Forudsætninger til statistisk inferens - stikprøvefordelinger

I denne øvelse undersøger vi måderne, hvorpå statistikkerne fra en tilfældig stikprøve af data kan tjene som punktestimater for populationsparametre. Vi er interesseret i at lære om estimatets egenskaber.

Data

Vi anvender ejendomsdata fra byen Ames, Iowa. Oplysningerne om enhver fast ejendomstransaktion i Ames registreres af City Assessor's kontor. Vores særlige fokus i denne øvelse er salg af boliger i Ames mellem årene 2006 og 2010. Vi vil gerne undersøge boligsalget ved at tage en mindre stikprøve fra hele befolkningen.

```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.
RData")

load("ames.RData")
```

Vi ser, at der er mange variable i datasættet, at vi derfor kan lave en dybtgående analyse. Her vil vi begrænse vores opmærksomhed til to af variablene: boligarealet over huset på kvadratmeter (Gr.Liv.Area) og salgsprisen (SalePrice). For at gøre det lettere for os vil vi oprette to variable med korte navne, der repræsenterer disse to variable.

```
area <- ames$Gr.Liv.Area

price <- ames$SalePrice
```

Lad os se på fordelingen af området i vores befolkning af hjemmesalg ved at beregne et par opsummerende statistikker og lave et histogram.

```
summary(area)

hist(area)
```

1. Beskriv fordelingen af populationen.

Den ukendte prøveudtagningsfordeling

I netop dette tilfælde har vi adgang til hele populationen, hvilket sjældent er tilfældet i det virkelige liv. Det er ofte ekstremt dyrt eller umuligt at indsamle oplysninger om hele populationen. Derfor tager vi ofte blot en stikprøve til at udtale os om populationens egenskaber.
Hvis vi er interesseret i at estimere det gennemsnitlige areal i en stikprøve, kan vi bruge følgende kommando til at undersøge dette.

```
samp1 <- sample(area, 50)
```

Med denne kommando samler vi en simpel tilfældig prøve i størrelse 50 fra vektor `area`, som er tildelt `samp1`. Dette svarer til at gå ind i City Assessors database og hente data fra 50 tilfældige hjemmesalg. Det vil være betydeligt nemmere kun at arbejde med disse 50 filer fremfor alle 2930 hjemmesalg.

2. Beskriv fordelingen af denne prøve. Hvordan kan det sammenlignes med befolkningens fordeling?

Hvis vi er interesseret i at estimere det gennemsnitlige areal i huse i Ames ved hjælp af prøven, er vores bedste enkleste gæt gennemsnittet.

```
mean(samp1)
```

Afhængigt af hvilke 50 hjem, der vælges, kan dette estimat ligge lidt over eller lidt under det sande populationsgennemsnit på 1499,69 kvadratfod. Generelt viser det sig, at gennemsnittet af stikprøver er et ret godt skøn over det gennemsnitlige areal, og vi var i stand til at få dette estimat blot ved at tage en stikprøve på under 3% af befolkningen.

2. Tag en anden prøve, også i størrelse 50, og kalder den `samp2`. Hvordan sammenlignes gennemsnittet af `samp2` med gennemsnittet af `samp1`? Antag, at vi tog to prøver mere, en bestående af 100 enheder og en i størrelsesorden af 1000. Hvilken tror vi vil give det mest nøjagtige skøn over populationsgennemsnittet?

Ikke overraskende, hver gang vi tager en anden tilfældig prøve, får vi et andet gennemsnit. Det er nyttigt at få en fornemmelse af, hvor meget variation vi skal forvente, når vi estimerer befolkningen. Fordelingen af stikprøvefordelingen kan hjælpe os til at forstå denne variation. Da vi her har kendskab til hele populationen, kan vi udtage flere forskellige stikprøver ved at gentage ovenstående trin mange gange. Her vil vi generere 5000 prøver og beregne gennemsnittet.

```
sample_means50 <- rep(NA, 5000)

for(i in 1:5000){
   samp <- sample(area, 50)
   sample_means50[i] <- mean(samp)
   }

hist(sample_means50)
```

Hvis du gerne vil justere bredden på histogrammet for at vise det lidt mere detaljeret, så kan du gøre det ved at ændre `breaks`.

```
hist(sample_means50, breaks = 25)
```

Her bruger vi R til at tage 5000 prøver i størrelse 50 fra populationen, beregne middelværdien af hver prøve og gemme hvert resultat i en vektor kaldet `sample_means50`. På den næste side gennemgår vi, hvordan dette kodesæt fungerer.

3. Hvor mange elementer er der i `sample_means50`? Beskriv stikprøvefordelingen, og sørg for specifikt at bemærke dens middelværdi. Kan du forvente, at fordelingen ændrer sig, hvis vi i stedet indsamler 50.000 stikprøvegennemsnit?

Interlude: For `for` loop

Vi har lige kørt den første `for` loop. Ideen bag `for` loop er *iteration*: hvor formålet med *iterationer* er at udføre de samme koder så mange gange du vil uden at skulle programmere enhver iteration. I tilfældet ovenfor ønskede vi at gentage de to kodelinjer inde i tuborgklammerne, hvor vi tager en tilfældig prøve af størrelse 50 fra `area` og gemme gennemsnittet heraf i `sample_means50`-vektoren. Uden `for` loop vil det kræve meget arbejde:

```
sample_means50 <- rep(NA, 5000)

samp <- sample(area, 50)
sample_means50[1] <- mean(samp)

samp <- sample(area, 50)
sample_means50[2] <- mean(samp)

samp <- sample(area, 50)
sample_means50[3] <- mean(samp)

samp <- sample(area, 50)
sample_means50[4] <- mean(samp)
```

og så videre ….

Med `for` loopen komprimeres disse tusinder af kodelinjer til en få linjer. Vi kan tilføje en ekstra linje til koden herunder, der udskriver variablen i under hver iteration af `for`-loop. Kør denne kode.

```
sample_means50 <- rep(NA, 5000)

for(i in 1:5000){
   samp <- sample(area, 50)
   sample_means50[i] <- mean(samp)
   print(i)
   }
```

Lad os overveje denne kode linje for linje for at finde ud af, hvad den gør. I den første linje `initialiserede vi en vektor`. I dette tilfælde oprettede vi en vektor på 5000 nuller kaldet `sample_means50`. I denne vektor gemmes værdierne genereret i `for`-loopen.

Syntaksen i den anden linje skal `for` loopen forstås således: Syntaksen kan læses som "for hvert element `i` fra 1 til 5000, kør følgende kodelinjer". Du kan tænke på `I` som tælleren, der holder styr på, hvilken loop du er på. Derfor, mere præcist, løkken løber en gang, når `i = 1`, derefter én gang, når `i = 2`, og så videre op til `i = 5000`.

Indholdet i `for`-loop er den del inden i de krøllede klammer, og dette sæt koder køres for hver værdi af `i`. Her på hver løkke tager vi en tilfældig prøve i størrelse 50 fra `area`, tager dens gennemsnit og opbevarer den som \ (i \) det element i `sample_means50`.

Ud over at `for`-loop giver mulighed for at køre koden 5000 gange, så er formålet også, at vi indsamler resultaterne i den tomme vektor, som vi dannede i starten.

5. For at være sikker på, at du forstår dette, skal du prøve at lave en loop i en mindre version. Start med at lave en vektor, bestående af 100 nuller, hvilken kaldes `sample_means_small`. Kør en loop bestående af 50 stikprøver fra `area` og lager dem i `sample_means_small`, men lav kun iterationer fra 1 til 100. Print outputtet til din skærm (skriv `sample_means_small` i konsollen og tryk enter). Hvor mange elementer er der I dette objekt kaldet `sample_means_small`? Hvad præsenterer hvert objekt?

Stikprøvestørrelse og stikprøvefordeling

Lad os anvende et `for loop`: til at beregne stikprøvefordelingen. Lav først et histogram over stikprøvegennemsnittene:

```
hist(sample_means50)
```

Den stikprøvefordeling, vi beregner, fortæller os, hvordan det gennemsnitlige areal i Ames fordeler sig. Da stikprøvegennemsnittet er en unbiased estimator, så er stikprøvefordelingen centreret omkring den sande gennemsnitlige værdi og spredningen af fordelingen indikerer, hvor stor variation der er fremkaldt ved kun at udtage prøver bestående af 50 boliger.

For at få en fornemmelse af den effekt, som stikprøvestørrelsen har på vores fordeling, så lad os udtage yderligere to stikprøvefordelinger: en baseret på en stikprøvestørrelse på 10 og en anden baseret på en stikprøvestørrelse på 100.

```
sample_means10 <- rep(NA, 5000)
sample_means100 <- rep(NA, 5000)

for(i in 1:5000){
  samp <- sample(area, 10)
  sample_means10[i] <- mean(samp)
  samp <- sample(area, 100)
  sample_means100[i] <- mean(samp)
}
```

Her er vi i stand til at bruge en enkelt loop til at skabe to fordelinger ved at tilføje ekstra linjer inden for de krøllede parenteser. Her er det muligt at anvende `samp` til to forskellige objekter. I den anden kommandolinje, hvor `for` loop anvendes, gemmes gennemsnittene i vektoren `sample_means10`. Når gennemsnittet er gemt, kan objektet `samp` overskrives med den nye stikprøve, denne gang en stikprøve af størrelsen 100. Generelt, hver gang vi laver et objekt med et navn, hvor navnet allerede anvendes, så vil det gamle objekt blive erstattet af et nyt objekt.

For at se hvilken effekt forskellige størrelser af stikprøver har på stikprøvefordelingen, så plot 3 fordelinger oven på hinanden.

```
par(mfrow = c(3, 1))

xlimits <- range(sample_means10)

hist(sample_means10, breaks = 20, xlim = xlimits)
hist(sample_means50, breaks = 20, xlim = xlimits)
hist(sample_means100, breaks = 20, xlim = xlimits)
```

Første kommando specificerer at du ønsker at dele området med at plotte ind I 3 rækker og 1 kolonne, (ønsker vi at komme tilbage til default setting, så skrives `par(mfrow = c(1, 1))`). Kommandoen `breaks` specificerer antallet af søjler i histogrammet. Argumentet `xlim` specificerer området af x - aksen I histogrammet, og ved at sætte det lig med `xlimits` for hvert histogram, så sikrer vi, at alle 3 histogrammer vil blive plottet inden for samme grænser på x - aksen.

6. Når stikprøvestørrelsen bliver større, hvad sker der med middelværdien? Hvad sker der med spredningen?

Opgave

I det ovenstående er der kun blevet set på variable `area`. Nu vil vi se på priserne.

- Tag en simple tilfældig stikprøve på 50 fra `price`. Ud fra denne stikprøve, hvad er dit bedste gæt på et punktestimat af populationsmiddelværdien?
- Da du har adgang til populationen, så simuler stikprøvefordelingen for $\bar{x}_{price}$ ved at tage 5000 stikprøver fra population med en størrelse på 50. Lager disse gennemsnit i en vektor kaldet `sample_means50`. Plot data, og beskriv formen af stikprøvefordelingen. Baseret på denne stikprøvefordeling, hvad er dit gæt på middelværdien af prisen? Beregn nu den sande middelværdi af populationen.
- Ændr stikprøvestørrelsen fra 50 til 150, beregn stikprøvefordelingen ved brug af ovenstående metode, lager disse gennemsnit i en ny vektor kaldet `sample_means150`. Beskriv formen af denne stikprøvefordeling, og sammenlign den med en stikprøvefordeling bestående af stikprøver med 50. Hvad er dit gæt af middelværdien af prisen for boliger I Ames?
- Ud fra en stikprøvefordeling fra 2 og 3, hvilken har den mindste spredning? Hvis vi ønsker os at lave estimater, der er tættest muligt på den sande værdi, ønsker vi så en stor eller lille spredning?