Cvičení z NSTP097

Jednovýběrové a párové testy

XI.

Úloha 1: Jednovýběrové testy — hladina, síla, p-hodnota

Uvažujme náhodný výběr $X_1, \ldots, X_n$ z rozdělení $\mathrm{N}(0.5, 2)$ o rozsahu $n = 60$. Vygenerujte jeden takový výběr příkazy

```
nobs = 60
x = rnorm(nobs,mean=0.5,sd=sqrt(2))
```

Nyní provedeme jednovýběrový Kolmogorovovův-Smirnovův test hypotézy $H_0 : X_i \sim \mathrm{N}(0.5, 2)$ proti alternativě, že X_i mají libovolné jiné rozdělení. V R se takový KS test na výběru `x` provede příkazem

```
ks.test(x,y="pnorm",mean=0.5,sd=sqrt(2))
```

(`pnorm` specifikuje distribuční funkci normálního rozdělení). Prozkoumejte výstup z této funkce: kde je testová statistika, kde je p-hodnota? Rozhodněte, zdali došlo k zamítnutí nulové hypotézy.

Nakreslíme si obrázek empirické distribuční funkce spolu s distribuční funkcí za hypotézy:

```
od = min(x)*0.9
do = max(x)*1.1
plot(ecdf(x),xlim=c(od,do))
body = seq(od,do,length=500)
lines(body,pnorm(body,mean=0.5,sd=sqrt(2)),col="blue")
```

Dá se z obrázku okem odhadnout hodnota testové statistiky KS testu?

Nyní na ten samý výběr `x` proveďte postupně test hypotézy $H_0 : X_i \sim \mathrm{N}(\mu, 2)$ pro $\mu = 0.6, 0.65, 0.7, 0.75, \ldots$. Jak se mění výsledek testu? Pokračujte v oddalování střední hodnoty hypotetického rozdělení od střední hodnoty skutečného rozdělení dat po stejných krůčcích, dokud nedojde k zamítnutí hypotézy. Pak si nakreslete obrázek empirické distribuční funkce dat a distribuční funkce za platnosti hypotézy.

Zopakujte to samé zadání s t-testem: nejdříve proveďte t-test hypotézy $H_0 : \mathsf{E}\, X_i = 0.5$ proti alternativě $H_1 : \mathsf{E}\, X_i \neq 0.5$ na původním datovém souboru `x` příkazem `t.test(x,mu=0.5)` a prozkoumejte výstup z funkce `t.test`. Pak zkoumejte výsledky t-testu hypotéz $H_0 : \mathsf{E}\, X_i = \mu$ pro $\mu = 0.6, 0.65, 0.7, 0.75, \ldots$, dokud nedojde k zamítnutí H_0. Došlo k němu dříve nebo později než u KS testu?

Úloha 2: Jednovýběrové testy — simulace

Nyní budeme simulovat hladinu a sílu jednovýběrových testů. Vyrobíme si jednoduchou funkci, která provede test na data `x` a vrátí pouze p-hodnotu.

```
vem.ph = function(x,test,...)
  {
    test(x,...)$p.value
  }
```

Vyzkoušejte si ji na původní data s KS testem a t-testem:

```
vem.ph(x,ks.test,y="pnorm",mean=0.5,sd=sqrt(2))
vem.ph(x,t.test,mu=0.5)
```

Teď vygenerujeme 1000 výběrů o rozsahu 60 z rozdělení $\mathrm{N}(0.5, 2)$:

```
nvyb = 1000
x.vyb = matrix(rnorm(nobs*nvyb,0.5,sqrt(2)),nrow=nobs,ncol=nvyb)
```

Nevypisujte si je ani omylem! Na každý výběr provedeme KS test hypotézy $H_0 : X_i \sim \mathrm{N}(0.5, 2)$ a získáme jeho p-hodnotu:

```
ks.ph = apply(x.vyb,2,vem.ph,ks.test,y="pnorm",mean=0.50,sd=sqrt(2))
```

Aplikovali jsme funkci `vem.ph` s testem `ks.test` na sloupce matice `x.vyb` a získali vektor `ks.ph` p-hodnot pro 1000 výběrů z rozdělení za platnosti hypotézy. Nakreslete si jejich histogram `hist(ks.ph)`. Výsledné p-hodnoty jsou náhodné veličiny: jaké je v tomto případě jejich rozdělení? Váš úsudek si teoreticky zdůvodněte a ověřte provedením KS testu na výběr 1000 p-hodnot. Spočtěte, jaký podíl p-hodnot je menších než 0.05, tj. `mean(ks.ph<0.05)`. Co odhaduje toto číslo?

Zopakujte tuto úlohu za následujících podmínek:

- Generujte výběry z $\mathrm{N}(\mu, 2)$ pro $\mu = 0.7$ a $\mu = 0.9$, testujte stále hypotézu $H_0 : X_i \sim \mathrm{N}(0.5, 2)$ KS testem. Jak se mění rozdělení p-hodnot? Jak se mění počet p-hodnot menších než 0.05 a co to znamená?
- Generujte výběry z $\mathrm{N}(0.5, 2)$ a provádějte t-test hypotézy $H_0 : \mathsf{E}\, X_i = 0.5$. Interpretujte výsledky.
- Generujte výběry z $\mathrm{N}(\mu, 2)$ pro $\mu = 0.7$ a $\mu = 0.9$, testujte hypotézu $H_0 : \mathsf{E}\, X_i = 0.5$ t-testem. Interpretujte výsledky.

Úloha 3: Párové testy, simulace

Nyní uvažujme náhodný vektor $\binom{X}{Y}$, kde $X \sim \mathrm{N}(0.5, 2)$ a $Y = X + \varepsilon$, kde $\varepsilon \sim \mathrm{N}(0, 1)$ je nezávislé na X. Jaké je sdružené rozdělení $\binom{X}{Y}$? Spočítejte $\mathsf{cov}\,(X, Y)$ a $\mathsf{cor}\,(X, Y)$.

Vygenerujme jeden výběr dvojic $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ pro $n = 60$:

```
nobs = 60
x = rnorm(nobs,mean=0.5,sd=sqrt(2))
y = x+rnorm(nobs,mean=0,sd=1)
```

Spočítejte výběrový korelační koeficient mezi X_i a Y_i (`cor(x,y)`) a porovnejte jej se skutečnou korelací $\mathsf{cor}\,(X, Y)$. Spočítejte průměry X_i a Y_i. Proveďte párový t-test hypotézy $H_0 : \mathsf{E}\, X_i = \mathsf{E}\, Y_i$ proti alternativě $H_1 : \mathsf{E}\, X_i \neq \mathsf{E}\, Y_i$:

```
t.test(x,y,paired=T)
```

a prozkoumejte jeho výstup (identifikujte všechny komponenty výstupu, zejména testovou statistiku a p-hodnotu). Nyní generujte znovu `y` při pevném `x`, přičemž po krocích velkých 0.05 postupně zvětšujte střední hodnotu $\mathsf{E}\, Y_i$, dokud nedojde k zamítnutí H_0.

Nyní znovu provedeme simulace, ale musíme upravit funkci pro výpočet p-hodnot na párový test. Data budeme ukládat do matice o `nvyb=1000` sloupcích tak, aby prvních 60 prvků každého sloupce obsahovalo X_i a druhých 60 prvků odpovídající Y_i. Funkce pro výpočet p-hodnot bude

```
vem.ph.2 = function(data,test,...)
  {
    test(x=data[1:nobs],y=data[-c(1:nobs)],...)$p.value
  }
```

Vyzkoušejte `vem.ph.2(c(x,y),t.test,paired=T)`, musíte dostat stejnou p-hodnotu jako z `t.test(x,y,paired=T)`.

Nyní dáme dohromady výběry pro $X \sim \mathrm{N}(0.5, 2)$ a $Y = X + \varepsilon$, kde $\varepsilon \sim \mathrm{N}(0, 1)$,

```
nvyb = 1000
x.vyb = matrix(rnorm(nobs*nvyb,0.5,sqrt(2)),nrow=nobs,ncol=nvyb)
y.vyb = x.vyb+matrix(rnorm(nobs*nvyb,mean=0,sd=1),nrow=nobs,ncol=nvyb)
oba.vyb = rbind(x.vyb,y.vyb)
```

vypočteme 1000 p-hodnot

```
t.ph = apply(oba.vyb,2,vem.ph.2,t.test,paired=T)
```

nakreslíme si jejich histogram, otestujeme jejich rozdělení, a spočítáme, kolik jich bylo pod 0.05.

V dalším kroku opakujeme simulace se změnou střední hodnoty $\mathsf{E}\,\varepsilon$: místo 0 dáme 0.1, 0.2, 0.3, a 0.4. Jak se mění chování testu hypotézy $H_0 : \mathsf{E}\,X_i = \mathsf{E}\,Y_i$ proti alternativě $H_1 : \mathsf{E}\,X_i \neq \mathsf{E}\,Y_i$?

Úloha 4: Párové testy, data „Pneu“

Připomeňme, že datová tabulka `pneu` obsahuje měření životnosti pneumatik (v tisících km jízdy) dvěma metodami: měření pomocí úbytku hmotnosti pneumatiky (veličina `met.v`) a měření pomocí úbytku hloubky dezénu (veličina `met.d`) – viz minulé cvičení. Natáhněte si znovu data (a jednu drobnou funkci) příkazem

```
siteaddr = "http://www.karlin.mff.cuni.cz/~pesta/NSTP097"
datafile = paste(siteaddr,"cvic_k10_3.RData",sep="/")
load(url(datafile))
```

Spočítejte párový t-test srovnávající obě metody měření.

```
t.test(met.v,met.d,paired=T)
```

Najděte ve výstupu testovou statistiku a p-hodnotu. Co je nulová hypotéza? Zamítáme ji? Spočítejte kritickou hodnotu pro testovou statistiku. Proveďte vlastní výpočet p-hodnoty z testové statistiky. Návod: p-kvantil t-rozdělení s k stupni volnosti dostaneme jako `qt(p,df=k)`, distribuční funkci v bodě x jako `pt(x,df=k)`.

Proveďte znaménkový test. Máte k disposici funkci `sign.test`, kterou jste si natáhli s daty (původně pochází z knihovny BSDA). Zavolejte `sign.test(met.v,met.d)`. Co testuje tento test? Zamítá nulovou hypotézu? Spočítejte sami v R jeho testovou statistiku.