Datorlaboration 3

Ulf Högnäs

Översikt

Innehåll

- Hypotest och p-värden
  - med randomization
  - med sannolikhetsmodell
- Konfidensintervall
  - med the Bootstap
  - med sannolikhetsmodell
- t-fördelningen och funktionen `t.test()`

Vi kommer att fokusera på andelar, precis som i föreläsning 5 och 6. Slutet av labben kommer att behandla test och konfidensintervall för sådant som inte är andelar. Vi kommer att introducera normalfördelningens kusin, t-fördelningen.

Problem 1 - Hypotestest för andel

1.1 EUs medborgarinitiativ

Få mer att säga till om i frågor som berör dig direkt. Med ett europeiskt medborgarinitiativ kan du bidra till EU-politiken genom att uppmana EU-kommissionen att föreslå nya lagar.

medborgarinitiativ

För att få igenom ett medborgarinitiativ krävs namnunderskrifter

Du måste få stöd från minst en miljon EU-medborgare och samla in ett minsta antal underskrifter i minst sju EU-länder

För Sverige är det minsta antalet 15 120. Låt oss säga att vi driver intiativet **Stop Destroying Videogames** och att vi har samlat in 18 910 underskrifter. Vi vet att vissa av dessa

underskrifter är ogiltiga. Det kan till exempel vara för att den som skrivit på inte har fyllt 18 år. Om 80% eller fler av underskrifterna är giltiga så har vi nått spärren på 15 120 eftersom

$$18910 \cdot 0.80 = 15128$$

Vi har inte tid att kontrollera samtliga underskrifter, så vi kontrollerar istället ett slumpmässig urval på 200 underskrifter. I vårt slumpmässiga urval så finner vi att **176 underskrifter är giltiga**, medan resterande **23 är ogiltiga**.

> **i** Uppgift 1.1
>
> Använd R för att beräkna stickprovsandelen. Spara resultatet som en variabel.

Vi börjar med hypotestest med randomization.

1. Antag att 80% av namnunderskrifterna är ogiltiga
2. Skriv "giltig" på 160 kort och "ogiltig" på 40 kort
3. Dra ett kort med återläggning 200 gånger, blanda mellan varje dragning.
4. Beräkna andelen som blev "giltig"
5. Upprepa steg 3 och 4 tiotusen gånger.

Detta tar för lång tid, så vi gör detta i R istället. Funktionen `rbinom()` skapar den slumpmässig vektor med dragningar av detta slag. Namnet kommer från orden *random* och *binomial*. Binomialfördelningen är en sannolikhetfördelning som vi har valt att inte ha med i kursen, trots att den är viktig. Vi nöjer oss med att säga att binomialfördelningen besvarar frågor av typen *om jag genomför ett och samma försök tio gånger, oberoende av varandra, vad är sannolikheten att jag lyckas minst åtta gånger?*

Här är en rad kod som genomför randomizeringen 10 000 gånger och sparar resultaten under namnet `MI_random`.

```
# n = antalet simuleringar
# size = antalet "dragna kort" per simulering
# prob = sannolikheten att "lyckas", i vårt fall att få "giltig"
MI_random <- rbinom(n = 1e4, size = 200, prob = .8)
# titta på de första 20 resultaten
head(MI_random, 20)
```

```
 [1] 157 155 152 158 161 160 162 157 161 157 152 168 151 155 156 163 159 153 172
[20] 160
```

Nu vill vi gå från antal till andel. Andel i detta fall är ju hur stor del av 200 som blev "giltig". Vi delar däför varje antal i vår resultatvektor med 200:

```r
MI_proportions <- MI_random/200
head(MI_proportions, 20)
```

```
 [1] 0.785 0.775 0.760 0.790 0.805 0.800 0.810 0.785 0.805 0.785 0.760 0.840
[13] 0.755 0.775 0.780 0.815 0.795 0.765 0.860 0.800
```

Uppgift 1.2

Skapa ett histogram över de simulerade andelarna. Antingen med `hist()` i base-R eller med `histogram()` från `mosaic` i lab 2. Ändra antalet `breaks` tills du tycker att det ser bra ut.
Vi har ju simulerat resultat under antagandet att 80% av underskrifterna är giltiga. Titta på histogrammet och jämför med den stickprovsandel som vi beräknade i Uppgift 1.1. Vad har vi visat med denna simulering? Förklara!

Nu ska vi skatta (uppskatta) ett p-värde. Först behöver vi hypoteser.
Låt p vara andelen ogiltiga underskrifter bland de 18 890.

$H_0 : p = 0.80$
$H_0 : p > 0.80$

Vi har också valt en gräns för statistisk signifikans. Eftersom vi vill vara nästan säkra på att vi har tillräckligt många underskrifter valde vi en låg gräns, 1%.

Vårt p-värde blir svaret på frågan *om den verkliga andelen giltiga underskrifter är 80%, hur ofta hade vi sett 176 giltiga eller fler, i ett stickprov på 200?*

Uppgift 1.3

Använd resultatet från simuleringen (`MI_proportions`) för att skatta p-värdet. Jämför p-värdet med vår gräns på 5%. Dra en slutsats.

Tips!

Du kan beräkna antalet eller andelen platser i en vektor som är större eller lika med ett visst tal på följande sätt

```r
# skapa en vektor som du kallar "numbers", att testa med
numbers <- c(0, 7, 1, 2, 1, 8, 7, 1, 0, 0)
# antalet platser som är större än 5
sum(numbers>=5)
```

```
[1] 3
```

```r
# andelen platser som är större än 5
mean(numbers>=5)
```

```
[1] 0.3
```

Syftet med att noggrant gå igenom simulering av andel är att det ska öka er förståelse för sannolikheter och abstrakta begrepp som p-värde. I praktiken kan vi använda `prop.test()` istället.

> **Uppgift 1.4**
>
> Upprepa testet med ‘`prop.test()`.

Problem 2 - Hypotestest för andel