Centrala Gränsvärdessatsen (CGS)

Datorlaboration 2

Skriv ditt namn här

2024-03-26

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

Repetera 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 $Z_n = X_1 + X_2 + ... + X_n$ är approximativt normalfördelad oavsett vilket fördelning $X_1, ..., X_n$ har. Med några simuleringar ska du undersöka om detta tycks stämma, där vi utgår från några olika fördelningar för X-variablerna.

Uppgift 1.1

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

```
unif1 <- runif(1000,0,1)
# skriv din R-kod här
```

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

```
# skriv din R-kod här
```

Uppgift 1.3

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

```
# skriv din R-kod här
```

Svar:

Uppgift 1.4

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

```
# skriv din R-kod här
```

Svar:

> Uppgift 1.5
>
> 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?

```
# skriv din R-kod här
```

Svar:

2 CGS i praktiken

> På 35 patienter med Hodgkins sjukdom mätte man antalet T4 celler i blodet (antal/mm3). Samtidigt mätte man motsvarande antal hos 35 patienter som hade andra sjukdomar (Non-Hodgkins). Data ligger i filen `data/lab2_mini1_filer/Hodgkindata.RData`. Läs in data med `load()`. Du har nu fått två nya variabler Hodgkin och NonHodgkin.

```
load("data/lab2_mini1_filer/Hodgkindata.RData")
```

> Uppgift 2.1
>
> Undersök om antalet celler i blodet är normalfördelat för de båda grupperna.

```
# skriv din R-kod här
```

Svar:

> Uppgift 2.2
>
> Det kan vara intressant att jämföra grupperna genom att bilda differensen mellan de två gruppmedelvärdena. Använd dig av centrala gränsvärdessatsen för att säga något om vilken fördelning differensen i stickprovsmedelvärden har.

Modell:

Låt $\mu_{Hodgkin}$ och $\mu_{NonHodgkin}$ vara väntevärden i de två grupperna.

Låt $\sigma^2_{Hodgkin}$ och $\sigma^2_{NonHodgkin}$ vara varianser i de två grupperna.

Vi har två stickprov $(x_1, ..., x_{35})$ och $(y_1, ..., y_{35})$ från gruppenrna med Hodgkin respektive NonHodgkin.

Differensen mellan stickprovsmedelvärden är $\delta = \bar{x} - \bar{y}$ och har följande fördelning

$\delta \sim$ ... fyll i beskrivningen!

> Uppgift 2.3
>
> Skatta parametrarna (väntevärde och varians) i fördelningen för differensen δ.

```
# skriv din R-kod här
```

Svar:

> Uppgift 2.4
>
> Är det ett stort problem att antal celler i bloder inte är normalfördelad i de båda grupperna från början? Föreställ dig en situtation där det är ett problem och ge ett förslag på hur man kan åtgärda det?

Svar:

> Kör (Render) denna fil till en html (öppna i webbläsare och skriv ut till PDF) eller en PDF och ladda upp din PDF som en laborationsrapport på Canvas. Forsätt sedan med Miniprojekt I under laboration 2!