Cvičení z NSTP097
13. 12. 2010

Empirická distribuční funkce, intervalové odhady

Doporučený postup práce (*pro ty, kteří si s R-em ještě netykají*)

1. Spusťte program `R` a nastavte si pracovní adresář, který jste si vytvořili na minulém cvičení, např. `H:/NSTP097`.
2. Ze stránky `www.karlin.mff.cuni.cz/~omelka/Vyuka_stp097.php` stáhněte soubor `cviceni2.RData` a uložte jej do pracovního adresáře z bodu (1).
3. Pomocí `File/Load_Workspace...` si načtěte do `R`-ka soubor `cviceni2.RData`.
4. Pomocí `File/New script...` si založte nový zdrojový soubor a uložte si jej např. jako `cviceni2.R`. Do tohoto souboru si pište příkazy, které se Vám osvědčily pro řešení zadaných úkolů.
5. V případě jakýchkoliv nejasností se nebojte obrátit na cvičícího.

Poznámka k nápovědě: Bohužel v počítačové učebně pod operačním systémem Windows nefunguje standardní nápověda, kdy do řádku napíši přímo `?plot` nebo `help(plot)` . Jde využít html nápověda, kterou vyvoláme pomocí `help.start()` , anebo se k ní lze proklikat pomocí menu Help/... Ta však není příliš šikovná, pokud znám přesný název funkce. Jde však použít úkrok stranou. Pomocí programu `putty` (bude asi ve skupině Internet) se přihlásíte k serveru artax.karlin.mff.cuni.cz. Zde zadáte svůj login a heslo a po úspěšném přihlášení se ocitnete na svém unixovém účtu na serveru artax. Program `R` spustíte jednoduše tím, že napíšete do konzole `R` a dáte Enter. Na takto spuštěné instalaci programu `R` by měla nápověda fungovat již standardně, tj. nápovědu k příkazu plot získáte jako `?plot` nebo ekvivalentně `help(plot)`.

Touto barvou budou psány věci, nad kterými byste se měli zamyslet. Touto barvou budou psány vysvětlení k některým příkazům z `R`.

Úloha 1: Empirická distribuční funkce

Zopakování: Nechť $X_1, \dots, X_n$ je náhodný výběr. Potom empirická distribuční funkce je dána vzorcem

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\}$$

a je odhadem teoretické distribuční funkce

$$F(x) = P(X_1 \leq x),$$

kterou zpravidla neznáme.

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. Vzpomeňte si na přesnou definici empirické distribuční funkce $\hat{F}_n$. Uvědomte si, že jde o náhodnou veličinu. Jaké rozdělení má $\hat{F}_n(x)$? Co víte o jejím limitním chování?

Spočtěte a nakreslete si empirickou distribuční funkci náhodného výběru z **normovaného normálního rozdělení**, tj. $\mathsf{N}(0,1)$, o rozsahu 25 pozorování. Postupujte následovně:

1. Vygenerujte a uschovejte si náhodný výběr z normovaného normálního rozdělení
   `x = rnorm(25)`
   Tento příkaz generuje výběr z normálního rozdělení (víme z minula). Jestliže neupřesníme parametry `mean` a `sd`, tj. střední hodnotu a směrodatnou odchylku, bere `R` defaultně $\mathsf{N}(0,1)$. Připomínám možnost nastavení parametru `set.seed` před voláním pseudonáhodných čísel.
2. Spočtěte empirickou distribuční funkci tohoto výběru. Označím ji např. `eF` (můžete použít jiný název)
   `eF = ecdf(x)`
3. Objekt `eF` se chová jako funkce, tj. můžeme spočítat hodnoty v libovolném bodě nebo bodech. Zkuste spočítat hodnotu této empirické distribuční funkce v bodě 0
   `eF(0)`
   Jaká je hodnota skutečné distribuční funkce rozdělení $\mathsf{N}(0,1)$ v bodě 0? (Uvědomte si, že je toto rozdělení symetrické.)
4. Nakreslete obrázek: `plot(eF)`.

Nyní vše zopakujte pro **beta rozdělení** s parametry $\alpha = \beta = 0.5$ a rozsah výběru $n = 35$.
Abychom věděli, s jakým rozdělením pracujeme, vykreslete si nejprve hustotu $\mathsf{B}(0.5, 0.5)$
`curve(dbeta(x,alpha,beta),from=0,to=1)`
Náhodný výběr z $\mathsf{B}(\alpha, \beta)$ o rozsahu n dostanete příkazem
`x=rbeta(n,alpha,beta)`.
Nezapomeňte za parametry `alpha, beta` dosadit, anebo je definovat.
Spočtěte opět distribuční funkci a nakreslete obrázek.

Chování distribuční funkce při rostoucím počtu pozorování:

Úkol: Podívejme se, jak se při vzrůstajícím počtu pozorování přibližuje empirická distribuční funkce skutečné distribuční funkci. Nakreslíme si empirickou distribuční funkci pro čtyři výběry z $\mathsf{N}(0,1)$ 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.

Řešení: Jádrem výpočtu je připravená funkce `gnp`, kterou jste si natáhli ze souboru `cviceni2.RData`. Vypište si, jak vypadá tato funkce.
`print(gnp)`
Jejím jediným argumentem je rozsah výběru n. Funkce vygeneruje data z $\mathsf{N}(0,1)$ a vyrobí obrázek empirické distribuční funkce a skutečné distribuční funkce, přitom vrací maximální rozdíl mezi empirickou a skutečnou distribuční funkcí. Pokuste se z kódu pochopit, co vlastně tato funkce dělá. Chceme-li počítat rozdíl $\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|$, na které body x se můžeme omezit?

Obrázky zakreslíme do jednoho panelu tak, že před voláním funkce `gnp` napíšeme příkaz `par(mfrow=c(2,2))`. Měli jsme už minule. Nastavuje okno pro obrázky, aby se do něj nakreslily 4 obrázky v matici 2x2. Tento příkaz Vám otevře prázdné grafické okno, které nezavírejte.
Pak čtyřikrát zavoláme `gnp` s argumentem 10, 50, 500 a 2000
`gnp(10); gnp(50); gnp(500); gnp(2000)`

Co pozorujete na obrázku? Co o tom víte z teorie? Z které věty vyplývá, že $\hat{F}_n(x) \to F(x)$ při $n \to \infty$ pro všechna $x \in \mathbb{R}$? A která věta hovoří o chování $\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|$?

Nakonec uvedeme grafiku do původního stavu
`par(mfrow=c(1,1))`.

Úloha 2: Odhady

Data `pneu` obsahují 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`).

Zobrazte si celá data: Prostě napište `pneu` . Uvědomte si, že jde o měření stejné věci dvěma různými způsoby. Tj. jeden řádek odpovídá vždy měření na jedné pneumatice. Podle toho, co vidíte, co soudíte o daných dvou způsobech měření?
Kolik bylo celkem pneumatik? Stačí napsat `dim(pneu)` .
Jednotlivé veličiny (sloupce) můžete získat jako `pneu$met.v` a `pneu$met.d`. Abychom k nim měli přímý přístup, zavoláme
`attach(pneu)`
Nyní už stačí psát jen `met.v` a `met.d`.
Toto je mnohem pohodlnější. Už nemusíme psát název dat a $. Pokud by Vám to přestalo vyhovovat, napíšete `detach(pneu)` .

Spočítejte průměry a výběrovou rozptylovou matici obou veličin:
`mean(pneu)`
`var(pneu)`.
Co odhadují prvky výběrové rozptylové matice?
Spočítejte také výběrovou korelační matici
`cor(pneu)`.
Interpretujte tyto hodnoty. Co „měří“ korelace? Očekávali jsme takové hodnoty nebo ne?

Graficky porovnejte empirické distribuční funkce obou měření:

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

Nejprve vykreslíme graf empirické distribuční funkce prvního měření a pak do něj pomocí funkce `lines` přidáme druhou distribuční funkci. Parametr `lty` volí typ čáry — 1 je plná, 2 je přerušovaná, 3 tečkovaná atd. Parametr `col.hor` udává barvu čar a `col.points` barvu bodů.

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

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

pro $\alpha = 0.05$. Ujistěte se, že víte, co znamenají všechny uvedené znaky a písmena, a že je umíte počítat v R. Jak spočtete např. hodnotu $u_{1-\alpha/2}$?
Zavolejte
`ci.asym(met.v); ci.asym(met.d)`
Jaká je pravděpodobnost pokrytí skutečné střední hodnoty tímto intervalem?
Pozor, jedná se o asymptotický interval spolehlivosti, proto je potřeba vše interpretovat „asymptoticky“. V jakých situacích umíte počítat přesný interval spolehlivosti pro střední hodnotu? Přesvědčte se (graficky) o tom, že v tomto našem případě tyto předpoklady nejsou splněny.

Úloha 3: Intervaly spolehlivosti

Vygenerujte náhodný výběr o rozsahu $n = 20$ z normálního rozdělení s parametry $\mu = 2$ a $\sigma^2 = 1$, tj. $\mathsf{N}(2, 1)$, příkazem
`smp = rnorm(20,2,1)`.
Sestrojte asymptotický interval spolehlivosti pro μ pomocí funkce `ci.asym`. Líbí se Vám tento interval? Co od něj „čekáme“?

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)
```

Parametr `nobs` je n, `nvyb` je N. V řádcích matice `data.mat` jsou pozorování, ve sloupcích jsou jednotlivé výběry. Tj. rozměr této matice je `nobs` řádků a `nvyb` sloupců.

Nyní spočítáme intervaly spolehlivosti (pro $\mathsf{E}X = \mu = 2$) pro každý ze 100 výběrů:

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

Funkce `apply` použije na matici `data.mat` po sloupcích (volba 2, volba 1 by odpovídala řádkům) funkci `ci.asym`.

Můžete si tyto intervaly vypsat (pozor, je toho hodně)
`vs.ci`
a nakreslit pomocí funkce `plotCI` (kterou jste získali zavoláním `cviceni2.RData` na začátku této hodiny):

```
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")
```

Funkce `abline` kreslí lineární funkce (tj čáry). Při volbě `h=2` kreslí horizontální, tj. vodorovnou, čáru v bodě 2. Další volby zjistíte vyvoláním nápovědy `?abline`.

Vodorovná červená čára vyznačuje skutečnou střední hodnotu. Kolik intervalů by ji mělo pokrývat? Můžeme spočítat, jaké procento intervalů ji skutečně pokrývá:

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

Nejdříve pomocí funkce `sum` spočítáme, kolik intervalů splňuje podmínku, že dolní mez je menší než 2 a zároveň horní mez je větší než 2. Tím máme počet takových intervalů. Chceme-li procento, musíme to ještě vydělit počtem všech intervalů, tj. `nvyb`.

Jak to tedy je s pokrytím skutečné střední hodnoty? Jak byste vysvětlili rozdíl mezi tím, co vidíte, a tím, co očekáváte?

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

Nyní opravte počet výběrů z $\mathsf{N}(2,1)$ na $N = 1000$ (abychom lépe odhadli jejich pokrytí a střední délku) a udělejte 1000 intervalů spolehlivosti pro výběry o rozsahu 20, 100 a 1000. To znamená, že pro `nobs` rovno postupně 20, 100 a 1000 a `nvyb = 1000` proveďte

```
data.mat = matrix(rnorm(nobs*nvyb,2,1),nrow=nobs,ncol=nvyb)
vs.ci = apply(data.mat,2,ci.asym)
sum(vs.ci[1,]<2 & 2<vs.ci[3,])/nvyb
mean(vs.ci[3,]-vs.ci[1,])
```

Skutečné pokrytí a průměrnou délku intervalů spolehlivosti pro různé rozsahy výběrů přijďte napsat na tabuli.

Jak se mění pokrytí a délka těchto intervalů v závislosti na velikosti výběru? Připomeňte si, jaká je skutečná délka intervalu pro konkrétní n (viz vzorec výše). Jaké bude její chování při $n \to \infty$?

Nyní změňte rozdělení a opakujte celou úlohu s **exponenciálním rozdělením** $\mathsf{Exp}(5)$ místo $\mathsf{N}(2,1)$. Výběr z exponenciálního rozdělení o rozsahu n dostanete pomocí příkazu
`x=rexp(n,5)`.
Spočtěte nejdřív interval spolehlivosti pro jeden výběr o rozsahu $n = 20$. Jaká je skutečná střední hodnota tohoto rozdělení?
Spočítejte si opět 100 intervalů spolehlivosti a nakreslete obrázek, tj.

```
nobs = 20
nvyb = 100
data.mat = matrix(rexp(nobs*nvyb,5),nrow=nobs,ncol=nvyb)
vs.ci = apply(data.mat,2,ci.asym)
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=1/5,col="red")
```

Podívejte se opět, 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? Na čem je založen asymptotický interval spolehlivosti?

Přesný interval spolehlivosti. Pro normální rozdělení $\mathsf{N}(\mu, \sigma^2)$ známe předpisy pro přesné intervaly spolehlivosti pro střední hodnotu μ. V případě, že náš náhodný výběr pochází z normálního rozdělení, použijeme tento přesný interval spolehlivosti (a nikoliv asymptotický). Z přednášky víte, že jiný interval použijeme v situaci, kdy σ^2 známe, a jiný, jestliže rozptyl neznáme. V praxi většinou rozptyl neznáme, proto se budeme dívat na druhý zmíněný případ.

Při neznámém σ^2 je interval spolehlivosti pro μ na hladině spolehlivosti $1 - \alpha$ dán jako

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

Opět se ujistěte, že víte co jednotlivé hodnoty znamenají. Jak byste v `R` počítali $t_{n-1,1-\alpha/2}$?

Vygenerujte si výběr z normálního rozdělení $\mathsf{N}(3, 1)$ o rozsahu $n = 25$
`x=rnorm(25,3,1)`
Interval spolehlivosti pro $\alpha = 0.05$ můžeme dostat zavoláním
`t.test(x)$conf.int[1:2]`.
Funkce `t.test` provádí jednovýběrový a dvouvýběrový t-test, s tím se ještě blíže seznáme později. Teď si z toho pouze vytáhneme interval spolehlivosti, tj. `conf.int`.

Pro rozsahy $n = 25, 100, 1000$ a 5000 postupně vygenerejtu výběr z $\mathsf{N}(3, 1)$ a spočítejte přesný i asymptotický interval spolehlivosti a porovnejte.
Co vidíte? V čem se liší přesný a asymptotický interval (co se týče vzorce)? Co víte o chování t_n-rozdělení při rostoucím n?