Matematisk Statistik

SF1911 Statistik för bioteknik: Lp 2 2017
Lab 3 för
CBIOT3

Introduktion

Detta är handledningen till Laboration 3. Läs gärna handledningen två gånger. Försäkra dig om att du förstår hur de MATLAB-kommandon som finns i den bifogade koden fungerar. Laborationen bedöms som godkänd eller ej godkänd. Arbete i grupp är tillåtet med **högst två**, personer per grupp. Godkänd laboration ger 2 poäng till ordinarie tentamenstillfälle.

1 Problem 1 - Annova

Börja med att ladda ner clouds.txt ifrån hemsidan. Filen innehåller mätningar på mängden regn i fem områden i Tasmanien mellan 1964 och 1971. Man undersökte i experimentet om användningen av cloud-seeding påverkade mängden regn.

Använd tre-vägs ANOVA för att testa effekten av område, season och seeded. Jämför modellerna med och utan interaktionstermen. MATLABs funktion för n-vägs ANNOVA är **anovan**.

För att kunna använda funktionen så behöver mätningarna och deras kategorier läggas i vektorer. En vektor som innehåller allt mätdata ifrån alla områden och sen en vektor per variabel som innehåller vilken kategori som mätpunkten tillhör, dvs en för område, en för season och en för seeded. Kod för läsa in datat och spara det i varsin vektor finns nedan:

```
clouds=dlmread('clouds.txt','\t',1,0);
rain=[clouds(:,4) ; clouds(:,5) ; clouds(:,6) ; clouds(:,7) ; clouds(:,8)];
seeded=[clouds(:,2) ; clouds(:,2) ; clouds(:,2) ; clouds(:,2) ; clouds(:,2)];
season=[clouds(:,3) ; clouds(:,3) ; clouds(:,3) ; clouds(:,3) ; clouds(:,3)];
area=[zeros(108,1) ; zeros(108,1)+1 ; zeros(108,1)+2 ; zeros(108,1)+3 ...
  ; zeros(108,1)+4];
```

2 Problem 2 - P värden som stokastiska variabler

P värden är stokastiska variabler och i det här problemet kommer vi undersöka hur fördelningen för p-värden beter sig för t-test för ett stickprov. Vi kommer börja med två sidigt test och hypoteserna $H_0\mu = 0$ och $H_a\mu \neq 0$.

Simulerar 10 000 stickprov av storlek $n = 5$, dvs totalt 50 000 värden, med $\sigma = 1$ och med de fyra olika μ värdena, $\mu = -0.5, 0, 0.5$ och 1. Normalfördelningen simuleras med funtkionen *normrnd*. För varje stickprov använd t-test med hypoteserna ovan och okänd varians för att få fram p-värdet för varje enskilt stickprov. Plotta sedan p-värdena för de fyra olika sanna μ i varsitt histogram. T-test med ett stickprov och okänd varians görs i MATLAB med funtkionen *ttest*. Med *[h, p]=ttest(x, m)* så är p-värdet ifrån testet p och m är värdet på μ för nollhypotesen. Alternativ hypotesen är $\mu \neq m$. Beräkna också styrkan för t-testen, men inte när det sanna $\mu = 0$. Detta eftersom vi får α om vi beräknar power för en effektstorlek som är samma som nollhypotesen. Styrkan för t-test på nivån $\alpha = a$ beräknas med
sampsizepwr('t',[mu0 sigma0],mu1,[],n, 'Alpha',a)
där *mu0*, *sigma0* är ifrån nollhypotesen och *mu1* är storleken på μ som vi är intresserad av. I det här fallet är *muH0*=0 och *sigmaH0*=1 ifrån nollhypotesen och *mu1* är μ för den sanna fördelningen. Med de värdena får vi styrkan på att ett t-test förkastar H_0 givet den sanna fördelningen som vi drog vårt slumpdata ifrån.

Vad visar histogrammen? Hur stor andel av p-värdena för varje μ är statistiskt signifkanta på nivån $\alpha = 0.05$? Jämför andelen ni får mot testets styrka(power) och α. Vad händer om ni ökar antalet dragningar i stickproven, dvs ökar n?

Nu ska vi använda samma test men ensidigt istället för två sidigt. Våra hypoteser är då: $H_0\mu \leq 0$ och $H_a\mu > 0$. Det är ett ensidigt test med alternativ hypotesen större än 0, dvs vi är intresserade av sannolikheten för den högra sidan av fördelningen. Det testas i MATLAB genom att vi lägger till *'Tail','right'* till *ttest*. Plotta igen p-värdena i varsitt histogram för varje värde på μ.

Varför blir histogrammet med $\mu = -0.5$ i det ensidiga testen annorlunda än i det två sidiga medan histogrammen för de andra μ inte förändras?

3 Problem 3 - T-test, lika eller olika varians

Vi har tidigare pratat om att t-testet blir olika beroende på om vi antar lika varians eller inte. Det går att göra statistiskt test om det är lika varians eller inte. I det här problemet ska vi undersöka vad som händer om vi först testar för lika varians och sen använder det resultatet för t-testet. Test av lika varians kan görs med funktionen *vartest2*, t-test med två stickprov kan görs med *ttest2*.

De tre testerna ni ska jämför är

- T-test med antagande om lika varians
- T-test utan antagande om olika varians
- T-test med eller utan antagande om lika varians baserat på resultatet av testet av variansen

Vi kommer jämföra de tre testerna genom att simulera data ur två normalfördelningar,$X \approx N(\mu_x, 1)$ och $Y \approx N(\mu_y, \sigma_y^2)$. Dra ur varje fördelning 10 000 stickprov av storlek n_x respektive n_y. Plotta p-värdena för de tre modellerna i en graf med σ_y som x-axel. Variera σ_y ifrån 1 till 3 i steg om 0.5. Jämför mellan $n_x = n_y = 30$ och $n_x = 20, n_y = 40$, samt $\mu_x = \mu_y = 0$ och ett par olika värden på μ. Gör en separat graf över testernas p-värden för de olika μ_x, μ_y, n_x, n_y. Det kan ta ett litet tag för MATLAB att simulera alltihopa eftersom vi totalt drar många slumptal.

Vilket av testen har bäst egenskaper utifrån resultatet av simuleringarna?