Jednovýběrové testy: t-test, Kolmogorovův-Smirnovův test

9. cvičení

Úvodní nastavení.

- Stáhněte si data Iq.txt (a případně i Hosi.txt, pokud je již nemáte stažena z minulého cvičení).
- Otevřete si program R Studio.
- Změňte si pracovní adresář pomocí Session →Set working directory →Choose directory nebo napište přímo použijte příkaz setwd, do kterého zadáte cesto do tohoto adresáře.
- Vyčistěte si pracoviště od starých objektů, které zůstaly uloženy:
  `rm(list=ls())`
- Načtěte si data Hosi.txt a Iq.txt.

  ```
  Hosi=read.table("Hosi.txt",header=TRUE);
  Iq=read.table("Iq.txt",header=T, stringsAsFactors=T)
  ```

  Pro jistotu se podívejte na prvních několik řádků (příkaz head) a ujistěte se, že se Vám data dobře načetla. Můžete zavolat i summary.
- Do proměnné alpha is uložte testovací hladinu 0.05, na které budeme provádět většinu dnešních testů.
  `alpha <- 0.05;`

I. Jednovýběrový t-test

1. Opět se budeme zabývat porodní hmotností, ale tentokrát budeme pracovat pouze s (náhodným) podvýběrem o rozsahu $n = 100$ pozorování.

   ```
   set.seed(2022);
   n <- 100;
   hmot100 <- sample(Hosi$por.hmot, n)
   ```

   Abychom měli všichni stejný podvýběr, zvolili jsme si pevné nastavení generátoru pseudonáhodných čísel pomocí set.seed. Uložte si do n rozsah výběru hmot100.

2. Na internetové stránce ČSÚ se v zásadě uvádí, že je průměrná porodní hmotnost novorozených chlapců rovna 3,349 kg. Ověřte, zda jsou naše data v souladu s tímto tvrzením.
   (a) Zformulujte vhodný pravděpodobnostní model a pokuste se graficky posoudit, zda je vhodný pro naše data.
   (b) Zformulujte nulovou a alternativní hypotézu.
   (c) Proveďte test pomocí funkce t.test.
   `t.test(hmot100,mu=3349)`
   Společně si řekneme, co jednotlivé části výstupu znamenají.
   (d) Rozhodněte o zamítnutí/nezamítnutí nulové hypotézy. Zformulujte závěr.

3. Nyní si jednotlivé části z výstupu funkce t.test spočítáme „ručně“. Jak víme, testová statistika má tvar
$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \mu_0)}{S_n}$$
a má za nulové hypotézy t_{n-1} rozdělení. Spočteme ji tedy následovně:

```
(tstat=sqrt(n)*(mean(hmot100)-3349)/sd(hmot100))
```

a klasickým způsobem testování bychom její absolutní hodnotu porovnali s kritickou hodnotou

```
qt(1 - alpha/2, df=n-1)
```

Jaký závěr dostáváme z tohoto porovnání?
My ale máme ve výstupu p-hodnotu. Ta udává nejmenší hladinu testu, na které bychom nulovou hypotézu ještě zamítli. Zároveň vyjadřuje pravděpodobnost, s jakou bychom za nulové hypotézy dostali ještě „méně příznivý“ výsledek než je hodnota naší testové statistiky. Pokud jsme napozorovali hodnotu t, spočítáme ji tedy jako $p = \mathsf{P}(|W| > |t|) = 2\,(1 - F_W(|t|))$, kde W je náhodná veličina s t_{n-1} rozdělením a F_W je její distribuční funkce. V R :

```
2*(1-pt(abs(tstat), df=n-1))
```

Konečně, intervalový odhad parametru μ_X spočítáme

```
mean(hmot100)-qt(1 - alpha/2, df=n-1)/sqrt(n)*sd(hmot100)
mean(hmot100)+qt(1 - alpha/2, df=n-1)/sqrt(n)*sd(hmot100)
```

Vzpomeňte si na dualitu mezi intervalovým odhadem a testem hypotézy (viz přednáška).

4. Někdy se nám může hodit umět přistupovat k jednotlivým položkám výsledku funkce t.test.

```
tt=t.test(hmot100,mu=3349)
names(tt)
tt$stat
tt$p.val
```

Zkuste si takto nechat vypsat interval spolehlivosti pro μ a podívejte se, jakou test uvažuje alternativu.

5. Nechejte si vypsat interval spolehlivosti s pravděpodobností pokrytí $0,99$. Ve funkce t.test nastavte conf.level $= 0.99$. Pouze na základě tohoto intervalu rozhodněte o zamítnutí/nezamítnutí nulové hypotézy na testovací hladině $0,01$.

6. Podobně jako v bodě 2. otestujte, zda je pravdivé tvrzení, že je střední porodní hmotnost chlapců nižší než 3,5 kg.
   - Zformulujte nulovou a alternativní hypotézu a řádně interpretujte výsledek.
   - Všimněte si, jaký intervalový odhad nám nyní R nabízí.

7. Jak bychom ručně spočítali p-hodnotu z 6?

8. Ve 3. jsme konstruovali interval spolehlivosti na základě t-rozdělení uvedené statistiky T_n. Jaké je její limitní rozdělení pro $n \to \infty$? Spočtěte asymptotický intervalový odhad μ zkonstruovaný na základě tohoto limitního rozdělení. Porovnejte oba intervaly. Který z nich je širší?

9. Spočítejte také p-hodnotu příslušného asymptotického testu. Jaký model zde stačí předpokládat? A na základě kterého rozdělení se doporučuje počítat p-hodnotu (resp. kritickou hodnotu)?

10. Doposud jsme k testování používali jenom část dat, a to proto, abychom mohli provést následující porovnání: Proveďte ještě jednou testy z 2. a 6. pro celá data a porovnejte je s výsledkem pro náš podvýběr. Co pozorujeme? Co z toho vyplývá pro praxi?

II. Kolmogorovův-Smirnovův test

11. Načtěte si data `Iq.txt`, která se týkají hodnot IQ a známek na ZŠ náhodně vybraných žáků. Prohlédněte si data a proměnné, které máme k dispozici.

12. Bude nás zajímat IQ žáků, proto si hodnoty uložte do proměnné `IQ` a do `n` si uložte rozsah výběru.

```
IQ = Iq$iq;
n = length(IQ)
```

Jakými popisnými statistikami byste popsali data? Jaké obrázky ilustrují rozdělení dat?

13. Na wikipedii se uvádí, že IQ má v populaci normální rozdělení se střední hodnotou 100 a směrodatnou odchylkou 15. Zajímá nás, zda naše data podporují nebo vyvracejí toto tvrzení.
    (a) Jaký model předpokládáme pro naše data? Formulujte nulovou a alternativní hypotézu, kterou budeme testovat.
    (b) Provedeme Kolmogorovův-Smirnovův test
    `ks.test(IQ, y="pnorm", mean=100, sd=15)`
    Poznámka: R nás upozorňuje, že v datech se objevují shodná pozorování. Pomocí `table(IQ)` si můžeme nechat vypsat tabulku četností.
    (c) Jaký je náš závěr?

14. Z přednášky víme, že testová statistika má tvar $K_n = \sup_x |\widehat{F}_n(x) - F_0(x)|$, ale pro výpočet se používá

$$K_n^+ = \max_{1\le i\le n}\left(\frac{i}{n} - F_0(X_{(i)})\right), \quad K_n^- = \max_{1\le i\le n}\left(F_0(X_{(i)}) - \frac{i-1}{n}\right), \quad K_n = \max(K_n^+, K_n^-).$$

Ověříme tedy, že R počítá opravdu tuto testovou statistiku:

```
F0 <- function(x) pnorm(x, mean=100, sd=15)
IQ.sorted=sort(IQ)
Knplus <- max((1:n)/n - F0(IQ.sorted))
Knminus <- max(F0(IQ.sorted) - (0:(n-1))/n)
(Kn <- max(Knplus, Knminus))
```

Vše vychází, ale úplně v pořádku výše uvedené není. Proč?

Za nulové hypotézy má $\sqrt{n}K_n$ asymptoticky rozdělení s distribuční funkcí

$$G(y) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2y^2}.$$

Pomocí konečné aproximace této distribuční funkce můžeme p-hodnotu testu spočíst

```
G <- function(y){
  k <- 1:10000; # aproximace nekonecne sumy
  1 - 2*sum((-1)^(k+1)*exp(-2*k^2*y^2))
}
1 - G(sqrt(n)*Kn)
```

Kdybychom chtěli znát kritickou hodnotu, tak vlastně hledáme c takové, že $\mathsf{P}(\sqrt{n}K_n \geq c) = \alpha$ za platnosti H_0, tj. $G(c) = 1 - \alpha$. To musíme vyřešit numericky:

```
fce <- function(x) G(x) - (1 - alpha)
(krit <- uniroot(fce, c(0.1,10))$root)
sqrt(n)*Kn
```

15. Celou situaci si graficky znázorníme. Vzpomeňte si na odvození tzv. pásu spolehlivosti pro distribuční funkci, které jste měli na přednášce.

```
empir.df <- ecdf(IQ);
plot(empir.df, ylab="", main="Porovnani distribucnich funkci",do.points=F)

# pas spolehlivosti
xgrid <- seq(min(IQ)-10, max(IQ)+10, length=10000);
dolni.pas <- pmax(empir.df(xgrid)-krit/sqrt(n),0)
horni.pas <- pmin(empir.df(xgrid)+krit/sqrt(n),1)
polygon(c(xgrid, rev(xgrid)), c(horni.pas, rev(dolni.pas)), col = "gray90", border="gray70",lty=2)

plot(empir.df,  verticals=FALSE,  add=T, lwd=2,do.points=FALSE)
 lines(xgrid, F0(xgrid), col="blue", lwd=2); # graf F_0

# Bod nejvetsiho rozdilu
k0 <- which.max(F0(IQ.sorted) - (0:(n-1))/n);
abline(v=IQ.sorted[k0], col="red");

legend("bottomright", col=c("black", "grey70", "blue"),  lty=c(1,2,1),
   legend=c("empiricka d.f.", "pas spolehlivosti", "d.f. za nulove hypotezy"))
```

16. Připomeňte si, že Kolmogorovův-Smirnovův test předpokládá, že je F_0 specifikovaná úplně, včetně hodnot všech parametrů. V případě, že tomu tak není (pokud bychom parametry odhadovali z dat), tak nám K-S test nedává správnou p-hodnotu (rozdělení testové statistiky za nulové hypotézy je jiné). Ukážeme si to na malé simulaci, kdy budeme generovat data o stejném rozsahu, tedy $n = 111$, z $\mathsf{N}(100, 15^2)$ a budeme provádět K-S test jednak se zadanými parametry a jednak s parametry odhadnutými z dat.

```
nopak <- 10000
pval.zname <- numeric(nopak)
pval.nezname <- numeric(nopak)
n<-111

for(i in 1:nopak){
  Y <- rnorm(n, mean=100, sd=15);
  pval.zname[i] <- ks.test(Y, y="pnorm", mean=100, sd=15)$p.value;
  pval.nezname[i] <- ks.test(Y, y="pnorm", mean=mean(Y), sd=sd(Y))$p.value;
}

mean(pval.zname <= 0.05)
mean(pval.nezname <= 0.05)
```

Poslední dvě hodnoty nám odhadují skutečnou hladinu testu (měla by být 0,05). Vidíme, že dostáváme dvě velmi rozličné hodnoty. Co lze tedy říci o K-S testu v případě, že bychom ho špatně používali s parametry odhadnutými z dat?

Můžeme si porovnat i histogramy p-hodnot pro obě situace. Podle tvrzení 4.1. by za platnosti H_0 měla mít $p(\mathbf{X})$ rovnoměrné rozdělení na $[0, 1]$.

```
hist(pval.zname,prob=TRUE,main="p-hodnota pro zname parametry")
hist(pval.nezname,prob=TRUE,main="p-hodnota pro nezname parametry")
```

17. Dalším důležitým předpokladem K-S testu je spojité rozdělení dat. Pomocí následující simulace se podíváme na skutečnou hladinu testu, použijeme-li nesprávně K-S test (s distribuční funkcí G) pro test shody s binomickým rozdělením $\mathsf{Bi}(N, p)$ na základě dat o rozsahu n.

```
nopak=10000
pval=numeric(nopak)
n=100

N=10
p=1/3
F0=function(x) pbinom(x, size=N, prob=p)

set.seed(123)
for(i in 1:nopak){
  x = rbinom(n,size=N,prob=p)
  xi=unique(c(x,0:N))
  Kn=max(abs(ecdf(x)(xi)-F0(xi)))
  pval[i]=1-G(sqrt(n)*Kn)
}
mean(pval<=0.05)

hist(pval)
```

Co lze na základě předchozí simulace říci o K-S testu, použijeme-li jej pro F_0 s diskrétním rozdělením?

Zkuste v předchozí simulaci zvýšit N na 2 000. Co sledujeme?

III. Simulace

18. Provedeme simulace z rovnoměrného a exponenciálního rozdělení a budeme zkoumat skutečnou hladinu testu při použití asymptotického testu (tj. při použití t-testu na nenormální data).

```
n=100
opak=1000
p.rovn=rep(NA, opak)
p.exp=rep(NA,opak)

for(i in 1:opak){
  x1=runif(n)
  p.rovn[i]=t.test(x1,mu=1/2)$p.val
#
  x2=rexp(n,rate=1)
  p.exp[i]=t.test(x2,mu=1)$p.val
}

mean(p.rovn<=0.05)
mean(p.exp<=0.05)


hist(p.rovn,prob=TRUE)
hist(p.exp,prob=TRUE)
```

Zkuste změnit počet pozorování n (zmenšit a zvětšit). Pro jaké n je rozumné použít asymptotický test?

19. Podobné simulace provedeme za alternativy a budeme sledovat sílu testu.

```
n=100
opak=1000
p.rovn=rep(NA, opak)
p.exp=rep(NA,opak)

for(i in 1:opak){
  x1=runif(n,0,1.2);
  p.rovn[i]=t.test(x1,mu=1/2)$p.val;
#
  x2=rexp(n,rate=1.2);
  p.exp[i]=t.test(x2,mu=1)$p.val
}
```

```
mean(p.rovn<=0.05);
mean(p.exp<=0.05)

hist(p.rovn,prob=TRUE)
hist(p.exp,prob=TRUE)
```

Vyzkoušejte měnit parametry rozdělení a sledovat, jak se mění síla testu, pokud se od nulové hypotézy vzdalujeme. Podobně, nechejte parametry rozdělení fixní a sledujte, jak se mění síla, pokud zvyšujeme počet pozorování. Učiňte nějaký obecný závěr z tohoto zkoumání.