MASB11: BIOSTATISTISK GRUNDKURS
DATORLABORATION 2, 10 APRIL 2019
CGS OCH MINIPROJEKT I

Syfte

Syftet med dagens laboration är att du ska

- illustrera centrala gränsvärdessatsen (CGS) med hjälp av simulering
- använda CGS i en praktisk situation
- arbeta med kursens Miniprojekt I

Förberedelseuppgifter

Du måste ha arbetat ordentligt med de väsentligaste begreppen i kapitel 4.5 och 5.1 i kursboken. Repetera vid behov begreppen *normalfördelning, väntevärde och varians för en summa (eller medelvärde) av oberoende slumpvariabler*.
Du skall ha gjort följande uppgifter *innan* du kommer till laborationen.

Hemuppgift 1:

Gör **Dig:3.3.2_5** och **3.96** och **Dig:3.4.5_1** på övning 5.

Hemuppgift 2:

Försök att med egna ord beskriva vad centrala gränsvärdessatsen säger.

Hemuppgift 3:

Repetera vad du gjorde vid laboration 1 och vilka metoder du använder för att simulera slumptal och för att anpassa en fördelning till data.

1 Centrala gränsvärdessatsen

Adderar man (eller beräknar medelvärdet) av flera oberoende normalfördelade slumpvariabler är summan också normalfördelad. Men vad händer om man lägger ihop flera variabler som alla är rektangelfördelade? Vilken fördelning fås om man adderar exponentialfördelade variabler?

Centrala gränsvärdessatsen (CGS) säger att om man adderar ett stort antal oberoende variabler från en godtycklig fördelning blir summan (eller medelvärdet) normalfördelad. I formler: om n är tillräckligt stort gäller att $Z_n = X_1 + X_2 + \ldots + X_n$ är approximativt normalfördelad **oavsett vilket fördelning** $X_1, \ldots X_n$ **har**. Med några simuleringar ska du undersöka om detta tycks stämma, t.ex. om du utgår från att X-variablerna är rektangelfördelade.

Uppgift 1.1:

Simulera 1000 slumptal från en rektangelfördelning, R(0, 1) och lägg dem i variabeln `unif1` (`unif1 <- runif(1000,0,1)`). Använd `hist()` och `qqnorm()` för att konstatera att slumptalen är rektangelfördelade och definitivt inte passar till en normalfördelning.

Uppgift 1.2:

Simulera 1000 nya slumptal från en rektangelfördelning, R(0, 1) och lägg dem i variabeln `unif2`. Summera sedan de gamla och de nya slumptalen genom `sum12 <- unif1+unif2`. Resultatet är 1000 slumptal från $Z_2 = X_1 + X_2$. Använd `hist()` och `qqnorm()` för att undersöka fördelningen hos denna summa.

Uppgift 1.3:

Skapa även variabler `unif3`, `unif4` och `unif5` på motsvarande sätt och studera fördelningen för $X_1 + X_2 + \ldots + X_5$. Verkar det rimligt att ju större n är, desto bättre kan fördelningen för summan anpassas till en normalfördelning?

Uppgift 1.4:

`I mån av tid:` Pröva vad som händer då du summerar exponentialfördelade slumvariabler med väntevärde 1 (`rexp(1000,1)`). Hur många variabler behövs summeras innan summan kan approximeras med en normalfördelning?

1.1 CGS i praktiken

På 35 patienter med Hodgkins sjukdom mätte man antalet **T4** celler i blodet (antal/mm^3). Samtidigt mätte man motsvarande antal hos 35 patienter som hade andra sjukdomar (Non-Hodgkins). Data ligger i filen **Hodgkindata.RData** som du hittar på kursens hemsida. Läs in data via Environment-fönstrets Öppna-ikon (Workspace i äldre versioner). Du har nu fått två nya variabler **Hodgkin** och **NonHodgkin**.

Uppgift 1.5:

Undersök om antalet celler i blodet är normalfördelat för de båda grupperna.

Uppgift 1.6:

Det är möjligt att jämföra grupperna genom att bilda differensen mellan de två gruppmedelvärdena. Kan du använda dig av centrala gränsvärdessatsen i detta fall? Kan du säga något om vilken fördelning differensen i medelvärden har? Är det ett stort problem att variabeln inte är normalfördelad i de båda grupperna från början? Kan man åtgärda detta på något sätt?

FORTSÄTT MED MINIPROJEKT I PÅ NÄSTA SIDA!

MASB11: BIOSTATISTISK GRUNDKURS VT-19
MINIPROJEKT I – PROVTAGNINGSTIDER
DEADLINE TISDAGEN DEN 16 APRIL KL 16.00

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.

Arbeta i grupper om två studenter. Redovisningen görs i form av en skriftlig rapport:

- Maila in rapporten senast **tisdagen den 16 april kl. 16.00**.
- som **en pdf-fil till masb11@matstat.lu.se**.
- **Ämnesraden ska skrivas som** :
  `Miniprojekt1 av studid1 och studid2`
  där studid1 och studid2 är era StiL-identiteter.

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 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.

Provtagningstider – problemställningar från labbet

Vid vårt kemilaboratorium analyserar vi bland annat en rad prover från den näraliggande provtagningscentralen. Det har framkommit starka önskemål om att vissa patienter som tar prover ska kunna träffa en läkare vid samma besök och inte behöva boka in olika dagar för provtagning och läkarbesök. För att detta önskemål ska uppfyllas måste vi förstås veta hur lång tid det tar från det att provet tagits på patienten till anlyssvaret är klart och undersöka om det är rimligt att låta patienter vänta på provsvar samma dag. Det finns flera moment att ta hänsyn till: provet kan få vänta på provtagningscentralen tills det blir hämtat till vårt laboratorium, det behövs en viss manuell handläggningstid av provet och slutligen har vi själva processtiden i maskinen. Dessutom har vi lite olika hanteringstider beroende på vilken dag i veckan det är och om det är för- eller eftermiddag.

I datafilerna `proverfm.Rdata` och `proverem.Rdata` finns det tider (minuter) som det tog ”från patientarm till analyssvar”. Provet är på så kallad ”allmän kemi” och, som ni märker, har vi har delat upp data i två filer, en för prover tagna på förmiddagen och en för prover tagna på eftermiddagen.

Nu till våra frågor: När vi gjorde ett histogram på data slogs vi av att det inte alls liknade en normalfördelning. Det kan väl i och för sig vara rimligt, men kan vi dra några slutsatser då? Finns det andra fördelningar som kan användas för att modellera tiden? Mer specifikt:

(a) Hur sannolikt är det att en förmiddagspatient får vänta mer än två timmar på analyssvar?

(b) Vi vill kunna säga att ”95 % av förmiddagspatienterna kommer att ha sitt provsvar snabbare än x minuter”. Vad är då x? Om vi vill göra samma uttalande för eftermiddagspatienterna, vad är x då?

(c) Vi beräknar att ca 40 % av proverna kommer på förmiddagen och resten på eftermiddagen. Hur troligt är det att ett patientprov, taget någon gång under dagen, tar mer än två timmar att analysera?

(d) Om vi har 50 patientprover på förmiddagen, vad är sannolikheten att **genomsnittstiden** för dessa 50 prov överstiger en timme?

Tips på arbetsgång

- Titta på data (histogram, empirisk fördelningsfunktion). beräkna enkla mått (medelvärde, standardavvikelse). Jämför förmiddags- och eftermiddagstider.
- Anpassa en lämplig standardfördelning till förmiddagstiderna och skatta parametrarna i den fördelningen.
- Gör samma sak för eftermiddagstiderna.
- Svara på frågorna (a)–(c) genom att utnyttja de anpassade fördelningarna.
- Fundera på vad medelvärdet av 50 förmiddagstider har för fördelning. Använd de ”enkla mått” du beräknade tidigare för att hitta rätt parametrar i fördelningen.