Laboration 2: Statistisk hypotesprövning

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. Döp den tilll "labb2-efternamn1-efternamn2.Rmd". Öppna den i `RStudio`. Skriv er rapport i denna fil.

Krav för laboration 2

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 figurer** måste förses med **numrering och beskrivande text**, och refereras till i rapportens vanliga text på rätt sätt. Figurer måste ha lämpliga rubriker på axlarna och tabeller lämpliga rubriker på kolumner.

*Tidigare versioner av Benjamin Kjellson och Fredrik Olsson.

Sammanfattning av laboration 2

Huvudsyftet med denna andra datorlaboration är att träna förmågan att genomföra statistiska tester genom att sätta upp lämpliga hypoteser och välja ett bra test baserat på beskrivningen av ett visst problem och egenskaperna hos ett datamaterial. Laborationen består av ett avsnitt där det beskrivs hur man använder `R` för några av de statistiska test som vi har gått igenom i kursen. Sedan följer två uppgifter som skall genomföras och redovisas. Varje uppgift består av en teoridel, som skall lösas innan man sätter sig framför datorn, och en praktisk del som skall lösas i `R`. Labbrapporten skall skrivas i `R Markdown`.

Avsnitt 1: Statistisk hypotesprövning i R

1.1: Test av väntevärdet för ett stort stickprov

Om n är tillräckligt stort för ett stickprov $X_1, X_2, ..., X_n$ bestående av oberoende och likafördelade stokastiska variabler så kan teststatistikan

$$Z = \frac{\bar{X} - \mu}{D} \sim \text{approx. } \mathcal{N}(0,1)$$

användas för test av väntevärdet, som här tar värdet μ under nollhypotesen. Här menas att teststatistikan är approximativt normalfördelad. Talet $D = \sqrt{Var(\hat{X})}$ är standardavvikelsen för stickprovsmedelvärdet. Kritiska gränser för standardnormalfördelningen kan fås i `R` med funktionen `qnorm(p)` som ger kvantilen för sannolikheten p. Kvantilen är alltså det tal q för vilket arean (=sannolikheten) under normalfördelningens täthetsfunktion till vänster om punkten q är lika med p. Vill man exempelvis få kritiska gränsen för ett tvåsidigt test på 5%-nivån skriver man alltså

```
qnorm(0.975)
```

```
## [1] 1.959964
```

Gränserna är alltså ± 1.959964.

För kritiska gränser för ensidiga test på 5%-nivån skriver man

```
qnorm(0.95)
```

```
## [1] 1.644854
```

eller

```
qnorm(0.05)
```

```
## [1] -1.644854
```

beroende på riktining hos mothypotesen.

För att beräkna p-värdet så kan kommandot `pnorm(z)` användas, där `z` är det observerade värdet på teststatistikan Z. Om vi exempelvis fått $z = 1.72$ får vi p-värdet för ett tvåsidigt test enligt

```
z <- 1.72
2 * (1- pnorm(z))
```

```
## [1] 0.08543244
```

och ett ensidigt test enligt

```
z <- 1.72
1 - pnorm(z)
```

```
## [1] 0.04271622
```

Observera att vi måste ta `1 - pnorm(z)` för att få rätt värde, vilket beror på att pnorm ger den kumulativa fördelningsfunktionen för normalfördelningen. Skriv `?Normal`, `?qnorm`, eller `?pnorm` i konsollen i `R` för att få detaljerad information om dessa funktioner och några till (de ligger på samma hjälpsida).

1.2: Test av väntevärdet för ett normalfördelat stickprov

Om stickprovet kan antas följa en normalfördelning bör vi istället genomföra ett t-test utifrån teststatistikan

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

I R finns funktionen `t.test` som underlättar genomförandet av alla former av t-test. Som ni ser (om ni skriver `?t.test` i Console) finns det en lång rad argument som kan användas för att specificera saker som värdet på $\mu = \mu_0$, vilka typer av hypoteser ni vill testa, konfidensgrad och mycket annat. Om ni exempelvis har lagrat stickprovet i vektorn x och vill testa hypoteserna

$$H_0: \ \mu = 4.5$$
$$H_1: \ \mu > 4.5$$

på 10%-nivån skriver ni

```
t.test(x, alternative = "greater", mu = 4.5, conf.level = 0.9)
```

Vill ni ha en ensidig mothypotes åt andra hållet så byter man ut textsträngen `greater` mot `less`, och vill ni ha en tvåsidig mothypotes byter ni ut den mot 'two.sided. Som en bonus får ni motsvarande konfidensintervall när ni använder denna funktion.

1.3: Test av skillnader i väntevärden för två oberoende stora stickprov

Om vi har två oberoende stickprov

- $X_1, X_2, \ldots, X_{n_1}$ med väntevärde μ_1, och
- $Y_1, Y_2, \ldots, Y_{n_1}$ med väntevärde μ_2,

och vi vill testa skillnaden $\mu_1 - \mu_2$ använder vi för stora stickprov teststatistikan

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_1)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}},$$

där S_1 och S_2 är stickprovsvarianserna för respektive stickprov. Kritiska gränser och p-värden beräknas på samma sätt som i Avsnitt 1.1 ovan.

1.4 Test av skillnader i väntevärden för två oberoende normalfördelade stickprov

För att få `R` att genomföra t-test för två stickprov räcker det att anropa funktionen `t.test`med två datavektorer `x` och `y` enligt

```
t.test(x, y, alternative = "greater", mu = 0, conf.level = 0.9)
```

Här anger argumentet `mu = 0` att skillnaden mellan väntevärdena är noll under nollhypotesen. Om man inte särskilt specificerar något annat så använder `R` teststatistikan

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim \text{approx. } t_\nu$$

där antalet frihetsgrader ν beräknas med Welch–Satterthwaites metod. För att få ett exakt t-test under förutsättningen att varianserna σ_1^2 och σ_2^2 är lika baserat på teststatistikan

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t_{n_1+n_2-2}$$

måste man ange det särskilt

```
t.test(x, y, alternative = "greater", mu = 0, var.equal = TRUE, conf.level = 0.9)
```

1.5: Test av skillnader i väntevärden för två parvist beroende stickprov

Om vi har två lika stora stickprov $X_1, X_2, \ldots, X_n$ och $Y_1, Y_2, \ldots, Y_n$ där X_i och Y_i är beroende för alla $i = 1, 2, \ldots, n$ kan vi betrakta problemet som ett enstickprovsproblem baserat på de parvisa skillnaderna $D_i = X_i - Y_i$. I `R` kan vi hantera detta genom att först beräkna

```
d <- x - y
```

och sedan hantera problemet som i avsnitt 1.1 eller 1.2 beroende på vilka förutsättningar som är uppfyllda. Man kan även genomföra ett t-test på parvist beroende stickprov genom att lägga till argumentet `paired = TRUE` enligt

```
t.test(x, y, alternative = "greater", mu = 0, paired = TRUE, conf.level = 0.9)
```

1.6: Wilcoxons teckenrangtest

Om vi har ett stickprov som inte kan anses vara normalfördelat bör vi istället genomföra ett icke-parametriskt test. Vi kommer att hoppa över teckentestet i den här laborationen, på grund av att det inte finns[1] någon färdig funktion i `R` för detta test och att det blir svårt att hantera när vi har många "ties". Istället koncentrerar vi oss på Wilcoxons teckenrangtest av medianen. I `R` kan vi använda funktionen `wilcox.test`. Skriv `?wilcox.test` i konsollen för en detaljerad beskrivning av hur denna funktion används. Som ni ser finns liknande argument som för `t.test` plus några till som styr approximativ beräkning av kritiska gränser och p-värde. Dessutom måste ni specifikt ange om ni vill ha ett konfidensintervall för medianen genom argumentet `conf.int = TRUE`.

1.7: Wilcoxon-Mann-Whitneytest

Samma funktion `wilcox.test` kan användas för test av två stickprov på liknande sätt som för `t.test` genom att ange två datavektorer istället för en som argument.

Sammanfattning

Ovanstående avsnitt visar hur `R` kan användas för att genomföra några av de statistiska test som vi har gått igenom i kursen. Det är nu dags för er att själva välja ut lämpliga test för att lösa uppgifterna nedan.

Uppgift 1: Molnsådd i Arizona

Under somrarna 1957–1960 genomfördes ett antal försök i Arizonas bergstrakter för att se om molnsådd kunde öka mängden nederbörd i torra ökenområden. Molnsådd ("cloud seeding" på engelska) innebär att moln beströs från flygplan med kristaller bestående av kolsyresnö. Meteorologerna som ansvarade för försöket hade anledning att tro att kolsyran skulle öka kondensationen i molnen och att detta skulle framkalla regn.

Försöket lades upp på så sätt att försöksperioden under varje sommar delades in i ett antal mindre tvådagarsperioder. Under varje sådan tvådagarsperiod valdes en av dagarna ut slumpmässigt, och på den dagen

[1] Nja, det går nog att hitta i något `R`-paket eller på nätet.

utfördes molnsådd. På den andra dagen utfördes ingen molnsådd, för att på så sätt få ett jämförelsematerial. De dagar då molnsådd genomfördes startades arbetet klockan 12 och molnen beströddes under två timmar. Nederbörden mättes sedan under eftermiddagen med hjälp av 29 stycken mätstationer. I filen `arizona.txt` på kursens hemsidan finns resultatet (i inches) av dessa mätningar i kronologisk ordning. Varje rad avser en tvådagarsperiod där första kolumnen anger årtal, andra kolumnen anger nederbörd under den dag då molnsådd genomfördes, och tredje kolumnen anger nederbörd under den dag då molnsådd inte genomfördes.

Uppgift 1.1: Teoretisk uppgift

Besvara följande frågor:

1. Vad är vitsen med att dela in hela sommaren i tvådagarsperioder istället för att från början bestämma på vilka dagar molnsådd ska genomföras? Ett alternativt tillvägagångssätt skulle ju ha kunnat vara att varje morgon singla slant och genomföra molnsådd om utfallet var klave.
2. Varför väljs vilken dag i en tvådagarsperiod molnsådd ska genomföras slumpmässigt istället för att alltid välja exempelvis den första dagen?
3. Vilket eller vilka test är lämpliga att använda för att testa om molnsådd ökar mängden nederbörd?

Uppgift 1.2: Praktisk uppgift

Börja med att gå in på kurshemsidan och leta upp och spara filen `arizona.csv` (med exakt det namnet och formatet) i samma mapp som ni sparar `.Rmd`-filen ni skriver denna labb i; förslagsvis i `Documents/Kurser/statan/Labb2` eller liknande. Läs in filen genom att skriva:

```
arizona <- read.csv("arizona.csv", header = FALSE)
```

Skapa sedan variablerna:

```
year <- arizona$V1
seed <- arizona$V2
nonseed <- arizona$V3
```

Undersök detta datamaterial grafiskt med hjälp av *histogram*, *boxplottar* och *normalfördelningsplottar* för att få en uppfattning om en eventuell fördelning. Genomför sedan ett (eller flera) lämpliga hypotestest för att avgöra om det kan anses statistiskt säkerställt att molnsådd ökar nederbörden.

Besvara följande frågor:

1. Hade datamaterialet någon fördelning? Vilken?
2. Vad är er nollhypotes samt alternativa hypotes?
3. Vilket test utför ni? Vad gav testet för resultat? Förkastar ni nollhypotesen?

Uppgift 2: Molnsådd i Oregon

Ett annat försök med molnsådd under ungefär samma period genomfördes i delstaten Oregon på ett något annorlunda sätt. Varje morgon fick en meteorolog göra en bedömning om förutsättningarna för nederbörd var lämpliga senare under dagen. Om så var fallet fattades beslut med hjälp av slumptalsgenerator om att genomföra ett försök den aktuella dagen, där sannolikheten för försök var 2/3 och sannolikheten att avstå från försök var 1/3. Detta resulterade i 22 dagar då molnsådd genomfördes och 13 dagar då molnsådd inte genomfördes. Nederbörden mättes sedan i tre olika områden, där data från två av områdena finns med i datamaterialet. Den första typen av område var stora områden i vindriktningen från de moln som behandlades och den andra typen av område var mindre delområden som av olika skäl ansågs särskilt känsliga för molnsådd.

I filen `oregon.csv` på kursens hemsida finns data över försöket. I första kolumnen anger 1 att molnsådd *inte* genomfördes och 2 att det genomfördes, andra kolumnen anger nedebörd i områden av första typen och tredje kolumnen nederbörd i områden av andra typen.

Uppgift 2.1: Teoretisk uppgift

Svara på följande frågor.

1. Vad är en rimlig anledning till varför en slumptalsgenerator användes, istället för att låta exempelvis meteorologen avgöra när ett försök skulle genomföras?
2. Vilket eller vilka test är lämpliga att använda för att testa om molnsådd ökar mängden nederbörd? Varför?

Uppgift 2.2: Praktisk uppgift

Börja med att spara filen oregon.csv på samma sätt som tidigare, och läs in filen i `R`. Skapa sedan variablerna

```
trial <- oregon$V1
typ1 <- oregon$V2
typ2 <- oregon$V3
```

För att få data för dagar då försök genomfördes och dagar då försök inte genomfördes måste ni dela upp variablerna `typ1` och `typ2` enligt följande:

```
nonseed1 <- typ1[trial == 1]
seed1 <- typ1[trial == 2]

nonseed2 <- typ2[trial == 1]
seed2 <- typ2[trial == 2]
```

Att skriva ett logiskt uttryck (här t.ex. ifall `trial` är lika med 1, vilket kan vara `TRUE` eller `FALSE`) inom hakparenteser efter en variabel i `R` medför att endast de värden i variabeln där det logiska uttrycket är sant tas med. Undersök även dessa data grafiskt med hjälp av *histogram*, *boxplottar* och *normalfördelningsplottar* för att få en uppfattning om eventuell fördelning. Genomför sedan ett (eller flera) lämpliga hypotestest för att avgöra om det kan anses statistiskt säkerställt att molnsådd ökar nederbörden.

Besvara följande frågor:

1. Hade datamaterialet någon fördelning? Vilken?
2. Vad är er nollhypotes samt alternativa hypotes?
3. Vilket test utför ni? Vad gav testet för resultat? Förkastar ni nollhypotesen?

Skriftlig laborationsrapport

Uppgifter 1 och 2 ovan skall redovisas skriftligt i en strukturerad och genomtänkt laborationsrapport. Utöver *kraven* som gavs överst i denna instruktion, så kommer här några till:

- Redogör ordentligt för era svar på de teoretiska frågorna och motivera ordentligt vilket eller vilka test ni har valt att använda i de båda uppgifterna.
- Bifoga gärna någon eller några illustrativa figurer, men inte för många, och referera ordentligt i texten vilka slutsatser ni drar av respektive figur.
- Sammanfatta de viktigaste resultaten från körningar av testen i `R` och de slutsatser ni kan dra från dem.
- Använd rösten och läs upp det ni har skrivit för någon annan eller för er själva och fundera på om er meningsbyggnad är vettig eller ej.
- 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.