Laboration 2

Förnamn Efternamn

2024-12-16

Sammanfattning

Uppgift 1

Eftersom datan är parvist beroende skulle det vara trevligt att ha ett nytt stickprov som bara består av differansen mellan paren av dagar. Dessa nya datapunkter som består av differansen kommer då att vara sinsemellan obereonde. Vad vi är intresserade av är ju om det finns en skillnad i medelvärde mellan de två stickproven. Det vill säga, vår nollhypotes blir $H_0 : \mu_0 = \mu_{nonseed} - \mu_{seed} = 0$, och vår alternativa hypotes, $H_1 : \mu_0 \neq 0$.

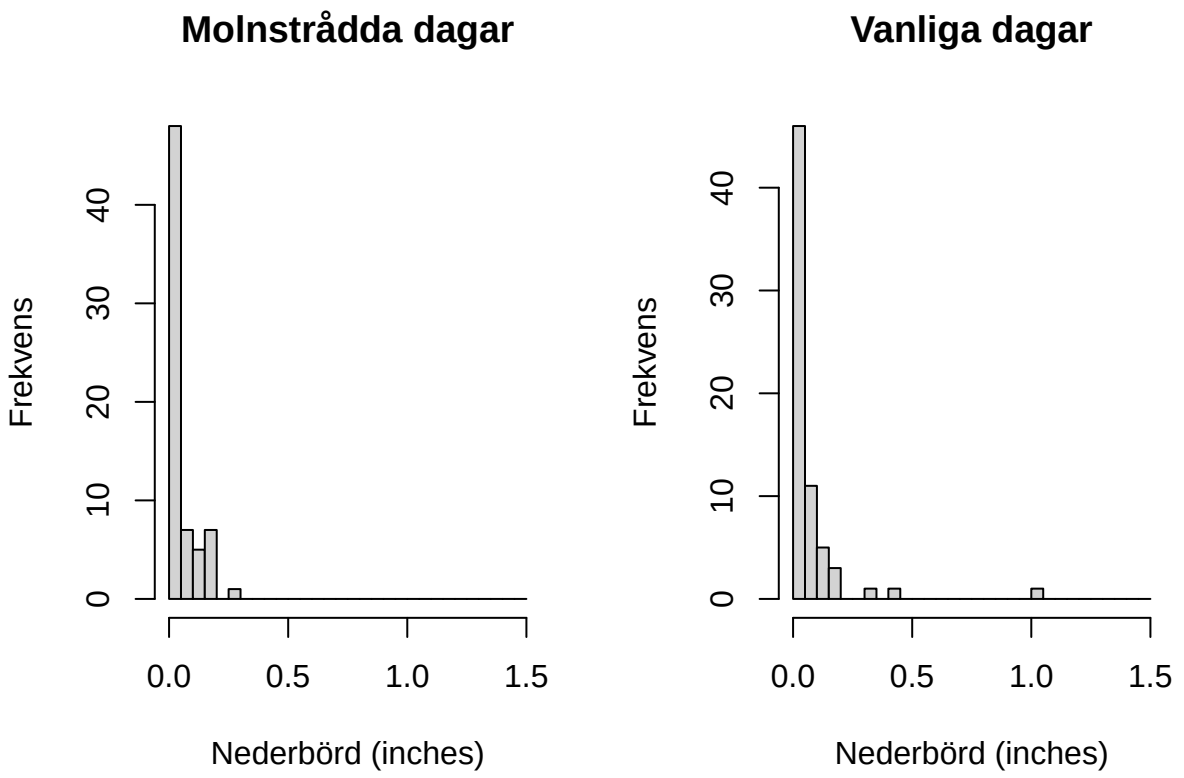
Vi är nu redo att avläsa datan vi har till hand, och vi gör detta enligt instruktionerna, men vi passar även på att skapa en differans-vektor som består av skillnaden i nederbörd mellan dagarna i varsin tvådagarsperiod.

```
arizona <- read.csv("arizona.csv", header = FALSE)
year <- arizona$V1
seed <- arizona$V2
nonseed <- arizona$V3
 difference <- nonseed - seed
```

Vi börjar med att undersöka om båda stickproven följer samma fördelning, och isåfall, vilken. Vi anar att histogramen kommer duga för detta, så vi plottar de sida vid sida (tillsammans hjälps de åt).

```
old_par <-par(mfrow = c(1,2))
hist(seed, breaks = seq(from = 0, to = 1.5, by = 0.05), xlab = "Nederbörd (inches)",
     ylab = "Frekvens", main = "Molnströdda dagar")

hist(nonseed, breaks = seq(from = 0, to = 1.5, by = 0.05),
     xlab = "Nederbörd (inches)",
     ylab = "Frekvens",
     main = "Vanliga dagar")
```

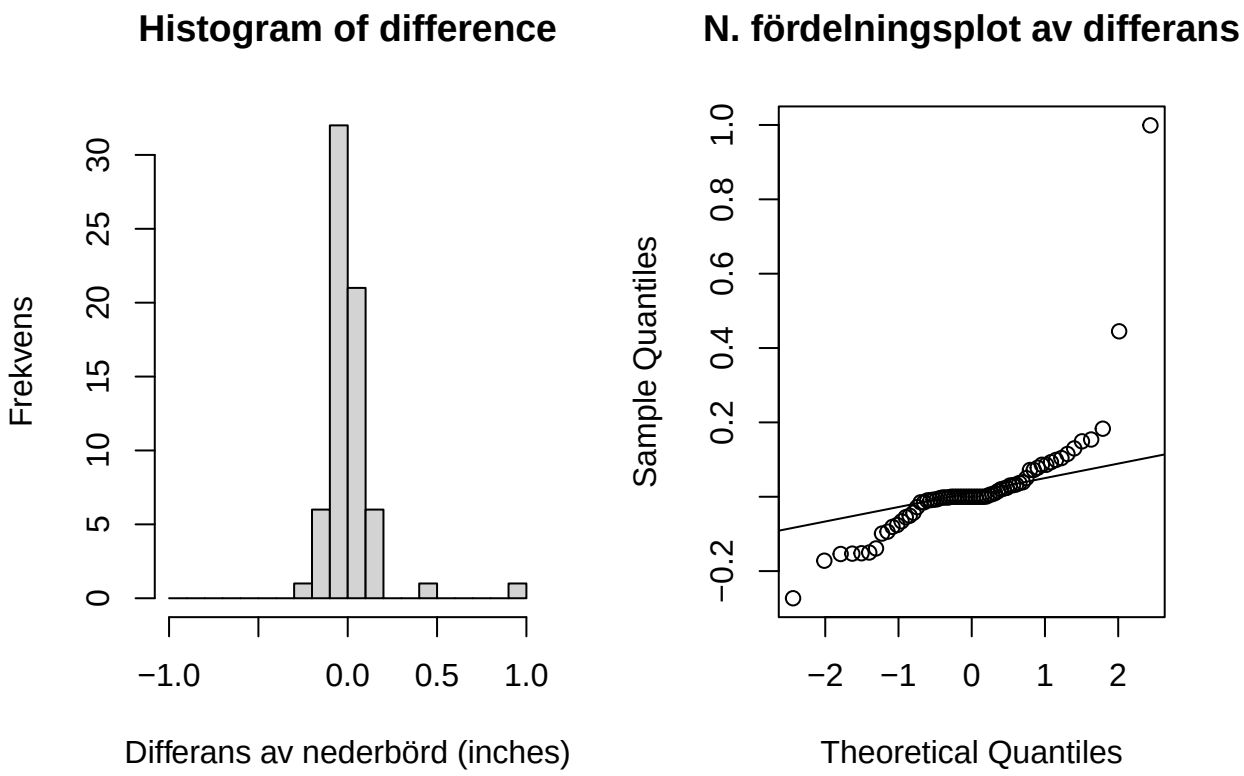


I figur 1 avläser vi rätt säkert att datan, oberoende av om molnsådd genomförs eller inte, är grovt exponentialfördelat. Detta är inte så konstigt, iom att de flesta dagarna regnar det inte i Arizona, molnströssel eller ej. Vidare ser vi att dagarna då det regnar mycket är väldigt få, och att när det väl regnar, så regnar det inte så mycket. Det råder nu att undersöka vilken fördelning differansen av nederbörd mellan dagarna i tvådagarsperioder har. Egentligen är vi bara intresserade av om fördelningen är normal eller inte, eftersom det kommer att avgöra vilket typ av test som är mest lämpligt. Detta undersöker vi med hjälp av histogram, men vi slänger även in en normalfördelningsplott för säkerhetens skull.

```
old_par <- par(mfrow = c(1,2))

hist(difference, breaks = seq(from = -1, to = 1, by = 0.1),
     xlab = "Differans av nederbörd (inches)",
     ylab = "Frekvens")

qqnorm(difference, main = "N. fördelningsplot av differans")
qqline(difference)
```



Det känns inte rimligt, givet informationen i figur 2, att differansen i nederbörd har följer en normalfördelning, eftersom majoriteten av datan är otroligt koncentrerad i mitten, och det finns i princip inga värden någon annanstans. Detta förstärks ännu mer av normalfördelningsplotten i figur “, då vi ser tydligt att våran data inte följer en rät linje. Från detta får vi en idé om vilket typ av test vi ska utföra.

Eftersom differansen mellan tvådagarsperioderna är oberoende, och fördelningen behåller en viss symmetri mellan dagarna (en differans ligger långt ut till höger men det är acceptabelt), så tror vi att vägen framåt här är att utföra ett Wilcoxon teckenrang test där vi testar om medianen av fördelningen differansen har ligger på talet 0. Alternativa hypotesen blir då att detta ej är fallet, och vi testar detta med en konfidensgrad av 95. För att utföra testet använder vi oss av Rs inbyggda funktion `wilcox.test()`.

```
wilcox.test(difference,
            alternative = c("two.sided"),
            mu = 0,
            conf.int = TRUE,
            conf.level = 0.95)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  difference
## V = 849, p-value = 0.5107
## alternative hypothesis: true location is not equal to 0
## 95 percent confidence interval:
##  -0.01654121  0.03498019
## sample estimates:
## (pseudo)median
##    0.008020371
```

Eftersom vårat $p_{obs} = 0.5107 > 0.05$, kan vi inte förkasta nollhypotesen. Det vill säga, vi kan inte med god säkerhet dra slutsatsen att molnsådd har någon effekt på nederbörd.

Uppgift 2.1 Problemformulering och teoretisk svar till teoretiska frågor.

I denna del ska vi undersöka samma fråga som i förra uppgiften, men med en annan uppsättning. I oregon gjorde man nämligen ett liknande försök i att undersöka huruvida molnsådd ökade mängden nederbörd eller inte. Men istället för att skapa parvist beroende stickprov, kollade man först och främst på dagar då *förutsättningarna* för nederbörd var uppfyllda, och sen avgjorde aman med slumpen huruvida man skulle utföra molnsådd eller inte. På så sätt hade man inte lika många dagar då det inte hände något, vilket vi i förra uppgiften såg ledde till att mycket av datan var koncentrerad kring att det inte var någon nederbörd alls (Figur 1). Anledningen till att man sen lät slumpen avgöra om molnsådd skulle genomföras eller inte, hade förmodligen att göra med att man önskade undvika systematiskt fel som hade kunnat uppstå, om t.ex. en person istället hade fått bestämma. Eftersom den personen hade kunnat ha någon bias för att experimentet skulle gå åt ett håll eller annat, och det hade kunnat ha en effekt på deras beslut att genomföra molnsådd eller inte på en given dag. Genom att låta slumpen avgöra detta undviker man det felet.

Vidare så delade man upp datan i detta experiment efter vilket typ av område det var man undersökte. Endast två av dessa är de vi kollar på, och den ena typen var stora områden som befann sig i vindriktning från molnen man strödde, och den andra var mindre områden man ansåg vara "särskilt känsliga för molnsådd"[1]. Given denna uppsättning vill vi lista ut om datan som samlades in har något att säga om effektiviteten av molnströssling.

Vi har skäl att anta att nederbörden ser olika ut beroende på vilket typ av området vi undersöker, därför kan det vara fördelaktigt att inte slå ihop datan så att vi kan betrakta den enskilt, och göra testet där variansen antas vara lika (men okänd).

[1] Enligt labbinstruktionerna på sida 5.