Laboration 1: Deskriptiv statistik

Statistisk analys

Emil Erikson och Leo Levenius*

2024-06-12

Viktigt: Innan ni läser vidare

Gör först följande:

1. (Om ni inte gjort det tidigare) I den övre menyn i `RStudio`, tryck på `Tools -> Global Options -> Code -> Saving`. Under "Default text encoding", tryck på "Change" och välj "UTF-8".
2. Gå in på kurshemsidan och ladda ner mallen för denna laboration. Öppna den i `RStudio`. Skriv er rapport i denna fil.

Krav för laboration 1

Tänk på följande **krav** på er rapport:

- Rapporten måste kunna läsas av någon som inte har läst labbinstruktionerna. Så ni måste skriva vad det är ni ska göra innan ni gör det, och berätta vad syftet är.
- Rapporten måste vara skriven i `R Markdown`.
- All kod som används måste synas i labbrapporten, men ska inte beskrivas i detalj i rapporten.
- Alla **tabeller och diagram** måste förses med **numrering och beskrivande text**, och refereras till i rapportens vanliga text på rätt sätt. Diagram måste ha lämpliga rubriker på axlarna och tabeller lämpliga rubriker på kolumner.

*Tidigare versioner av Erik Thorsén, Benjamin Allévius och Fredrik Olsson.

Denna laboration

Denna första datorlaboration i kursen **Statistisk analys** består väsentligen av tre olika delar:

1. En introduktion (eller minnesuppfräschning) till `R`, som kommer att användas på samtliga laborationer.
2. En övning i att avgöra huruvida ett stickprov kan anses normalfördelat.
3. En deskriptiv (beskrivande) analys av ett litet datamaterial.

Som förberedelse till laborationen ska den teoretiska Uppgift 1 lösas. Uppgift 2 och 3 skall redovisas skriftligt senast det datum som anges på schemat.

Introduktion till Deskriptiv statistik

I kursen Sannolikhetsteori I fick ni en introduktion till `R`. Alla dokument som användes i denna introduktion finns att hitta även på hemsidan för denna kurs, så ni kan kika i dem för att fräscha upp minnet. Här ska vi istället kika på vad `R` har att erbjuda när det kommer till att beskriva datamängder. För att få en uppfattning om hur ni kan beskriva ett datamaterial numeriskt och grafiskt ska ni genomföra följande enkla uppgift.

I staden Grötköping mättes kroppslängden på 11 av stadens invånare och resultaten (i cm) blev

```
 [1] 174.6 173.2 189.6 167.7 179.2 179.6 170.5 168.5 185.3 164.1 178.4
```

Börja med att mata in data i variabeln `x`. De vanliga läges- och spridningsmåtten kan enkelt fås med hjälp av funktionerna `mean` (för medelvärdet), `var` (för stickprovsvariansen), `sd` (för stickprovsstandardavvikelsen), och `summary` (för minimum, maximum, median och kvartiler).

Ett träd-bladdiagram fås med `stem`, histogram med `hist`, boxplot med `boxplot`, och normalfördelningsplot med `qqnorm`. I samtliga fall räcker det att ange vektorn `x` som enda argument så ritar `R` upp diagrammet automatiskt. Dock vill vi att ni sätter passande rubriker på axlarna när ni skapar diagram—för att se hur ni gör det, titta på hjälpsidan för plotfunktionen, t.ex. genom att skriva `?hist` i `Console` i `RStudio`. För att skapa figurrubrik använd `fig.cap`.

När det gäller histogram kommer ni ibland vilja ha en annan klassindelning än den som `R` ger automatiskt. Pröva exempelvis

```
hist(x, breaks = seq(from = 162, to = 192, by = 5))
```

och

```
hist(x, breaks = seq(from = 162, to = 190, by = 4))
```

och se vad som händer. Här ser vi att vi anger brytpunkterna (argumentet `breaks`) för histogrammet genom att specificera en vektor som i det andra fallet är en talsekvens med 162 som första värde, 190 som sista värde, och med tal däremellan som har avståndet 4 till talen strax före och efter.

När det gäller normalfördelningsplotten så är den konstruerad på ett sådant sätt att data kommer att ligga längs en rät linje om data verkligen är normalfördelade. En sådan jämförelse underlättas om ni ritar ut en rät linje, vilket enkelt kan åstadkommas med kommandot `qqline(x)`. Dvs, ni skriver

```
qqnorm(x)
qqline(x)
```

Uppgift 1: Två teorifrågor (behöver ej redovisas skriftligt)

1. Om en exponentialfördelad slumpvariabel har väntevärde a, vad är då dess standardavvikelse?
2. Den stokatiska variabeln X är likformigt fördelad med väntevärde a. Vad ska fördelningens övre och undre gränser vara, uttryckt i a, för att standardavvikelsen ska bli lika stor som väntevärdet? Ledning: Om $X \sim \mathrm{U}[\alpha, \beta]$, så gäller att $\mathrm{Var}(X) = (\beta - \alpha)^2/12$.

Notera: beteckningen $\mathrm{U}[\alpha, \beta]$ är ekvivalent med beteckningen $\mathrm{Re}[\alpha, \beta]$ eller $\mathrm{Re}(\alpha, \beta)$ ni har stött på tidigare. Den likformiga fördelningen kallas "the uniform distribution" på engelska, därav U:et. Ibland ser man också $\mathrm{Uniform}(\alpha, \beta)$ eller $\mathrm{Uniform}[\alpha, \beta]$.

Uppgift 2: Kommer data från en normalfördelning?

När ni i framtiden kommer att syssla med praktiska tillämpningar av matematisk statistik kommer ni troligtvis ställas inför frågan om en uppsättning data kan anses komma från en normalfördelning (eventuellt med viss approximation). I det fallet har ni således (eller förhoppningsvis) ett stickprov med n oberoende observationer från en okänd sannolikhetsfördelning. Frågan är om den okända fördelningen kan vara en normalfördelning.

Vi ska här jämföra olika metoder (främst grafiska) som kan användas för att besvara ovanstående fråga. Vi ska också försöka svara på frågan om hur stort n behöver vara för att vi med rimlig approximation ska kunna avgöra om data är normalfördelade eller inte. För detta ändamål ska vi simulera data från för oss kända fördelningar, både normalfördelade och icke-normalfördelade data. Alla fördelningar som ni ska jämföra ska ha väntevärde och standardavvikelse lika med a, där a är de två sista siffrorna i erat personnummer (om ni jobbar i par, välj den enas personnummer att jobba med).

Frågor uppgift 2

De frågor ni ska besvara i denna uppgift är följande:

1. Vilken är det minsta stickprovsstorleken som behövs för att den **fördelning** ni simulerar från skall avslöja sig som normal eller icke-normal?
2. Vilken grafisk metod anser ni är mest effektiv för att avgöra om ett stickprov är normalfördelat eller inte? Motivera med de olika grafiska metoderna (boxplot, histogram, normalfördelningsplott).

I uppgifter 2.1–2.3 som följer nedan simulerar vi från olika **fördelningar**. I deluppgift 1 introducerar vi också er till de grafiska metoder ni behöver för att kunna slutföra uppgiften. Svara på ovan ställda frågor i skrift för deluppgifter 2.1—2.3 nedan.

- **Svara på frågorna i löpande text, inte i listform**.
- **Motivera era svar**.

Uppgiftskrav

Vi ställer följande **krav** på uppgift 2:

1. **Visa endast plottar för *ett* värde på er stickprovsstorlek** för var och en utav *normal-, likformig-*, och *exponential*-fördelningarna. (något av dem som ni har att välja bland). Dock måste ni på egen hand titta på plottar för olika värden på n för att kunna svara på frågorna.
2. I de diagram ni tar med i rapporten måste ni ange passande **rubriker på axlarna** och **numrera dessa diagram samt ge dem beskrivande texter!** *Tips*: Använd `fig.cap` för automatisk numrering.

Uppgift 2.1: Normalfördelade data

Normalfördelade data kan enkelt simuleras i `R` med funktionen `rnorm(n, m, s)`, där `n` anger antal värden som skall simuleras, `m` anger väntevärde, och `s` anger standardavvikelse för normalfördelningen som skall simuleras från. Börja med att simulera ett stickprov med storlek 10 enligt

```
set.seed(19690420) # fyll i ditt egna födelsedatum. Om ni jobbar i par, välj den enas.
x1 <- rnorm(10, a, a)
```

där ni har tilldelat `a` värdet enligt ovan (och fyllt i ert födelsedatum som argument till funktionen `set.seed`).

Vi kan nu göra en jämförelse mellan slumpdatan vi just simulerade och den sanna normalfördelningen, genom att plotta ett histogram med densitet på y-axeln, och ovanpå detta lägga grafen för normalfördelningens täthetsfunktion:

```
hist(x1, prob = TRUE)
x <- seq(from = low, to = high, length.out = 100)
lines(x, dnorm(x, a, a))
```

Här måste ni definiera värdena på `low` och `high`, som är undre respektive övre gränserna mellan vilka normalfördelningen kommer att ritas för i histogrammet. Hitta passande värden för dessa. När vi skriver `lines(x, dnorm(x, a, a))` ritar vi en kurva med x-koordinater definierade av variabeln `x` och y-koordinater `dnorm(x, a, a)`, vilket är värdet av täthetsfunktionen i punkterna `x` för en normalfördelning med väntevärde och standardavvikelse båda lika med `a`. Rita även en boxplot och en normalfördelningsplot för data:

```
# Boxplot: https://en.wikipedia.org/wiki/Box_plot
boxplot(x1)

# Normalfördelningsplot (Q-Q plot): https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot
qqnorm(x1)
qqline(x1)
```

Simulera sedan ytterligare sju stickprov `x2`, `x3`, ..., `x8` av storlek 10 med samma väntevärde och standardavvikelse som ovan:

```
set.seed(19690420) # fyll i ditt egna födelsedatum. Om ni jobbar i par, välj den enas.
x1 <- rnorm(10, a, a) # kommer att bli samma stickprov som ovan då vi har samma seed
x2 <- rnorm(10, a, a) # detta blir ett stickprov annorlunda från x1, likaså de nedan
x3 <- rnorm(10, a, a)
# osv (fyll i resten själv)
```

Börja med att jämföra stickproven med en gemensam boxplot enligt

```
boxplot(x1, x2, x3, x4, x5, x6, x7, x8)
```

Lägg märke till att de åtta stickproven verkar skilja sig åt ganska markant trots att de alla är simulerade från samma fördelning. För att få en motsvarande jämförelse mellan histogram måste vi först tala om för `R` att dela upp det grafiska fönstret i mindre delfönster, sedan ge kommandona för att rita upp histogram, och till sist säga åt `R` att sluta förvänta sig fler plottar till samma fönster:

```
old_par <- par(mfrow = c(2, 4)) # 2 rader, 4 kolonner
hist(x1)
hist(x2)
hist(x3)
# osv (fyll i resten själv)
par(old_par) # säg åt R att sluta förvänta sig fler plottar till samma fönster
```

Uppgift 2.2: Likformigt fördelade data

För simulering av likformigt fördelade stickprov används i `R`-funktionen `runif(n, min, max)`, där `n` anger stickprovsstorleken, `min` och `max` anger undre respektive övre gräns i det intervall som fördelningen är definierad för. Lösningen på Uppgift 1 ger er intervallgränserna för det värde på a som ni använder. Använd funktionen `runif` för att simulera fem oberoende stickprov `u1`, `u2`, osv av storleken 10 och jämför dem grafiskt; dels med varandra och dels med de normalfördelade stickproven ovan. Rita exempelvis upp fem normalfördelade och fem likformigt fördelade stickprov i samma plot.

När ni simulerar de fem oberoende stickproven, använd först funktionen `set.seed` på samma sätt som för de normalfördelade stickproven ovan, d.v.s. använd `set.seed` med ert födelsedatum **en** gång ovanför definitionerna av `u1`, `u2` etc, i samma stycke kod. Upprepa er analys med stickprovsstorlekarna $n = 20$, $n = 100$ och ytterligare något värde på n som ni väljer själva. Besvara sedan frågorna som tidigare ställts.

Uppgift 2.3: Exponentialfördelade data

Genomför samma jämförelse även för exponentialfördelade stickprov med hjälp av funktionen `rexp(n, r)`, där `n` är stickprovsstorleken och `r` är intensiteten (1 genom väntevärdet) för exponentialfördelningen. Hitta på lämpliga namn för dessa stickprov och se till att använda `set.seed` på samma sätt som ovan. Öka er stickprovsstorlek succesivt och besvara frågorna som tidigare ställts.

Uppgift 3: Explorativ dataanalys

Ni ska nu undersöka ett verkligt datamaterial med de grafiska metoderna ovan samt med så kallade scatterplots (spridningsdiagram) som illustrerar beroenden mellan variabler på ett bra sätt. Ett spridningsdiagram är helt enkelt det ni får genom att skriva

```
plot(x, y)
```

och som ni vet från kursen Sannolikhetsteori I så kan ni ge fler argument till denna funktion för att göra plotten snyggare. **Kom till exempel ihåg att ni måste ange passande rubriker på axlarna, och glöm absolut inte att numrera diagrammet och att ge det en beskrivande text!** *Tips*: Använd `fig.cap` för automatisk numrering.

På kursens hemsida finns filen `olvinsprit.csv` som innehåller data över genomsnittlig konsumtion av öl, vin och starksprit i några OECD-länder. Börja med att spara filen i **samma mapp** som ni placerat er `.Rmd`-fil för labbrapporten i, exempelvis `Documents/Kurser/statan1/Labb1` eller liknande. Se till att filen sparas som en `.csv`-fil och inget annat (som t.ex. `.php`). För att vara övertydlig: spara filen som `olvinsprit.csv`. Vi kan därefter läsa in den i `R` genom

```
data <- read.csv("olvinsprit.csv", header = TRUE)
```

Skapa därefter de fyra variablerna

```
land <- data$Land
beer <- data$beer
vin <- data$vin
sprit <- data$sprit
```

Nu kan vi rita en scatterplot med kommandot

```
plot(beer, vin) # När ni gör detta själv, ange passande axelrubriker etc
```

vilket i det här fallet ger ölkonsumtionen och vinkonsumtionen för samtliga länder. Vill ni tydligare se vilka länder punkterna motsvarar skriver ni

```
plot(beer, vin) # När ni gör detta själv, ange passande axelrubriker etc
text(beer, vin, land)
```

så dyker namnen på länderna upp istället för punkterna.

Frågor uppgift 3

Illustrera nu fördelningen för de tre variablerna var för sig. Uppgift 3 ska redovisas skriftligt. Svara på följande frågor **i löpande text, inte i punktform**.

1. Vilka alkoholtyper i `olvinsprit.csv` kan anses komma från en normalfördelning?
2. Hur ligger Sverige till i förhållande till andra länder inom de olika alkoholtyperna? Är Sverige extremt åt något håll?
3. Vilka länder kan anses extrema? Åt vilket håll är de extrema?
4. Finns det gemensamma drag hos de extrema länderna?
5. Medför hög konsumtion av en typ av alkohol en högre eller lägre konsumtion av de andra typerna? Påverkar de varandra alls?

6. Sammanfatta era slutsatser om alkoholkonsumtionen i OECD-länderna.

- Kom ihåg att **numrera och ge en beskrivande text** till **alla** diagram. *Tips*: Använd `fig.cap` för automatisk numrering.