Cvičení z NSTP097

Dvouvýběrové testy, analýza rozptylu

XII.

Vše potřebné k dnešnímu cvičení natáhnete z webu do R příkazy

```
siteaddr <- "http://www.karlin.mff.cuni.cz/~pesta/NSTP097"
datafile <- paste(siteaddr,"cvic_k10_4.RData",sep="/")
load(url(datafile))
```

Vyzkoušejte, že se vše povedlo: příkaz `ls()` musí mezi vypsanými objekty ukázat `skewness` (předdefinovaná funkce) a `plat` (datový soubor).

Datový soubor

Datový soubor `plat` obsahuje údaje z výběrového šetření 534 zaměstnaných osob žijících na různých místech v USA v roce 1985 (Berndt, ER. *The Practice of Econometrics.* 1991. NY: Addison-Wesley). Každý řádek obsahuje údaje pro jednoho respondenta šetření. Budeme vyšetřovat závislost výše mzdy na některých vybraných faktorech.

Veličiny a jejich kódování

`vzdel`	Vzdělání (počet let ve škole)
`jih`	Geografická poloha (Jih/jinde)
`pohl`	Pohlaví (Žena/Muž)
`leta`	Délka praxe (roky)
`odbory`	Členství v odborech (Člen/Nečlen)
`mzda`	Hodinová mzda v US $
`vek`	Věk (roky)
`rasa`	Rasa (Běloch/Hispanik/Jiná)
`zaraz`	Pracovní zařazení (Vedoucí prac./Obchod/Administrativa/Služby/Odborný prac./Ostatní)
`sektor`	Sektor (Stavebnictví/Průmysl/Ostatní)
`stav`	Rodinný stav (Ženatý–vdaná/Ostatní)

Zadání

(i) Vypište jména veličin (`names(plat)`). Prohlédněte si prvních dvacet pozorování (`plat[1:20,]`). Vypište si základní charakteristiky jednotlivých veličin souboru – `summary(plat)`. Zajistěte si k nim přímý přístup příkazem `attach(plat)`. Zobrazte histogram mezd (`hist(mzda)`).

(ii) Nejprve se budeme zabývat otázkou, zdali jsou na Jihu v průměru stejné platy, jako jinde ve Spojených státech. Ve značení používaném na přednášce máme dva nezávislé výběry $X_1, \ldots, X_n$ z nějaké distribuční funkce F_X (platy na Jihu) a $Y_1, \ldots, Y_m$ z obecně různé distribuční funkce F_Y (platy jinde). Zajímá nás rozdíl mezi distribučními funkcemi F_X a F_Y, zejména pak rozdíl jejich středních hodnot $\mathsf{E}\, X_i = \int_{-\infty}^{\infty} x\, dF_X(x)$ a $\mathsf{E}\, Y_j = \int_{-\infty}^{\infty} y\, dF_Y(y)$. Jelikož hodinové mzdy nabývají mnoha kladných hodnot, můžeme rozdělení F_X a F_Y považovat za spojitá, s nosičem $(0, \infty)$.

(iii) `table(jih)` prozradí, že $n = 156$ a $m = 378$. Zjistěme výběrové momenty rozdělení F_X a F_Y: odhady středních hodnot pomocí `tapply(mzda,jih,mean)`, odhady rozptylů pomocí `tapply(mzda,jih,var)`, odhady šikmosti pomocí `tapply(mzda,jih,skewness)`. Získané odhady nám dávají hrubou představu o tom, zdali se F_X a F_Y liší střední hodnotou či rozptylem a zdali jsou obě rozdělení přibližně symetrická.

(iv) Porovnejte rozdělení obou výběrů pomocí krabicových diagramů (*box plot*)[1]
`boxplot(split(mzda,jih))`.

(v) Porovnejte histogramy obou výběrů[2]:

```
par(mfrow=c(2,1))
hist(mzda[jih=="Jih"],breaks=5*(0:10),xlim=range(mzda))
hist(mzda[jih=="jinde"],breaks=5*(0:10),xlim=range(mzda))
par(mfrow=c(1,1))
```

(vi) Nyní spočtěme a nakresleme empirické distribuční funkce $\hat{F}_{X,n}$ a $\hat{F}_{Y,m}$[3]:

```
a <- ecdf(mzda[jih=="Jih"])
b <- ecdf(mzda[jih=="jinde"])
oddo <- range(mzda)*c(0.9,1.1)
par(col="blue")
plot(a,xlim=oddo,main="Empiricka distribucni funkce mzdy",cex=0.3)
par(col="black")
lines(b,cex=0.3,lty=2)
legend(30,0.4,lty=c(1,1),col=c("blue","black"),legend=c("Jih","Jinde"))
```

Co prozrazují empirické distribuční funkce o platech?

(vii) Proveďme formální test hypotézy $H_0 : F_X = F_Y$ proti $H_1 : \exists t : F_X(t) \neq F_Y(t)$, dvouvýběrovým Kolmogorovovým-Smirnovovým testem: `ks.test(mzda[jih=="Jih"], mzda[jih=="jinde"])`.[4] Zamítá se hypotéza o rovnosti distribucí?

[1] Krabicový diagram mj. zakresluje vybrané výběrové kvantily: obvykle medián (tlustá čára uvnitř krabice) a kvartily (okraje krabice).

[2] Volby `xlim` a `breaks` zajišťují porovnatelnost obou histogramů co do rozmezí osy x a šířky a rozmístění sloupců.

[3] Příkazy `par` střídají barvy kreslených funkcí, aby bylo možné je rozlišit. Volba `cex=0.3` zmenšuje zakreslované body.

[4] `ks.test` si stěžuje, že nelze spočítat p-hodnotu kvůli shodným pozorováním (porušení předpokladu o spojitém rozdělení). Příkazem `sum(table(mzda)>1)` zjistíme, že data obsahují 74 shodných pozorování. Nyní tento problém ignorujme.

(viii) Proveďme klasický dvouvýběrový t-test (za předpokladu shodnosti rozptylů):
`t.test(x=mzda[jih=="Jih"],y=mzda[jih=="jinde"],var.equal=T)`
Jakou hypotézu tento test testuje? Zamítá ji? Co je testová statistika?

(ix) Vrátíme-li se k výsledkům bodu iii, můžeme si být jisti, že předpoklad rovnosti rozptylů není porušen? Co kdybychom provedli asymptotický z-test, který tento předpoklad nemá? Spusťme
`t.test(x=mzda[jih=="Jih"],y=mzda[jih=="jinde"],var.equal=F)`[5]
Liší se testová statistika a p-hodnota od klasického t-testu? Liší se rozhodnutí o zamítnutí H_0?

(x) Proveďme ještě dvouvýběrový Wilcoxonův test: `wilcox.test(mzda[jih=="Jih"], mzda[jih=="jinde"])`. Podívejte se na p-hodnotu. Zamítá se nulová hypotéza? Co vlastně říká nulová hypotéza dvouvýběrového Wilcoxonova testu.[6]

(xi) Připomeňme, že Wilcoxonova testová statistika byla na přednášce definována jako $W_{n,m} = \sum_{i=1}^{n} R_i$, kde R_i jsou pořadí pozorování náhodného výběru $X_1, \ldots, X_n$ mezi všemi $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$. Testová statistika ve výstupu z R není přímo $W_{n,m}$, ale $W_{n,m} - n(n+1)/2$. Ověřte to (statistiku $W_{n,m}$ spočítáte jako `sum(rank(mzda)[jih=="Jih"])` – chápete, jak tento příkaz funguje?).

(xii) Problém různých rozptylů u dvouvýběrového t-testu lze řešit i transformací zkoumané veličiny nějakou funkcí. Opakujte body iii, viii a ix, kde místo mzdy budete analyzovat logaritmus mzdy o základu 10, tj. všude místo `mzda` pište `log10(mzda)`. Zlepšila se rovnost rozptylů? Liší se po transformaci výsledky t-testu a z-testu více nebo méně než před transformací? Jsou hypotézy testované t-testem a z-testem po transformaci ekvivalentní hypotézám testovaným na netransformovaných datech? Pokud ne obecně, za jakých předpokladů by byly?

(xiii) Má smysl opakovat body vii a x pro transformovaná data? Co by vyšlo?

(xiv) Nyní přistoupíme k porovnání středních hodnot několika výběrů pomocí analýzy rozptylu. Podívejme se třeba na vztah mezi rasou a mzdou. Rasa je veličina, která rozděluje pozorování na tři skupiny: Běloch/Hispanik/Jiná. Mají tyto tři skupiny stejnou střední mzdu? Začneme prozkoumáním velikosti skupin (`table(rasa)`) a grafickým znázorněním pozorovaných rozdělení mezd:

```
boxplot(mzda~rasa)

par(mfrow=c(3,1))
hist(mzda[rasa=="Beloch"],breaks=5*(0:10),xlim=range(mzda))
hist(mzda[rasa=="Hispanik"],breaks=5*(0:10),xlim=range(mzda))
hist(mzda[rasa=="Jina"],breaks=5*(0:10),xlim=range(mzda))
par(mfrow=c(1,1))
```

[5] Tento test je zde nazýván Welchův dvouvýběrový t-test, ale nejedná se o nic jiného než o testovou statistiku našeho z-testu s kritickými hodnotami spočtenými z t-rozdělení s aproximovaným počtem stupňů volnosti. Tato statistika nemá t-rozdělení, jedná se vskutku o aproximaci. Pro velké m a n je to totéž, jako používat limitní normální rozdělení.

[6] To, že zkoušíme řadu různých testů stejné nebo podobné hypotézy na jedněch datech, má pouze pedagogické důvody. V praxi bychom si měli předem vybrat jeden určitý test a ten provést.

(xv) Porovnejme aritmetické průměry a odhady rozptylu mzdy ve třech rasových skupinách: `tapply(mzda,rasa,mean)`, `tapply(mzda,rasa,var)`. Vypadají průměrné mzdy podobně? Proč je důležité porovnávat odhadnuté rozptyly?

(xvi) Proveďme analýzu rozptylu. K tomu slouží funkce `aov` (Analysis Of Variance), ale její výsledek musíme ještě zpracovat funkcí `summary`. Uděláme to najednou: `summary(aov(mzda~rasa))`[7]. Rozhodněte, jestli se hypotéza o rovnosti středních hodnot zamítá nebo ne.

(xvii) Porovnejte výstup z funkce `aov` s tabulkou analýzy rozptylu z přednášky:

Zdroj měnlivosti	Součet čtverců	Stupně volnosti	Podíl	F
Skupina	SS_A	$k-1$	$\frac{SS_A}{k-1}$	$\frac{SS_A}{k-1}/\frac{SS_e}{n-k}$
Residuální	SS_e	$n-k$	$\frac{SS_e}{n-k}$	
Celkový	SS_C	$n-1$		

Poznáte, kde se tyto údaje nacházejí v tabulce analýzy rozptylu získané z R, popř. co tam chybí a co přebývá? Dopočítejte $SS_C = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2$.

(xviii) Hodnota distribuční funkce F-rozdělení s n a m stupni volnosti v bodě x se spočítá jako `pf(x,n,m)`; q-kvantil téhož rozdělení jako `qf(q,n,m)`. Pomocí těchto dvou funkcí spočítejte kritickou hodnotu pro zamítání H_0 u právě provedeného testu a ověřte p-hodnotu.

(xix) Opakujte body xv a xvi po zlogaritmování mezd transformací `log10(mzda)`. Změnil se výsledek testu? Která analýza je vhodnější, ta s původními mzdami nebo po zlogaritmování?

(xx) Prozkoumejte, zda mzda souvisí s pracovním zařazením a se sektorem, v němž je respondent zaměstnán. Opakujte pouze body xiv, xv, xvi a xix.

[7] Na levé straně operátoru ~ je veličina obsahující měření Y_{ij}, na pravé straně je veličina definující skupiny.