TAIU06/9GMA05 DATORÖVNING 2

Datorövningen behandlar konstruktion av konfidensintervall för $\mu_1 - \mu_2$, undersökning av konfidensgrad vid normalapproximation, enkel linjär regression

Förberedelser

Glöm inte att rensa minnet och alla fönster mellan varje uppgift så att det inte finns något gammalt sparat som kan störa dina resultat.

Kopiering av filer: Du kommer att använda filerna `lab2_uppg3.m` och de finns på kurshemsidan.

Redovisning

Datorövningen ska göras, individuellt eller i par om två.

Spara alla filer med kod så att du är beredd att visa den eller testköra under redovisningstillfället. Spara även, eller skriv ut alla plottar.

Uppgift 1.
Konfidensintervall

Fortsättning på uppgift 3 från laboration 1.

De båda s-värdena var ganska lika och vi antar att vi i u och v har observationer från $N(\mu_1, \sigma)$ respektive $N(\mu_2, \sigma)$. Vi vill jämföra μ_1 och μ_2.

Använd funktionen `ttest2` för att bilda ett nedåt begränsat 95% konfidensintervall för $\mu_1 - \mu_2$. Skriv `help ttest2` för information om hur funktionen `ttest2` fungerar. Fundera speciellt på hur man matar in signifikansnivån α och hur man får ett nedåt begränsat konfidensintervall.

Vad blir din slutsats beträffande de båda börsernas volatilitet mätt på det här sättet?

Hur kan man motivera att man gör ett nedåt begränsat intervall? Diskutera.

Uppgift 2.
Konfidensgrad vid normalapproximation

Öppna en ny m-fil och namnge filen till *uppg1.m*.

Då vi konstruerar konfidensintervall i binomialfördelningsfallet så utnyttjar vi normalapproximation och vi kräver att $npq > 10$ där $q = 1 - p$. Det är intressant att se vad som händer med konfidensgraden om detta villkor inte är uppfyllt.

Ställ upp formeln för ett 95% konfidensintervall för p.

a) Vi ska nu skapa 1000 observationer från Bin(16,0.3).

```
n = 16;
p = 0.3;
x = binornd(n,p,1000,1);
```

Nu kan vi beräkna skattningarna för p (1000 gånger)

```
phat = x/n;
```

och de båda gränserna för konfidensintervallet

```
lower_lim = phat - 1.96*sqrt(phat.*(1-phat)/n);
upper_lim = phat + 1.96*sqrt(phat.*(1-phat)/n);
```

Om det sanna värdet $p = 0.3$ ligger nedanför den undre konfidensgränsen eller ovanför den övre, så har intervallet missat sin parameter. Följande kommandon hjälper oss att räkna ut hur många gånger $p = 0.3$ hamnade nedanför den undre konfidengränsen respektive ovanför den övre

```
missar = sum(lower_lim > p) + sum(upper_lim < p)
```

Nu är `missar` antalet intervall som missade det sanna parametervärdet. Jämför detta antal med det förväntade antalet för den aktuella konfidensgraden.

b) Gör om hela proceduren för 1000 observationer från Bin(80,0.3).

Jämför resultaten.

Spara m-filen och fyll i resultaten på resultatbladet.

Uppgift 3.
Enkel linjär regression

En gejser är en varm källa, som mer eller mindre regelbundet får utbrott, varvid vattnet kan spruta högt upp i luften. Old Faithful Geyser i Wyoming är en sådan källa som blivit en turistattraktion. Eftersom det är ganska långt mellan utbrotten, är man intresserad av att kunna prediktera tiden till nästa utbrott och man tror att denna tid beror på längden av det närmast föregående utbrottet. För att konstruera en prognosformel har man registrerat samhörande värden på

$$\begin{aligned} x &= \text{längden av det senaste utbrottet (enhet:min) och} \\ y &= \text{tiden till nästa utbrott (enhet:min),} \end{aligned}$$

under ett antal dagar. Vi skall analysera datamaterialet enligt modellen

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$$

där

$$\varepsilon_1, ..., \varepsilon_n \text{ är oberoende och } N(0, \sigma).$$

På filen *uppg21.m* finns samhörande värden på x och y.

Börja med att plotta y mot x för att se att en linjär regression kan vara intressant att göra.

Även korrelationskoefficienten är ett mått på graden av linjärt samband. Beräkna korrelationen med hjälp av kommandot `corr`.

Gör sedan en regressionsanalys med kommandot `regstats`; enligt

```
stats = regstats(y,x,'linear','all');
betahat = stats.tstat.beta
se = stats.tstat.se
t = stats.tstat.t
s2 = stats.mse
fstat = stats.fstat
```

b) Hur ser den skattade regressionslinjen ut? Skriv upp ekvationen. Plotta den skattade regressionslinjen enligt

```
figure
scatter(x,y,'*')
xlabel('x'), ylabel('y')
hold on
lsline % ls = least square, dvs. minsta-kvadrat anpassning
```

c) I datautskriften hittar du de skattade standardavvikelserna $d(\hat{\beta}_0)$ och $d(\hat{\beta}_1)$, (se = standard error) dvs. $s\sqrt{h_{00}}$ och $s\sqrt{h_{11}}$.

Skriv upp formeln på hur variansen σ^2 skattas, vilken kvadratsumma ska användas. Vad är frihetsgraderna för σ^2-skattningen?

d) Testa på nivån 0.01

$$H_0 : \beta_1 = 0 \text{ mot } H_1 : \beta_1 \neq 0$$

med hjälp av t-test; ledning: teststorheten har beräknats enligt formeln $(\hat{\beta}_1 - 0)/d(\hat{\beta}_1)$.

För att se om normalfördelningsantagandet i modellen är ok kan en residualanalys göras. Residualerna är beräknade i MATLAB. Ta fram dem genom kommandot `stats.r` och plotta med `scatter`, `hist` och `normplot`, tex. enligt

```
residualer = stats.r;

figure
scatter(x,residualer,'filled')
title('Residualer')

figure
hist(residualer)
title('Histogram för residualer')

figure
normplot(residualer)
```

e) Vad visar de olika residualplottarna? Är du nöjd med residualplottarna?

f*) Ett utbrott på 4 minuter har just avslutats. Beräkna ett 95% **prediktionsintervall** för tiden till nästa utbrott. Matrisen $(\mathbf{X}^T\mathbf{X})^{-1}$ kan beräknas enligt

```
XtXinv = stats.covb/stats.mse
```

Fyll i redovisningsbladet, spara koden och spara eller skriv ut plottar.

REDOVISNINGSBLAD

Fyll i namn och personnr med bläck.

1)

.....

2)

.....

Uppgift 1

Konfidensintervall $I_{\mu_1-\mu_2} =$ OK .....

Uppgift 2

Låt x vara observation av X för $X \sim Bin(n, p)$. Ställ upp formeln för ett 95% konfidensintervall för p.

$I_p =$

Det förväntade antalet intervall som missar p om konfidensgraden är 95%: .....

a) Värde på npq: .....

Antalet intervall som missade parametern då $n = 16$: .....

b) Värde på npq: .....

Antalet intervall som missade parametern då $n = 80$: ..... OK .....

Uppgift 2

a) Skattad korrelation mellan y och x: .....

b) Skattad regressionslinje .....

c) $d(\hat{\beta}_0) =$..... $d(\hat{\beta}_1) =$.....

Formel $s^2 =$ —— Frihetsgrad för s^2: .....

d) Teststorheten $T =$..... ; kritisk gräns för $|T|$: ..... ; Förkastas H_0? .....

e) Är residualplottarna OK? ..... OK .....

f) Prediktionsintervall för tiden till nästa utbrott: .....

$(\mathbf{X}^T\mathbf{X})^{-1} =$