Neparametrické jednovýběrové testy a párové testy
10. cvičení

Úvodní nastavení.

- Můžete si stáhnout zdrojový kód k dnešnímu cvičení cviceni10.R.
- Otevřete si program R Studio, změňte si pracovní adresář a vyčistěte pracoviště.
- Načtěte si data Hosi.txt a ujistěte se, že se Vám data dobře načetla.
- Do proměnné alpha is uložte testovací hladinu $0,05$.
- Opět se budeme zabývat porodní hmotností a opět budeme nejprve pracovat pouze s (náhodným) podvýběrem o rozsahu $n = 100$ pozorování.

```
set.seed(2022)
hmot100 = sample(Hosi$por.hmot,100)
n=length(hmot100)
hmot=Hosi$por.hmot
```

Jednovýběrový znaménkový test

1. Minule jsme se zabývali otázkou, zda je střední porodní hmotnost chlapců rovna 3,3 kg. Nyní se budeme snažit zodpovědět tutéž otázku pomocí **znaménkového testu**.
   - Připomeňte si předpoklady tohoto testu (tj. uvažovaný model pro naše data).
   - Formulujte nulovou a alternativní hypotézu. Co je nyní testovaný parametr?
   - Testová statistika má tvar

$$Y_n = \sum_{i=1}^{n} \mathrm{I}[X_i > m_0],$$

   kde m_0 je námi testovaná hodnota 3 300 g. Jaké má přesné a jaké asymptotické rozdělení?
   - Asymptotický test provedeme nejprve manuálně podle návodu z přednášky. Nejprve ale vyřadíme pozorování, která jsou přesně rovna 3 300, a spočítáme si hodnotu Y_n

```
sum(hmot100 == 3300)

(nLess <- sum(hmot100 < 3300))
(nMore <- sum(hmot100 > 3300))

(N <- nLess + nMore)
```

   Naše analýza tedy bude založena na N pozorováních, v proměnné nMore je hodnota testové statistiky Y_n. Pro provedení asymptotického testu si ji příslušně znormujeme. Následně ji porovnáme s kvantilem normálního rozdělení, nebo spočítáme přímo p-hodnotu:

```
(Z = (nMore - N/2) / sqrt(N*1/4))
qnorm(1-alpha/2)

# p-hodnota (dve moznosti, jak spocitat):
2*(1-pnorm(abs(Z)))
2*pnorm(-abs(Z))
```

   K jakému závěru jsme došli?

2. Celý výpočet můžeme provést rychleji pomocí funkce prop.test, která slouží k testování pravděpodobnosti (proporce) v alternativním rozdělení. (S touto funkcí se proto ještě setkáme na jednom z dalších cvičení.)

```
prop.test(x=nMore, n=N, p=1/2, correct=FALSE)
# nebo kdybychom nMore a N nemeli ulozene:
prop.test(x=sum(hmot100>3300),n=sum(hmot100!=3300),p=1/2,correct=F)
```

   - Označíme-li $p_X = \mathsf{P}(X_i > m_0)$, pak výše uvedený test provádí test hypotézy $H_0 : p_X = 1/2$ proti $H_1 : p_X \neq 1/2$.
   - Všimněte si, že jsme dostali opravdu stejnou p-hodnotu. Testová statistika je ale jiná. Najdete souvislost s naším výsledkem uloženým v Z? Jaké asymptotické rozdělení má tedy testová statistika funkce prop.test?
   - Nastavení correct=FALSE zakazuje tzv. korekci pro spojitost, která je jinak defaultně zapnutá. Více v Doplňující informace (i).
   - Znaménkový test lze provést také jako přesný test založený na binomickém rozdělení, viz Doplňující informace (iv) (funkce `binom.test`).

3. Vyzkoušejte, co by se stalo, kdybychom celý výpočet prováděli s nLess namísto nMore.

4. Rozmyslete si jednostrannou variantu testu pro případ, že bychom chtěli testovat, zda je medián m_X porodní hmotnosti menší než 3,3 kg.

   (a) Jak ekvivalentně vyjádříme hypotézu $m_X \geq m_0$ pomocí p_X? Jakou alternativu tedy musíme nastavit do funkce prop.test?

   (b) Je i nyní jedno, jestli zadáme nLess nebo nMore?

5. Jelikož znaménkový test převádí naše naměřená data na binární, podíváme se na to, jak vizualizovat kategoriální veličinu:

```
Y <- hmot100[hmot100 != 3300]; # vybereme hmotnosti ruzne od 3300
Y <- factor(Y > 3300) # udelame z nich 0-1 veliciny a chapeme je jako "faktor"
levels(Y) <- c("< 3300", "> 3300") # pojmenujeme kategorie
summary(Y)   # kontrola s nLess a nMore

plot(Y,col=2:3);
# totez jako:
barplot(table(Y),col=2:3)
# nebo
barplot(c(nLess, nMore), col=2:3,legend=c("< 3300", "> 3300"))

# proporcionalne:
pie(table(Y),col=2:3)
# nebo
barplot(prop.table(as.matrix(c(nLess, nMore))), col=2:3, legend=c("< 3300", "> 3300"))
```

JEDNOVÝBĚROVÝ WILCOXONŮV TEST

6. Stále se budeme zabývat stejnou otázkou jako v bodě 1., ale nyní k testu použijeme **Wilcoxonův test**.

(a) Jaký model předpokládáme pro naše data nyní? Jak budeme teď formulovat testované hypotézy a co je testovaný parametr?
   - Porovnejte předpoklady Wilcoxonova testu s předpoklady přesného t-testu a s předpoklady asymptotického t-testu.
   - Porovnejte, kdy jsou testované hypotézy ekvivalentní a kdy naopak nejsou.
   - Uveďte příklad situace, pro kterou můžeme použít Wilcoxonův test a t-test by nebyl vhodný, a situace opačné.

(b) Připomeňte si testovou statistiku tohoto testu: Je-li $Z_i = X_i - \delta_0$, kde δ_0 je pro nás 3300, pak

$$W_n = \sum_{i \in I} R_i,$$

kde R_i jsou pořadí $|Z_i|$ a I je množina indexů i takových, že $Z_i > 0$. Test můžeme provést jak přesný (pro rozumně malé n; viz Doplňující informace (ii)) nebo asymptotický založený na tom, že statistika

$$\frac{W_n - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{1}$$

má asymptoticky $\mathsf{N}(0,1)$ rozdělení.
   - Pro provedení testu musíme opět nejprve odstranit pozorování rovné přímo δ_0.
   - V případě výskytu shod je potřeba jmenovatel zmenšit o tzv. korekci rozptylu (viz skripta str. 101). V R se toto děje defaultně.
   - Navíc se v R defaultně uvažuje ještě korekce pro spojitost, kdy se čitatel snižuje o 1/2 podobně jako u znaménkového testu. Toto lze vypnout pomocí correct=FALSE.

(c) Provedeme test

```
wilcox.test(hmot100, mu = 3300, correct = FALSE)
```

Jaký je náš závěr?

7. Ověříme, že uvedená testová statistika je skutečně W_n uvedené výše

```
x <- hmot100 - 3300
x <- x[x != 0]     ### nebudeme pouzivat porodni hmotnosti 3300
(R <- rank(abs(x)))
(Splus  <- sum(R[x > 0]))
```

Výpočet p-hodnoty se provede pomocí (1), kde ale navíc musíme v našem případě zohlednit korekci rozptylu. Pro zájemce viz Doplňující informace (iii).

8. Porovnáme sílu t-testu, znaménkového testu a Wilcoxonova testu pro normální rozdělení pomocí simulací. Budeme tedy generovat náhodný výběr o rozsahu n=100 z $\mathsf{N}(0.2, 1)$ a budeme testovat $H_0 : \mu = 0$ proti $H_1 : \mu \neq 0$. Test provedeme pomocí všech tří testů a uložíme si příslušné p-hodnoty. Celý postup zopakujeme nopak=1000 krát a odhad síly spočteme jako relativní četnost případů, kdy se nám podařilo H_0 zamítnout.

```
opak=1000
n=100
```

```
p.t=numeric(opak)
p.z=numeric(opak)
p.w=numeric(opak)
for(i in 1:opak){
 x=rnorm(n,mean=0.2,sd=1)
 p.t[i]=t.test(x,mu=0)$p.val
 p.w[i]=wilcox.test(x,mu=0)$p.val
 p.z[i]=prop.test(sum(x>0),n,p=1/2)$p.val
}

mean(p.t<=0.05)
mean(p.w<=0.05)
mean(p.z<=0.05)
```

Zkuste změnit velikost rozsahu `n` nebo alternativu, za které simulujeme, a sledujte, jak se mění síla. Lze udělat nějaký obecný závěr o síle zkoumaných tří testů?

9. Společně se podíváme na porovnání síly uvedených tří testů v závislosti na velikosti parametru polohy pro různá rozdělení, viz zdrojový kód `cviceni10_obrazky.R`. Budeme testovat nulovost parametru polohy a sledovat sílu jednotlivých testů pro $n = 100$ a náhodný výběr $X_1, \ldots, X_n$ z rozdělení:
   (a) $\mathsf{N}(m_j, 1)$,
   (b) Cauchyho rozdělení s hustotou
   $$f(x) = \frac{1}{\pi\left(1 + (x - m_j)^2\right)}, \quad x \in \mathbb{R},$$
   (c) $X_i = Y_i - 1$, kde $Y_1, \ldots, Y_n$ je náhodný výběr z $\mathsf{Exp}\left(\frac{1}{1+m_j}\right)$,

   kde $0 \leq m_j \leq 0.5$ pro $j = 1, \ldots, 20$. Rozmyslete si, jak pro daná rozdělení vypadají $\mathsf{E}X$, m_X a střed symetrie δ_X a zda jsou splněny předpoklady všech použitých testů.

Párový problém

10. Některé zdroje uvádějí, že chlapci v průběhu prvního roku života přiberou v průměru více než 6,5 kg. Máme za úkol ověřit, zda jsou naše data v souladu s tímto tvrzením. Použijeme k tomu data z celého našeho datového souboru.
    (a) Jsou porodní hmotnost a hmotnost v jednom roce nezávislé veličiny? Uveďte další možné příklady párového problému.
    (b) Spočítáme si váhový přírůstek během prvního roku:
    ```
    hmot <- Hosi$por.hmot
    hmot1 <- Hosi$hmotnost
    rozdil = hmot1-hmot
    ```
    (c) Nejprve použijeme t-test.
    - Jaký model předpokládáme (uvažte pomocí vhodných grafů) a jak zní nulová a alternativní hypotéza?
    - Proveďte test a interpretujte výsledek.

```
# postup, ktery jiz zname:
t.test(rozdil, alternative="greater",mu=6500)
# nebo pomoci volby paired:
t.test(hmot1, hmot,  alternative="greater", mu=6500, paired=T)
# pozor na poradi!
```

(d) Nyní na stejný problém použijeme znaménkový test.
   - Co nyní předpokládáme? A co testujeme?
   - Provedeme test

```
n.more=sum(rozdil>6500)
n=sum(rozdil!=6500)
prop.test(n.more, n, alternative = "greater")
```

   Jaký je náš závěr?

(e) Nakonec si na tomto problému vyzkoušíme ještě i Wilcoxonův test.
   - Jaký je zde předpokládaný model a příslušné hypotézy?
   - Proveďte test.

Samostatná práce

11. Zajímá nás věkový rozdíl mezi rodiči chlapců. Konkrétně bychom rádi věděli, zda jsou otcové v průměru o více než 2 roky starší než matky.
   (a) Nejprve si prohlédněte vhodné popisné statistiky a grafy, které ilustrují rozdělení věkového rozdílu.
   (b) Zvolte test, který Vám příjde pro tato data nejvhodnější a proveďte jej. Uvědomte si, jaký uvažujete model, jaké testujete hypotézy a jak budeme interpretovat závěr.

12. V rámci tohoto cvičení jsme pro test hypotézy, zda je průměrná porodní hmotnost chlapců rovna 3,3 kg, použili několik rozličných testů, které v praxi mohou vést k různým p-hodnotám a různým závěrům. Kterému výsledku tedy máme „věřit“?

13. Pomocí znaménkového testu proveďte test hypotézy, zda je průměrná porodní hmotnost nižší než 3,5 kg. Zformulujte řádně nulovou a alternativní hypotézu a proveďte test.
   Totéž proveďte i pro jednostranný Wilcoxonův test.

Doplňující informace

(i) Vyzkoušejte si znaménkový test se zapnutou korekcí pro spojitost. Ověřte, že v tomto případě se počítá testová statistika

$$Z^C = \frac{Y_n - \frac{n}{2} - \frac{1}{2}\operatorname{sgn}(Y_n - \frac{n}{2})}{\sqrt{\frac{n}{4}}}$$

(resp. její druhá mocnina). Použijte postup analogický tomu v bodě 2.
Korekce pro spojitost zde vychází z toho, že Y_n má diskrétní rozdělení a $\mathsf{P}(Y_n \le y) = \mathsf{P}(Y_n < y+1)$. Tudíž při přechodu k asymptotické aproximaci spojitým normálním rozdělením s distribuční funkcí F lze $\mathsf{P}(Y_n \le y)$ aproximovat jak $F(y+\varepsilon)$ pro všechna $\varepsilon \in [0,1]$. Jako určitý kompromis tedy vezmeme $F(y + 1/2)$.

(ii) Pro Wilcoxonův test můžeme použít i přesný test, založený na přesném rozdělení W. Je však nutné, aby v datech nebyly shody a žádné pozorování nebylo rovno přesně testované hodnotě. Pro vyšší n může být ale výpočet časově náročný. Vyzkoušejte si to na simulovaných datech z rovnoměrného rozdělení:

```
x=runif(100,0,1)
wilcox.test(x,mu=1/2,exact=TRUE)
```

Postupně zvyšujte počet použitých dat ze 100 na 200, 500, 1000, 2000. Pro jaký rozsah dat je ještě možné použít přesný test?

(iii) Manuální výpočet p-hodnoty Wilcoxonova testu: Vzorec pro korekci rozptylu z důvodu výskytu shod viz skripta str. 101 (pasáž před začátkem části 5.5).

```
ESplus <- N*(N+1)/4
# totez jako 1/2*sum(1:N)
varSplus_NoTies <- N * (N + 1) * (2 * N + 1) / 24

# vypocet korekce:
(tx <- table(R)) # Pocet jednotlivych hodnot
varCorrectWithTies <- sum(tx^3 - tx) / 48   # Korekce v pripade shod
varSplus <- varSplus_NoTies - varCorrectWithTies

(U <- (Splus - ESplus) / sqrt(varSplus))
(U.corr <- (Splus - ESplus - 0.5) / sqrt(varSplus))


### P-hodnoty
(P           <- 2*pnorm(-abs(U)))
(P.corr <- 2*pnorm(-abs(U.corr)))
```

První p-hodnota odpovídá testu bez korekce pro spojitost, druhá odpovídá korekci pro spojitost. Můžeme je srovnat s p-hodnotami:

```
wilcox.test(hmot100, mu = 3300, correct = FALSE)$p.val
wilcox.test(hmot100, mu = 3300)$p.val
```

Výsledné p-hodnoty zkuste sami porovnat s p-hodnotou, kterou bychom dostali bez korekce jmenovatele.

(iv) Znaménkový test má kromě asymptotické verze (kterou jsme dělali výše) i přesnou variantu založenou na binomickém rozdělení. V R můžeme použít funkce `binom.test`, parametry této funkce jsou pak stejné jako u `prop.test`. Proveďte pro porovnání test hypotézy, zda je průměrná porodní hmotnost nižší než 3,5 kg, i pomocí přesného testu. Vyzkoušejte použití našich 100 vybraných hodnot i všech dostupných dat.