Cvičení z NSTP097

Empirická distribuční funkce, intervalové odhady

X.

Postupujte podle zadání. 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_2.RData",sep="/")
load(url(datafile))
```

(Objekt `datafile` obsahuje úplnou http adresu souboru). Vyzkoušejte, že se vše povedlo: příkaz `ls()` musí mezi vypsanými objekty ukázat `gnp`, `plotCI` a `ci.asym` (předdefinované funkce) a `pneu` (datový soubor).

Úloha 1: Empirická distribuční funkce

Empirickou distribuční funkci počítá funkce `ecdf(x)`. Její argument `x` je vektor představující náhodný výběr. Jejím výsledkem je objekt, který můžeme pod nějakým jménem uschovat a dále zpracovávat.

Spočtěte a nakreslete empirickou distribuční funkci náhodného výběru z normovaného normálního rozdělení o rozsahu 25 následujícím způsobem:

(i) Vygenerujte a uschovejte náhodný výběr z normovaného normálního rozdělení (`x = rnorm(25)`).
(ii) Spočtěte empirickou distribuční funkci tohoto výběru (`edf = ecdf(x)`).
(iii) Objekt `edf` se chová jako funkce, tj. můžeme spočítat jeho hodnoty v libovolném bodě nebo bodech. Zkuste spočítat `edf(-0.5)`. Co vám tato hodnota říká?
(iv) Udělejte obrázek: stačí napsat `plot(edf)`.

Nyní zopakujte tento postup pro beta rozdělení s parametry $\alpha = \beta = 0.5$ a rozsah výběru $n = 35$ [`rbeta(n,alpha,beta)`].

Podívejme se, jak se při vzrůstajícím počtu pozorování přibližuje empirická distribuční funkce skutečné distribuční funkci. Nakreslíme si empirickou distribuční funkci pro čtyři výběry z $\mathrm{N}(0,1)$ o rozsahu 10, 50, 500 a 2000. Každým obrázkem proložíme skutečnou distribuční funkci a dáme si je na jeden list. Taktéž spočítáme maximální absolutní rozdíl mezi skutečnou a empirickou distribuční funkcí.

Jádrem výpočtu je připravená funkce `gnp`, kterou jste si natáhli z webové adresy. Vypište si, jak vypadá [`print(gnp)`]. Jejím jediným argumentem je rozsah výběru n. Funkce vygeneruje data z $\mathrm{N}(0,1)$ a vyrobí obrázek empirické d.f a skutečné d.f., přitom vrací maximální rozdíl mezi empirickou a skutečnou distribuční funkcí (to je vlastně hodnota Kolmogorovovy-Smirnovovy statistiky pro jednovýběrový test na normované normální rozdělení). Zkuste teď spustit `gnp(20)`.

Teď vyrobíme obrázky empirických distribučních funkcí pro $n = 10$, 50, 500 a 2000. Abychom mohli porovnat výsledky, nejdříve spustíme příkaz `par(mfrow=c(2,2))`, který způsobí, že se všechny čtyři grafy vykreslí do jednoho okna. Pak čtyřikrát zavoláme `gnp` s

argumentem 10, 50, 500 a 2000. Nakonec uvedeme grafiku do původního stavu příkazem `par(mfrow=c(1,1))`. Chápete, co vám výstupy z `gnp` říkají o konsistenci empirické distribuční funkce?

Úloha 2: Odhady

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`).

Kolik bylo pneumatik [`dim(pneu)`]? Zobrazte si celá data – prostě napište `pneu`. Jednotlivé veličiny můžete získat jako `pneu$met.v` a `pneu$met.d`. Abychom k nim měli přímý přístup, napíšeme `attach(pneu)` – nyní stačí napsat `met.v` a `met.d`.

Spočítejte průměry a výběrovou rozptylovou matici obou veličin: `mean(pneu)`, `var(pneu)`. Co odhadují prvky výběrové rozptylové matice? Spočítejte výběrovou korelační matici, tj. `cor(pneu)`.

Graficky porovnejte empirické distribuční funkce obou měření:

```
a = ecdf(met.v)
b = ecdf(met.d)
plot(a)
lines(b,lty=2,col="blue")
```

Spočítejte asymptotické intervaly spolehlivosti pro střední hodnotu obou měření. Použijte funkci `ci.asym(x)`, která počítá tři čísla:

$$\bar{X}_n - t_{n-1}(1-\alpha/2)\frac{S_n}{\sqrt{n}}, \quad \bar{X}_n, \quad \text{a } \bar{X}_n + t_{n-1}(1-\alpha/2)\frac{S_n}{\sqrt{n}}$$

pro $\alpha = 0.05$. Jaká je pravděpodobnost pokrytí skutečné střední hodnoty tímto intervalem?

Úloha 3: Intervaly spolehlivosti

Vygenerujte náhodný výběr o rozsahu $n = 20$ z normálního rozdělení s parametry $\mu = 2$ a $\sigma^2 = 1$ příkazem `smp = rnorm(20,2,1)`. Sestrojte interval spolehlivosti pro μ pomocí funkce `ci.asym` (pro normální rozdělení má tento interval přesně pokrytí $1 - \alpha$).

Nyní získáme $N = 100$ náhodných výběrů o rozsahu $n = 20$ a sestavíme je do matice:

```
nobs = 20
nvyb = 100
data.mat = matrix(rnorm(nobs*nvyb,2,1),nrow=nobs,ncol=nvyb)
```

V řádcích matice jsou pozorování, ve sloupcích výběry. Nyní spočítáme intervaly spolehlivosti (pro $\mathsf{E}\,X = \mu = 2$) pro každý ze 100 výběrů:

```
vs.ci = apply(data.mat,2,ci.asym)
```

[Příkaz `apply(data.mat,2,ci.asym)` spustí funkci `ci.asym` na jednotlivé sloupce matice `data.mat`.] Výsledky si můžete vypsat (`vs.ci`) a nakreslit

```
co = 1:100
plotCI(vs.ci[2,co],uiw=(vs.ci[3,co]-vs.ci[1,co])/2,gap=0.15,sfrac=0.002,
       ylab="Int. spol. pro str. hodnotu",xlab="Vyber")
abline(h=2,col="red")
```

[Grafické okno si můžete myší rozšířit, abyste intervaly lépe viděli.] Vodorovná červená čára vyznačuje skutečnou střední hodnotu. Kolik intervalů by ji mělo pokrývat? Můžeme spočítat, kolik jich ji skutečně pokrývá:

```
sum(vs.ci[1,]<2 & 2<vs.ci[3,])/nvyb
```

Také si můžeme odhadnout střední délku intervalu spolehlivosti:
`mean(vs.ci[3,]-vs.ci[1,])`.

Nyní opravte počet výběrů z $\mathrm{N}(2,1)$ na $N = 1000$ (abychom lépe odhadli jejich pokrytí a střední délku) a udělejte 1000 intervalů spolehlivosti pro výběry o rozsahu 20, 100 a 1000. Sledujte, jak se mění jejich pokrytí a délka v závislosti na velikosti výběru. [Dejte si pozor, abyste omylem nevypisovali matici 1000×1000, která se vám vytvoří. Kód pro vykreslování intervalů, který je uveden výše, kreslí pouze prvních 100 intervalů, nikoli všech 1000, což by bylo nečitelné.]

Nyní změňte rozdělení a opakujte celou úlohu s rozdělením $\exp\{(\}5)$ místo $\mathrm{N}(2,1)$. To můžete generovat příkazem `rexp(nobs*nvyb,5)`. Jaká je skutečná střední hodnota tohoto rozdělení? Jak se mění pokrytí a délka intervalů spolehlivosti pro střední hodnotu v závislosti na velikosti výběru $n = 20, 100, 1000$? Je situace v něčem jiná než u normálního rozdělení? Proč?