

Övningsuppgift 1

Specialskrivna m-filer bl a `jackknif.m`, `jackstat.m`, `bca.m`, `neigen.m`, `eigen.m` och `yuwaest.m` finns att hämta på nätet på kursens hemsida.

<http://www.math.kth.se/matstat/gru/5b1555/uppgift.html>

Typ av redovisning

Uppgifterna bör dokumenteras genom att man redovisar egna m-filer samt printar ut loggen från Matlab som lätt erhålls med kommandot "`diary filnamn`" (stängs av med `diary off`).

Dessutom vill jag att Du skriver en liten beskrivning av vad som gjorts respektive vilka slutsatser som kan dras.

1 Simulering

1)

Vi är intresserade av att undersöka egenskaper hos skattningen $\hat{\lambda} = 1/\bar{x}$ av felintensiteten för exponentialfördelade data då vi har $n = 10$ observationer. Funktionen `exprnd` kan användas (eventuellt `Eexprnd`). Simulera ett stort antal sådana stickprov och undersök fördelningen för $\hat{\lambda}$. Visa resultatet i form av ett histogram och beräkna väntevärde och varians.

Man kan visa att $E(\hat{\lambda}) = \lambda \cdot n/(n-1)$ - stämmer detta med Dina simuleringsresultat?

2)

Simulera stickprov bestående av 20 två-dimensionella normalfördelade variabler som har korrelation ρ . Att få en två-dimensionell observation (X, Y) som har $E(X) = E(Y) = 0$ och $V(X) = V(Y) = 1$ och där korrelationen är ρ kan enligt Choleski-metoden göras genom att skaffa sig två oberoende $N(0, 1)$ -fördelade variabler Z_1 och Z_2 och sen bilda $X = Z_1$ och $Y = \rho Z_1 + \sqrt{1 - \rho^2} Z_2$. Vi skattar ρ med plug-in-skattningen $\hat{\rho}$ dvs

$$\hat{\rho} = \frac{\sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{20} (x_i - \bar{x})^2 \sum_{i=1}^{20} (y_i - \bar{y})^2}}.$$

Simulera detta för $\rho = 0.9$ och gör histogram. Studera också fördelningen för $h(\hat{\rho})$ där

$$h(x) = \frac{1}{2} \ln \left(\frac{1+x}{1-x} \right)$$

dvs Fisher-transformationen (se avsnitt 5.5 i kompendiet). Gör ett histogram och beräkna väntevärde och varians i fördelningen för $h(\hat{\rho})$. Jämför med det teoretiska asymptotiska resultatet angående Fisher-transformationen.

2 Bootstrap - väntevärdeskorrigering

Låt $X_1, X_2, \dots, X_{10}$ vara oberoende likformigt fördelade på intervallet $[0, \theta]$ med $\theta = 1$. Vi låter $X_{(1)}, X_{(2)}, \dots, X_{(10)}$ vara dessa sorterade i storleksordning. Vi skattar θ med plug-in-skattningen $X_{(10)}$.

a) Härled fördelningen för $X_{(10)}$, dvs för maximum av 10 R(0,1)-fördelade. Beräkna väntevärde och varians. Vad är det systematiska felet (bias) för $\hat{\theta} = x_{(10)}$?

b) Visa att vid bootstrap från observationerna $x_1, x_2, \dots, x_{10}$ får vi för $\hat{\theta}^* = X_{(10)}^*$ bootstrapfördelningen

$$P(\hat{\theta}^* = x_{(i)}) = \left(\frac{i}{10}\right)^{10} - \left(\frac{i-1}{10}\right)^{10}, \text{ för } i = 1, 2, \dots, 10.$$

c) Härled den bias-korrigerade skattningen $\bar{\theta}$ av θ baserad på $\hat{\theta} = x_{(10)}$ som erhålls med bootstrap och beräkna dess väntevärde. Jämför med resultatet från a-delen.

d) Simulera fördelningen för $\bar{\theta}$ och beräkna ur detta dess väntevärde och varians samt jämför med motsvarande för $\hat{\theta} = X_{(10)}$.

3 Bootstrap och jackknife

Generera två dataserier med förslagsvis 20 respektive 100 data från någon fördelning, exempelvis $Exp(m)$ -fördelningen eller $N(m, \sigma^2)$. Detta kan göras med rutinerna `exprnd` respektive `normrnd` i Matlabs statistics toolbox (Statsmodulen). Om Du använder filer från som erhållits p annat stt heter de alltså `Eexprnd` respektive `Enormrnd`.

Allt nedanstående behöver inte dokumenteras utan tanken är att Du skall få en känsla för hur bootstrap respektive jackknife fungerar för enkla skattningar samt hur bootstrap-resultaten beror av antalet simulerade stickprov.

3.1 Jackknife

Skatta för de båda dataserierna följande kvantiteter med hjälp av `jackknif` och `jackstat`

Jackknife:

- 1) Medelfel med Jackknife
- 2) Biaskorrekturen med Jackknife

3.2 Bootstrap

Använd t ex $B = 50, 100, 500, 1000$ och (i mån av tid 10000). Notera att man kan utnyttja delserier av samma bootstrap-generering.

- 1) Medelfel med Bootstrap.
- 2) Biaskorrekturen med Bootstrap
- 3) Plotta fördelningen för $\hat{\theta}^*$ dvs bootstrapfördelningen. Detta görs lämpligen med proceduren `hist`.

Gör detta för följande parametrar efter att ha angett hur plug-in-skattningen ser ut.

- a) $\theta = E(X)$ dvs väntevärdet
- b) $\theta = D(X)$ dvs standardavvikelsen
- c) $\theta = \text{medianen}$ dvs 50%-punkten i fördelningen
- d) $\theta = \text{interkvartilavståndet}$, dvs skillnaden mellan de två punkter där 25% av massan ligger till vänster resp. till höger – kräver att man skriver en egen `m`-fil. Man kan använda `prctile` eller `Eprctile`. Ett annat alternativ är att sortera vektorn och ta rätt i storleksordning.

Tips: Man kan skriva en egen `m`-fil som ger de sökta storheterna och där man endast ändrar indata respektive skattningsfunktion när man löser uppgifterna a)-d).

4 Flerdimensionella observationer

Proceduren `neigen.m` (eller `Eneigen.m`) producerar en matris av flerdimensionellt normalfördelade observationer - en per rad - med en föreskriven A -matris. Detta betyder att rad nr k är en observation av $\mathbf{X} = A\mathbf{Z}$ där $\mathbf{Z}$ består av oberoende $N(0,1)$ -fördelade variabler. Alltså har $\mathbf{X}$ kovariansmatrisen $C = AA^T$. Proceduren `neigen.m` ger även egenvärdena för C samt värdet på

$$\theta = \frac{\max(\text{egenvärdena})}{\text{summan av egenvärdena}}$$

som alltså är parametern som skall skattas med hjälp av de genererade observationerna.

Angiv en A -matris samt generera lämpligt antal observationer ($n = 20$ eller $n = 100$ kanske) med hjälp av `neigen.m`.

Skatta sedan θ med hjälp av `eigen.m` (eller `Eeigen.m`) samt utför jackknife respektive bootstrap med hjälp av rutinen `eigen.m` och skatta därigenom medelfelet för $\hat{\theta}$.

Gör även konfidensintervall med percentil-metoden (direkt ur bootstrap-fördelningen alltså) och med BC_a -metoden (använd proceduren `bca.m` eller `Ebca.m`).

4.1 Korrelation

Generera två-dimensionella data - förslagsvis Efrons LSAT/GPA data som kan laddas genom kommandot `load lawdata` i Matlab varvid man får två vektorer `lsat` och `gpa` som innehåller data. Vill man få dessa i en enda vektor `x` kan man skriva

```
x=[lsat gpa]
```

Ett alternativ är att generera två-dimensionellt normalfördelade data med hjälp av `neigen.m` (se ovan).

Skatta ur dessa tvådimensionella data följande kvantiteter:

- Korrelationskoefficient. Proceduren `corrcoef` kan vara till nytta. Tänk dock på att den levererar hela korrelationsmatrisen.
- Skatta parametrarna a och b i enkel linjär regression $y = a + bx$ tillämpat på dessa data. Proceduren `regress` (`Eregress`) kan vara till nytta.

Gör för ovanstående parametrar skattningar av bias och medelfel med hjälp av jackknife och bootstrap (bootstrap av talpar).

5 Regression

Linjär regression med $y = a + bx + cx^2$ som teoretiskt regressions samband där vi är särskilt intresserade av $\theta = -b/2c$ (som ger maximum av regressionslinjen).

Generera t ex $y = 10 + 50x - x^2$ för $x = 1, 2, \dots, 50$ och med slumpfel som är $N(0, 20^2)$.

5.1 Jackknife

Skatta bias för $\theta = -b/2c$ med hjälp av jackknife

Skatta medelfel för $\theta = -b/2c$ med hjälp av jackknife

5.2 Bootstrap

Använd

a) metoden att bootstrappa talpar

b) metoden att bootstrappa residualer
för följande uppgifter:

a) Skatta medelfel och bias för $\theta = -b/2c$

b) Ge ett histogram över bootstrapfördelningen för $\theta = -b/2c$.

Beräkna konfidensintervall för $\theta = -b/2c$ med BC_a -metoden.

6 Pivot-baserade konfidensintervall

1) Generera t ex 20 $\text{Exp}(\theta)$ -fördelade slumpstal med t ex $\theta = 5$. Använd `exprnd`.

a) Beräkna med pivot-metoden ett t ex 90%-igt konfidensintervall för θ baserat på $\hat{\theta}/\theta$.

b) Beräkna ett konfidensintervall med BC_a -metoden.

c) Beräkna ett exakt konfidensintervall med utnyttjande av Γ -fördelningen. Proceduren `gaminv` (`Egaminv`) kan användas.

Jämför bootstrap-fördelningen i a-delen med den sanna fördelningen för $\hat{\theta}/\theta = \bar{X}/\theta$, dvs (i princip!) en Γ -fördelning - använd `gampdf` (`Egampdf`) eller `gamcdf` (`Egamcdf`).

2) Generera t ex 20 Normalfördelade observationer enligt $N(\theta, \sigma)$. Använd `normrnd` (`Enormrnd`).

a) Beräkna ett konfidensintervall för θ baserat på

$$\frac{\hat{\theta} - \theta}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \theta}{\frac{s}{\sqrt{n}}}$$

b) Beräkna ett konfidensintervall med BC_a -metoden

c) Beräkna ett exakt intervall baserat på t -fördelningen - använd `tinvs` eller tabell. Jämför t -fördelningen (använd `tpdf` eller `tcdf`) med bootstrapfördelningen.

3) Skapa två normalfördelade stickprov av längd 10 och 20 från $N(m_1, \sigma_1^2)$ resp $N(m_2, \sigma_2^2)$. Beräkna med pivot-metoden ett konfidensintervall för $m_1 - m_2$ baserat på en lämplig (approximativ) pivot-variabel. Du får inte anta att $\sigma_1 = \sigma_2$.

7 Tidsserier

Generera en $AR(1)$ -tidserie av t ex längd 100 dvs

$$y_{t+1} = ay_t + \epsilon_t, \quad t = 0, 1, 2, \dots, 100$$

med t ex $a = 0.9$ och där ϵ_t :na är oberoende $N(0, 5^2)$. Starta rekursionen med $y_0 = 0$.

Skatta a med m-filen `yuwaest.m` som skattar parametrarna i en $AR(p)$ -process (dvs vi har $p = 1$). Denna fil finns i `funk`-directoriet eller kan laddas ner från hemsidan. `yuwaest` skattar med hjälp av Yule-Walker-ekvationerna med anropet

```
[aest sest C]=yuwaest(y,1)
```

där **aest** är skattningen av $a = 0.9$, **sest** är en skattning av $\sigma = 5$ och **C** är den traditionella skattningen av medelfelet av a -skattningen som t ex lärs ut i kurser om tidsserier.

Du skall nu använda bootstrap för att få en uppfattning om fördelningen för a -skattningen. Generera alltså nya tidsserier genom att med det skattade värdet på a göra bootstrap av de skattade residualerna i den ursprungliga tidsserien. Är skattningen väntevärdesriktig? Vad är dess varians?