
LABORATION 3
MATEMATISK STATISTIK FÖR K, TMA073
DAVID BOLIN

1 Introduktion

Syftet med den här laborationen är att få en djupare förståelse för hypotesprövning och jämförelser samt lära oss hur vi kan använda Matlab för att beräkna dessa storheter. Börja med att gå in på kurshemsidan och ladda ner filen `airquality.mat` som vi kommer behöva senare. Ladda också ner filen `livslangd.mat` som vi använde i förra laborationen.

2 Hypotesprövning

Tidigare tittade vi på ett stickprov med mätningar av livslängder för en viss komponent. Läs in den datan i Matlab igen och plotta den

```
>> load livslangd.mat
>> plot(livslangd, '*')
>> hist(livslangd)
```

Komponenterna anses hålla tillräckligt hög kvalitet om dess förväntade livslängd är 380 dagar och vi vill nu testa om kvalitén är tillräckligt hög. Vi sätter upp ett ensidigt test

$$H_0 : \beta = 380$$

$$H_1 : \beta < 380$$

eftersom vi endast är intresserade av att påvisa om den förväntade livslängden är under acceptabel nivå. Vi beräknade tidigare konfidensintervall för skattningen av väntevärdet och vi beräknar nu ett ensidigt konfidensintervall

$$[-\infty, U] = \left[-\infty, \beta^* + z_\alpha \frac{\beta^*}{\sqrt{n}} \right].$$

Eftersom 380 inte ligger i intervallet kan vi förkasta nollhypotesen och data tycks alltså påvisa att komponenterna inte håller tillräckligt hög kvalitet.

- Beräkna konfidensintervallet med 99-procentig signifikansnivå. Vad blir intervallet? Kan vi förkasta nollhypotesen?
- Gör om testet med en 95-procentig signifikansnivå, kan vi förkasta nollhypotesen på den svagare signifikansnivån?

För att hitta den exakta nivån vi kan förkasta testet på kan vi använda direktmetoden. Vi använder oss av teststorheten

$$T = \frac{\beta^* - 380}{\beta^* / \sqrt{n}}$$

- Vad är det observerade värdet på teststorheten?
- Vad är p-värdet för testet?
- Vilken är den högsta signifikansnivå med vilken vi kan förkasta H_0 ?

3 Jämförelser

PM₁₀ är en typ av luftförorening som är skadlig för människor och som EU har bestämt ett maximalt tillåtet dagligt värde för på $50\mu\text{g}/\text{m}^3$ (EU Council Directive 1999/30/EC). I filen `airquality.mat` finns mätningar av den dagliga halten PM₁₀ i två städer i norra Italien under vintern 2006. Värdena i filen är logaritmerade halter. Börja med att läsa in datan

```
>> load airquality.mat
```

Titta först på halten i den första staden, som finns i vektorn `y1`. När vi gör skattningarna och konfidensintervall antar vi att vi har ett stickprov med oberoende likafördelade mätningar. Det är alltid bra att visualisera datan för att se om det finns trender i den som gör att oberoendeantagandet kanske inte är uppfyllt.

- Plotta datan, tycks det finnas några trender i datan?
- Finns det anledning till oro kring att antagandet om oberoende inte är uppfyllt?

En normalfördelningsmodell passar inte speciellt bra till halten PM₁₀, men logaritmerade halter följer en normalfördelning ganska bra. Vi kan undersöka detta genom att rita ett normalfördelningsdiagram:

```
>> normplot(y1)
```

Att en normalfördelningsmodell passar bra till logaritmerade halter betyder att halten PM₁₀ kan modelleras med en log-normalfördelning, så vi fortsätter jobba i logaritmerad skala och använder oss då av normalfördelningen.

Vi vill nu jämföra de två städerna för att se om det finns någon skillnad i luftkvalité. Låt μ_1 och μ_2 beteckna väntevärdena för (logaritmerad) luftförorening i de två städerna och låt σ_1^2 och σ_2^2 beteckna motsvarande varianser. Vi börjar med att testa om det är rimligt att anta att variansen för luftföroreningen är samma för de två städerna. Vi sätter upp hypotestestet

$$\begin{aligned}H_0 : \sigma_1 &= \sigma_2 \\H_1 : \sigma_1 &\neq \sigma_2\end{aligned}$$

Vi beräknar stickprovsmedelvärde och standardavvikelse för de två stickproven

```
>> mu1 = mean(y1);  
>> mu2 = mean(y2);  
>> s1 = std(y1);  
>> s2 = std(y2);
```

Vi använder oss av ett F-test och beräknar teststorheten

```
>> T = s1^2/s2^2
```

- Beräkna ett konfidensintervall för kvoten σ_1^2/σ_2^2 . Kvantiler till F-fördelningen kan beräknas med funktionen `finv`.
- Beräkna p-värdet för testet med hjälp av funktionen `fcdf`.
- Kan vi förkasta nollhypotesen?
- Testet kan också utföras direkt med hjälp av funktionen `vartest2` i Matlabs statistics toolbox. Se hjälptexten till funktionen för att förstå vad den gör och använd den till att beräkna p-värdet och konfidensintervallet. Får ni samma svar som ovan?

Vi vill nu testa om de två städerna har olika medelhalt luftförorening. Vi vill alltså testa

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

- Beräkna en poolad skattning av variansen.
- Beräkna teststorheten för testet givet att vi antar att varianserna i de två städerna är lika.
- Beräkna p-värdet. Kan vi förkasta nollhypotesen?

Som ni kanske märkt har vi lika många observationer för de två städerna, och mätningarna är parade på så sätt att de är gjorda på samma tidpunkter i de två städerna. Beroende på hur långt från varandra städerna befinner sig skulle detta kunna införa ett beroende mellan de två tidsserierna vilket skulle kunna göra att modellen stickprov i par kan vara lämplig.

- Gör om testet under modellen stickprov i par, förändras resultatet?

3.1 Extra uppgifter

Om ni har tid och vill träna mer på konfidensintervall och tester kan ni:

- Beräkna konfidensintervall för μ_1 och μ_2 .
- Beräkna konfidensintervall för σ_1 och σ_2 .

Baserat på konfidensintervallen för μ_1 och μ_2 , verkar den förväntade nivån på luftföroreningen överstiga gränsvärdet $\log(50)$ i de två städerna? Hur ska vi tolka detta resultat, betyder det att vi har tillräckligt låg halt av luftförorening?