MASB11: BIOSTATISTISK GRUNDKURS
DATORLABORATION 3, 14 MAJ 2019
STATISTISKA TEST OCH BÖRJAN PÅ MINIPROJEKT II

Syfte

Syftet med dagens laboration är att du ska

- träna på de grundläggande begreppen inom hypotesprövning (t.ex. signifikansnivå och styrka) samt vilka slutsatser man kan dra från analysen
- bekanta dig med lite av de funktioner som finns i R vad det gäller olika grundläggande statistiska test
- arbeta med kursens Miniprojekt II.

Förberedelseuppgifter

Du måste ha arbetat ordentligt med de väsentligaste begreppen i kapitel 6 och 7 i kursboken. Repetera vid behov begreppen *hypoteser, signifikansnivå, styrkefunktion* samt *modell med matchade data och modell med två oberoende stickprov*.
Du skall ha gjort följande uppgifter *innan* du kommer till laborationen.

Hemuppgift 1:

För att träna på de grundläggande begreppen i hypotest gör uppgifterna **Dig:4.4.1(felrisker)_3, _4 och _5** på övning 7.

Hemuppgift 2:

För att träna på olika modeller gör **Dig:5.2_1 och _2**.

Hemuppgift 3:

(A) Om du använder din egen laptop med R behöver du göra följande:

Installera de R-rutiner som följer med arbetsmaterialet. Tryck på ”ladda ner”-symbolen vid uppgiften så kommer du till bokens lerladdningssida. I högerspalten, ladda ner filen ”raknaMedVariation_0.9-0.tar.gz” och spara den någonstans, t.ex. i ”Hämtade filer”. *OBS!* Om du har en Mac kan du behöva spara ned filen i den map som R letar i som default. Det bör vara den som är default i File panes nere till höger i RStudio.

Starta R-Studio och gå till ”Tools → Install Packages. . . ”. I dialogrutan, välj ”Install from: Package Archive File (.zip, .targz)” och leta reda på filen med ”Browse. . . ”. Låt ”Install to Library:” vara som det är. Tryck ”Install”. Detta behöver du bara göra en gång.

För att kunna använda rutinerna behöver du aktivera dem, enklast genom att välja fliken ”Packages” nere till höger i RStudio och kryssa i rutan för ”raknaMedVariation”. För att se vilka rutiner som finns, ge kommandot `lsf.str(''package:raknaMedVariation'')`.

(B) Om du använder datorsalens datorer ska du *istället* ladda ner de två filerna `hypotes.R` och `styrkefkn.R`, öppna dem i R-studio och köra alla kommandona i dem.

1 Grundläggande begrepp vid hypotestestning

1.1 Exempel: Muntorrhet

Läkemedel kan ge en nedsatt salivkörtelfunktion, vilket är en riskfaktor för karies och andra sjukdomar i munhålan. På 7 slumpmässigt utvalda patienter som alla fick samma medicin mätte man under 5 minuter den så kallade tuggstimulerade saliven. Normal mängd saliv under dessa förhållanden är 1 ml/min och muntorrhet anses föreligga när mängden saliv understiger 0.7 ml/min. Som modell antog man att salivmängden är normalfördelad med väntevärde μ och standardavvikelse σ, där σ anses vara 0.5 ml/min. Intressanta frågeställningar är t.ex.:

- Stöder data vår misstanke att medicinen sänker salivproduktionen?
- Om medicinen ger upphov till en genomsnittlig salivproduktion på 0.8 ml/min, hur troligt är det att vi kommer att missa den nedsatta salivproduktionen med vårt test?
- Hur många patienter ska vi mäta på om vi vill att testet ska upptäcka en nedsatt salivproduktion på 0.7 ml/min med sannolikheten 0.95?

På kursens hemsida hittar ni data i filen `Saliv.RData`. Kortfattade svar till frågorna som ställs i uppgifterna finns i slutet på denna del av handledningen.

Uppgift 1.1:

Först vill man undersöka om data från de 7 patienterna stöder vår misstanke att medicinen sänker salivproduktionen. Ställ upp lämpliga hypoteser.

Uppgift 1.2:

Beräkna medelvärdet av mätningarna.

Uppgift 1.3:

Använd rutinen (A) `hypotes` eller (B) `hypotes1` för att illustrera testets kritiska område då testet utförs på signifikansnivå $\alpha = 0.05$. Det aktuella kommandot är `hypotes(`σ`, `n`, `μ_0`, `α`, H1-riktn)`, så om hypoteserna är $H_0 : \mu = 1$; $H_1 : \mu < 1$ blir kommandot `hypotes(0.5, 7, 1, 0.05, '<')`. (Negligera de felmeddelanden som eventuellt kommer och titta på figuren.)

Uppgift 1.4:

Rutinen markerar det kritiska området och anger ett värde k som är gränsen till området. Hur har k beräknats?

Uppgift 1.5:

Använd ditt beräknade medelvärde för att utföra testet. Vad är din slutsats om H_0?

Uppgift 1.6:

Vad är din konkreta tolkning av signifikansnivån $\alpha = 0.05$ i detta exempel?

Uppgift 1.7:

Undersök hur det kritiska område ändras då du ändrar signifikansnivån till $\alpha = 0.01$. Vad är din slutsats nu?

Testets styrka och styrkefunktion med rutinerna Hypotes.R och Styrkefkn.R

Antag nu att genomsnittlig salivutsöndring i riskgruppen är 0.8. Då är förstås $H_0 : \mu = 1$ falsk och vi vill att vårt test ska upptäcka detta och förkasta denna hypotes till förmån för hypotesen $H_1 : \mu < 1$. Sannolikheten att testet verkligen klarar av detta kallas för testets styrka i punkten 0.8. Använd rutinen (A) `hypotes` eller (B) `hypotes1` för att illustrera testets styrka i punkten 0.8. Kommandot är nu `hypotes(`σ`, `n`, `μ_0`, `α`, H1-riktn, sant `μ`)`, så i detta fall skriver du `hypotes(0.5, 7, 1, 0.05, '<', 0.8)`.

Uppgift 1.8:

Rutinen ger dig ytterligare en figur som, förutom signifikansnivån α (felrisk av typ I), även visar β (felrisk av typ II). Vad är den konkreta tolkningen av β i detta exempel? Hur hänger β ihop med testets styrka?

Mer generellt, testets styrka i punkten μ, är $P(H_0$ förkastas då μ är verklig genomsnittlig salivutsöndring i riskgruppen). Observera att styrkan beror på värdet μ. I detta exempel gäller att ju mindre μ är i förhållande till $\mu_0 = 1$ desto större är chansen att testet ska upptäcka att H_0 inte gäller. Därför är det intressant att studera styrkan som en funktion av μ, denna funktion betecknas ofta $S(\mu)$.

Rutinen (A) `styrkefkn` eller (B) `styrka1` ritar upp styrkefunktionen, kommandot är `styrkefkn(`σ`, `n`, `μ_0`, `α`, H1-riktn, sant `μ`)` om du vill rita upp funktionen och markera ett speciellt μ-värde. Använd alltså kommandot `styrkefkn(0.5, 7, 1, 0.05, '<', 0.8)`, vilket ger dig den tidigare figuren plus styrkan som en funktion av μ.

Uppgift 1.9:

Uppskatta utifrån styrkefunktionen hur stor sannolikheten är att vi med vårt test kommer upptäcka att en grupp som bör klassas som muntorra ($\mu = 0.7$) har en sänkt salivproduktion.

Uppgift 1.10:

Hur många patienter bör vi mäta på om vi med sannolikheten 0.95 verkligen ska upptäcka att muntorra har en sänkt salivproduktion? Tips: Testa med olika värden på n i `styrkefkn`.

Svar till exemplet med muntorrhet:

1.1 $H_0 : \mu = 1$; $H_1 : \mu < 1$

1.2 $k = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 1 - 1.6445 \cdot \frac{0.05}{\sqrt{7}} = 0.689$

1.5 Eftersom medelvärdet < 0.689 förkastas H_0 på nivå 0.05

1.6 Det är 5 % risk att vi påstår att riskgruppen har en sänkt salivproduktion när den i själva verket är normal

1.7 H_0 kan ej förkastas på nivå 0.01

1.8 β = P(ej förkasta H_0 då verklig genomsnittlig salivproduktion i riskgruppen är 0.8) = $1 - S(0.8)$, d.v.s. β är 1- styrkan i punkten 0.8

1.9 Styrkan i punkten 0.7 är $S(0.7)$ vilket enligt figuren kan uppskattas till 0.48

1.10 Det krävs $n = 30$ patienter för att styrkan ska vara 0.95 i punkten 0.7.

2 Några statistiska test i R

Kortfattade svar till frågorna som ställs i uppgifterna finns i slutet på denna del av handledningen. Datamaterialen som används är Albumin, Klorofyll och Dammar, vilka du hittar på kurshemsidan.

2.1 Test av väntevärde i en population (t-test)

Använd datamaterialet `Albumin.RData`. En blandning av blodserum innehåller exakt 42 g albumin per liter. Två laboratorier (A och B) får göra sex bestämningar var av koncentrationen. Vi vill undersöka om det finns någon systematisk avvikelse från det sanna värdet (42 g/l) (tvåsidig mothypotes). I R gör man två t-test med nedanstående kommandon i kommandofönstret (glöm inte att spara det som, t.ex. lab3.R) och kör dem.

```
t.test(AlbuminA, mu=42)
t.test(AlbuminB, mu=42)
```

Uppgift 2.1:

Tolka utskriften. Hur stora är *P*-värdena och vad blir slutsatserna? Vad blir konfidensintervallen för de förväntade koncentrationerna?

Om du vill ha ett ensidigt test använder du ett av kommandona

```
t.test(AlbuminB, mu=42, alternative="less")
t.test(AlbuminB, mu=42, alternative="greater")
```

Uppgift 2.2:

Tolka utskriften och jämför *P*-värde och intervall med det tvåsidiga alternativet.

2.2 Jämförelse av väntevärden i två populationer (t-test vid två oberoende stickprov)

Alger fick växa under ljusa respektive mörka förhållanden. Undersök med ett t-test om det finns skillnader i förväntad klorofyllhalt mellan de två grupperna genom att använda `t.test` igen men nu genom att lägga till gruppvariabeln

```
t.test(Klorofyll$Klorofyll ~ Klorofyll$Grupp)
```

Uppgift 2.3:

Tolka utskriften.

Om vi inte säger något annat förutsätter `t.test` att varianserna i de två grupperna är olika och kompenserar för det. Om vi vet (eller antar) att varianserna är lika kan vi utnyttja det:

```
t.test(Klorofyll$Klorofyll ~ Klorofyll$Grupp, var.equal=TRUE)
```

Om man vill testa att varianserna är olika skriver man

```
var.test(Klorofyll$Klorofyll ~ Klorofyll$Grupp)
```

Uppgift 2.4:

Undersök om varianserna är olika. Vilket test bör användas? Slutsatser från det använda testet?

En grafisk beskrivning av skillnaderna kan fås med en boxplot:

```
plot(Klorofyll$Klorofyll ~ Klorofyll$Grupp)
```

2.3 Test vid matchade data (stickprov i par)

I filen `Dammar.RData` finns mätningar av kvävebelastning vår respektive sommar på ett antal dammar. Om man vill jämföra kvävebelastningen mellan de två årstiderna är en rimlig modell ”stickprov i par”. För att tala om för `t.test` att data är matchade används kommandot

```
t.test(Dammar$N_belast_V, Dammar$N_belast_S, paired=TRUE)
```

Uppgift 2.5:

Finns det några skillnader mellan vår och sommar?

Svar till denna sektions frågor:

2.1 A: $P = 0.033$; B: $P = 0.081$; Vi kan påvisa en skillnad för A men inte för B. 95 % konfidensintervall: A: $(42,06,\ 42.94)$; B: $(35.69,\ 42.51)$

2.4 Varianserna är inte olika ($P = 0.48$); t-test: $P = 0.009$; Vi kan påvisa en skillnad i klorofyll.

2.5 $P = 0.262$ Vi kan inte påvisa någon skillnad

Miniprojekt II – Hade vår aktiveringskampanj effekt?

Målsättningen med denna uppgift är bl.a. att du:

- ska träna på att hämta ett problem ur verkligheten och med hjälp av ett insamlat material konstruera en rimlig statistisk modell samt göra en kritisk granskning av modellen och dess förmåga att beskriva verkligheten;
- ska tillämpa dina kunskaper och med hjälp av R analysera ett biostatistiskt datamaterial;
- ska träna på att skriftligt redovisa antaganden, modeller och slutsatser från en statistisk analys.

Redovisningen görs i form av en skriftlig rapport:

- Maila in rapporten senast **fredag 24 maj kl. 16.00**.
- Skicka rapporten som en pdf-fil till `masb11@matstat.lu.se`.
- **Ämnesrad ska skrivas som** `Miniprojekt2 av studid1 och studid2`

Utformning av projektredovisning

Målgruppen för rapporten är en person med bakgrundskunskaper som en student i samma årskurs, som läst den aktuella kursen men inte är insatt i detaljerna i den aktuella uppgiften.

Rapporten bör vara strukturerad enligt följande:

1. Titelsida med med författarnas namn
2. Kort bakgrund och syfte med undersökningen
3. Redovisning av de uppgifter som finns i problemställningen. Ange (om så är lämpligt):
   - vilka antaganden ni gör om data,
   - vilka hypoteser ni ställer upp,
   - resultatet av analysen och vilka tolkningar och slutsatser ni gör.

   Lämpliga figurer och tabeller ska vara med i rapporten.
4. Sammanfattning av era resultat.

Problemställningar i studien

De flesta forskare anser att hög kolesterolhalt i blodet är en riskfaktor för hjärt- och kärlsjukdomar. I en studie ville vi undersöka om man genom ett aktiveringsprogram bestående av flera faktorer (rökstopp, mental och fysisk träning) kan minska halten av kolesterol. Vi utgick från en grupp på 40 rökande män som samtliga hade något förhöjda kolesterolhalter i blodet. Av dessa 40 valde vi slumpmässigt ut 20 (A-grupp) som fick genomgå vårt aktiveringsprogram. De övriga 20 (B-grupp) levde som vanligt, åtminstone såsom vi uppfattade det. Efter ett halvår mätte vi kolesterolhalten (mmol/l) på samtliga 40 män igen.

I filen `kolesterol` finns samtliga data och variablerna heter `Afore`, `Aefter`, `Bfore` samt `Befter`. Vi har förstås en rad frågeställningar vi vill ha svar på och jag har försökt punkta ner dem:

(a) Eftersom vi slumpmässigt valt ut de 20 som ska genomgå aktiveringsprogrammet bör det inte finnas några skillnader mellan A- och B-gruppen beträffande genomsnittlig kolesterolhalt **innan** studien börjar. Men vi vill verkligen försäkra oss om detta så vi inte från början introducerar en systematisk skillnad mellan grupperna. Kan ni undersöka?

(b) Sänkte aktiveringsprogrammet den förväntade kolesterolhalten hos grupp A?

(c) Det är inte otroligt att patienterna i grupp B, även om de inte genomgår aktiveringsprogrammet, ändå påverkas i sin kolesterolhalt eftersom uppmärksamhet kring frågorna kan ge effekt. Verkar det vara så i vår undersökning?

(d) Kan vi dra slutsatsen att aktiveringsprogrammet påverkar de två grupperna på olika sätt?

(e) En del forskare menar att kolesterolhalten ökar med åldern medan andra anser att faktorer som rökning och stillasittande har större betydelse. I variablerna `Aalder` och `Balder` finns de undersökta männens ålder vid studiens start. Tyder våra data vid studiens start på att åldern påverkar kolesterolhalten?

Tips på arbetsgång

- Börja med att beräkna sammanfattande mått i de olika grupperna och eventuellt rita histogram för att få en överblick av data.
- I frågeställningarna (a)–(d) gäller det att sätta upp rätt modell för data, välj mellan ”två oberoende stickprov” och ”stickprov i par”. Ange modell, hypoteser, beräkningsgång och slutsatser i samtliga fyra frågeställningar.
- Frågeställning (e) gör du lämpligtvis inte förrän efter laboration 4 som handlar om regression. Plotta kolesterolhalt mot ålder och undersök om kolesterolhalten, bortsett från slumpmässiga faktorer, kan beskrivas med en funktion av ålder.