}(1-\alpha/2) = qt_{48}(0.975) = 2.010635$ (kde n značí počet pozorování). Vidíme, že absolutní hodnota testové statistiky $|T| = 2.235$ je větší než tento kvantil (testová statistika překročila kritickou hodnotu), a tudíž zamítáme nulovou hypotézu H_0, že by střední (tj. populační) hmotnost chlapců byla 8 200 g.

(b) Pomocí p-hodnoty: Připomeňme, že p-hodnota je nejmenší hladina, na které bychom H_0 zamítli. Zde tedy bychom zamítli pro hladinu 0.03011 a vyšší. Naše zvolená hladina ze zadání příkladu je $\alpha = 0.05$ a jelikož $0.05 > 0.03011$, tak naše rozhodnutí opět je zamítnout H_0. Jednoduché pravidlo zní: „Je-li $\alpha >$ p-hodnota $\Rightarrow$ zamítáme H_0.“

[1] Standarně se kvantil t-rozdělení značí pouze $t_{n-1}(1-\alpha/2)$, prefix q jsem přidala pouze pro názornost, aby se zdůraznilo, že jde o kvantil. Snad to není naopak matoucí :-(

(c) Pomocí intervalu spolehlivosti: Víme, že 95% interval spolehlivosti obsahuje s pravděpodobností 0.95 skutečnou hodnotu střední hmotnosti. Je-li tedy 8200 g skutečná střední hmotnost, měla by v našem intervalu spolehlivosti (7684.391; 8172.752) ležet. Jelikož tam neleží, tak to asi nebude ta skutečná střední hmotnost. Opět tedy závěr zní - zamítnout H_0.

Všechny tři způsoby vždy vedou ke stejnému závěru, takže je jedno, který si vyberete. První způsob je poněkud nepraktický, neboť si musíte ručně spočítat kvantil $qt_{48}(0.975)$, a uvedli jsme ho spíš pro úplnost. V praxi je asi nejpoužívanější způsob (b), popř. (c).

✧ Jak byste formulovali svůj závěr bez použití výrazů „zamítáme/nezamítáme H_0“?
„Na 5% hladině jsme prokázali, že střední hmotnost je různá od 8200 g.“ (Ten rozdíl je statisticky významný).

✧ Jaká jiná hodnota μ při nulové hypotéze by vedla k opačnému závěru?
No nějaká, která leží uvnitř intervalu spolehlivosti. Tak třeba 8 000 g.

✧ Je potřeba znovu počítat, jestliže bychom požadovali přísnější 1% hladinu významnosti, resp. benevolentnější 10% hladinu významnosti?
NE. Na základě výstupu výše lze rozhodnout o platnosti H_0 i na jiné hladině než 5 %. Sice už nelze použít 95% interval spolehlivosti, ale můžeme se řídit p-hodnotou. Je-li p-hodnota menší než zvolená hladina (0.1 nebo 0.01), tak H_0 zamítáme.

Pokud bychom ale přece jen chtěli výstup procedury t-test přepočítat pro hladinu 10 % (tj. $\alpha = 0.1$), použili bychom:

```
t.test(hmotnost, mu=8200, conf.level = 0.9)
```

Argument `conf.level` udává spolehlivost a je rovna vždy $1 - \alpha$.

Ruční výpočet testové statistiky a p-hodnoty

10) Pojď'me si osvěžit klasické programování a vypočítat hodnotu testové statistiky a p-hodnoty ručně. Nejprve zkuste příkazy vymyslet sami a až pak se podívat na řešení níže.

✧ Hodnota testové statistiky. (Příkazy jsou v závorkách, aby se současně s uložením do proměnné vypsala i její hodnota.)

```
(n <- length(hmotnost))              # počet pozorování
(prumer <- mean(hmotnost))           # výběrový průměr
(se.prumer <- sd(hmotnost)/sqrt(n))  # směrodatná chyba průměru
(odlisnost <- prumer - 8200)         # odchylka půrměru od testované hodnoty
(T <- odlisnost / se.prumer)         # hodnota t-statistiky
```

Ověřte si, že nám vyšla stejná hodnota jako ve výstupu procedury `t.test`.

✧ P-hodnota. Z definice p-hodnoty (text *ÚTH*, str. 4-5) je jasné, že p-hodnota je součtem ploch vymezených vypočtenými hodnotami testové statistiky: Tyto modré plochy odpovídají pravděpodobnostem $P(\text{testová statistika} \leq -|T|)$ a $P(\text{testová statistika} \geq |T|)$. V R tyto pravděpodobnosti spočítáme pomocí distribuční funkce t-rozdělení:

```
(P.vlevo <- pt(-abs(T), df=n-1))       # = P(testová statistika <= -|T|)
(P.vpravo <- 1 - pt(abs(T), df=n-1))   # = P(testová statistika >= |T|)
(P <- P.vlevo + P.vpravo)              # p-hodnota je součtem obou ploch
```

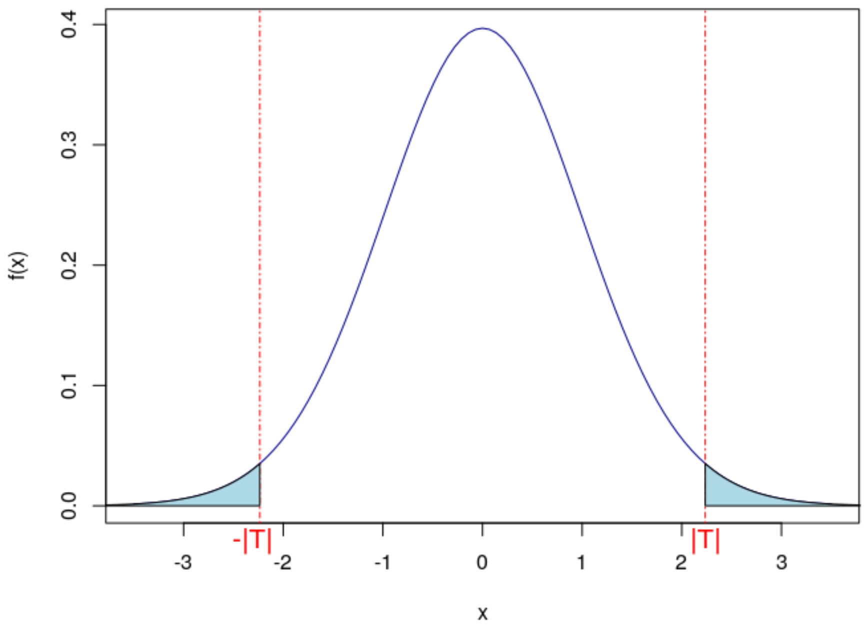
✧ Jak je vidět, obě modré plochy jsou stejné, a tak p-hodnotu lze spočítat najednou:

```
(P <- 2*pt(-abs(T), df=n-1))
```

Ověřte si, že nám vyšla stejná hodnota jako ve výstupu procedury `t.test`.

11) Graficky se lze na P-hodnotu podívat následovně:

```
curve(dt(x,n-1), xlim=c(-3.5, 3.5), xlab="x", ylab="f(x)", col="darkblue")
abline(v=c(-abs(t), abs(t)), col="red", lty=4)
```



4 Konec práce

Než zavřete všechna okna, nezapomeňte si uložit poslední změny ve skriptovém souboru:

File ➡ **Save**

nebo klávesovou skratkou `Ctrl-s`.