Statistisk Analys: Laboration 2

Ottilia Andersson & David Sermoneta

2024-12-18

Sammanfattning

Vi undersöker ett historiskt datamaterial, där meteorologer har undersökt huruvida det går att öka mängden nederbörd med mänsklig hjälp. I uppgift 1 undersöker vi data från Arizona, där vi har ett parvist beroende stickprov. Vi kommer fram till att vi inte med säkerhet kan säga att den prövade metoden ger ökad nederbörd. I uppgift 2 undersöker vi data från Oregon: här har vi två uppsättningar med data, från två olika geografiska områdestyper. Undersökningsmetoden är lite annorlunda och vi har oberoende stickprov. Men inte heller här kan vi dra slutsatsen att meteorologernas metod ökar mängden nederbörd.

Uppgift 1: Molnsådd i Arizona

1.1 Problemformulering

Under somrarna 1957–1960 genomfördes ett antal försök i Arizonas bergstrakter för att se om molnsådd kunde öka mängden nederbörd i torra ökenområden.[1] För att kunna besvara frågan utfördes ett experiment, där sommaren delades in i tvådagarsperioder. För varje tvådagarsperiod lät man slumpen avgöra vilken av de två dagarna som molnsådd skulle genomföras på. Nederbörden mättes och samlades i ett dataset som vi här ska undersöka för att kunna besvara meteorologernas ursprungliga fråga.

Vitsen med att dela in hela sommaren i tvådagarsperioder i stället för att från början bestämma på vilka dagar molnsådd ska genomgöras, exempelvis *helt* med hjälp av slump, är att vi får viss kontroll över spridningen över vår testperiod. Hade vi helt och hållet låtit slumpen avgöra vilka dagar vi skulle genomföra molnsådd på och inte, skulle vi ha kunnat råka få ett utfall där molnsådd till exempel endast genomfördes under dagar helt utan möjlighet till nederbörd. Om vi antar att molntäcken (dvs *möjligheten* till nederbörd) dröjer sig kvar mer än en dag, har genom införandet av våra tvådagarsperioder "garanterat" molnsådd någon av dessa dagar.

Genom att dela in hela sommaren i tvådagarsperioder garanteras också att det aldrig kommer vara mer än två dagars molnsådd i rad. Detta är ett sätt att skapa *oberoende* mellan observationerna.

Anledningen till att vi slumpmässigt väljer vilken av dagarna i respektive tvådagarsperiod vi väljer att genomföra molnsådd på, är att vi inte vet huruvida molnsådd påverkar nederbörden *dagen efter*. Tvärtom är det rimligt att anta att det *det föreligger* ett beroende mellan dag med molnsådd och dag utan molnsådd, när dessa följer på varandra. Varje tvådagarsperiod kommer således att utgöra en datapunkt bestående av *två* värden. Om vi hade valt att exempelvis alltid utföra molnsådd den första av de två dagarna, så skulle det finnas systematisk skevhet i våra data, vilket gör det svårt för oss att uttala oss om orsaken till en eventuell skillnad i stickproven. Genom att låta slumpen välja vilken av de två dagarna vi genomför molnsådd på, "neutraliserar" vi detta eftersom beroendet kommer att gå "åt båda håll".

Vi vill nu veta om molnsådd ökar mängden nederbörd. Utifrån vår testuppsättning får vi två *parvist beroende* stickprov: Inom varje tvådagarsperiod råder *beroende*, medan mellan det mellan tvådagarsperioderna råder *oberoende*. Det vi vid en sådan situation bör välja att göra är ett test av skillnader i väntevärden för två

[1] Mening kopierad ur laborationsbeskrivningen.

parvist beroende stickprov. Testet kommer hjälpa oss att besvara frågan huruvida molnsådd ökar mängden nederbörd, med andra ord: är det är någon skillnad i väntevärde (medelvärde) mellan dagar med molnsådd (det ena stickprovet), och dagar utan molnsådd (det andra stickprovet).

Testet går ut på att vi utifrån våra parvist beroende (och lika stora) stickprov $x_1, \cdots, x_n$ och $y_1, \cdots, y_n$, där x_i och y_i är beroende för alla $i = 1, 2, \cdots, n$, bildar ett *nytt* stickprov $d_1, \cdots, d_n$ genom att betrakta differensen mellan de båda stickproven, så att $d_i = x_i - y_i$. Vi får då ett (1) oberoende stickprov, vilket vi behandlar enligt kända metoder för enskilda stickprov. Mer om detta nedan.

1.2 Hämtning och undersökning av data

Eftersom data är parvist beroende skulle det vara trevligt att ha ett nytt stickprov som bara består av differensen mellan paren av dagar. Dessa nya datapunkter som består av differensen kommer då att vara sinsemellan obereonde. Vad vi är intresserade av är ju om det finns en skillnad i medelvärde mellan de två stickproven. Vår nollhypotes är således $H_0 : \mu_0 = \mu_{nonseed} - \mu_{seed} = 0$ (det är ingen skillnad i nederbörd mellan dagar med och utan molnsådd), och vår alternativa hypotes, $H_1 : \mu_0 \neq 0$ (det *är* skillnad i nederbörd mellan dagar med och utan molnsådd).

Vi är nu redo att avläsa data vi har till hand, och vi gör detta enligt instruktionerna. Vi passar även på att skapa differans-vektorn som introducerades ovan, 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
```

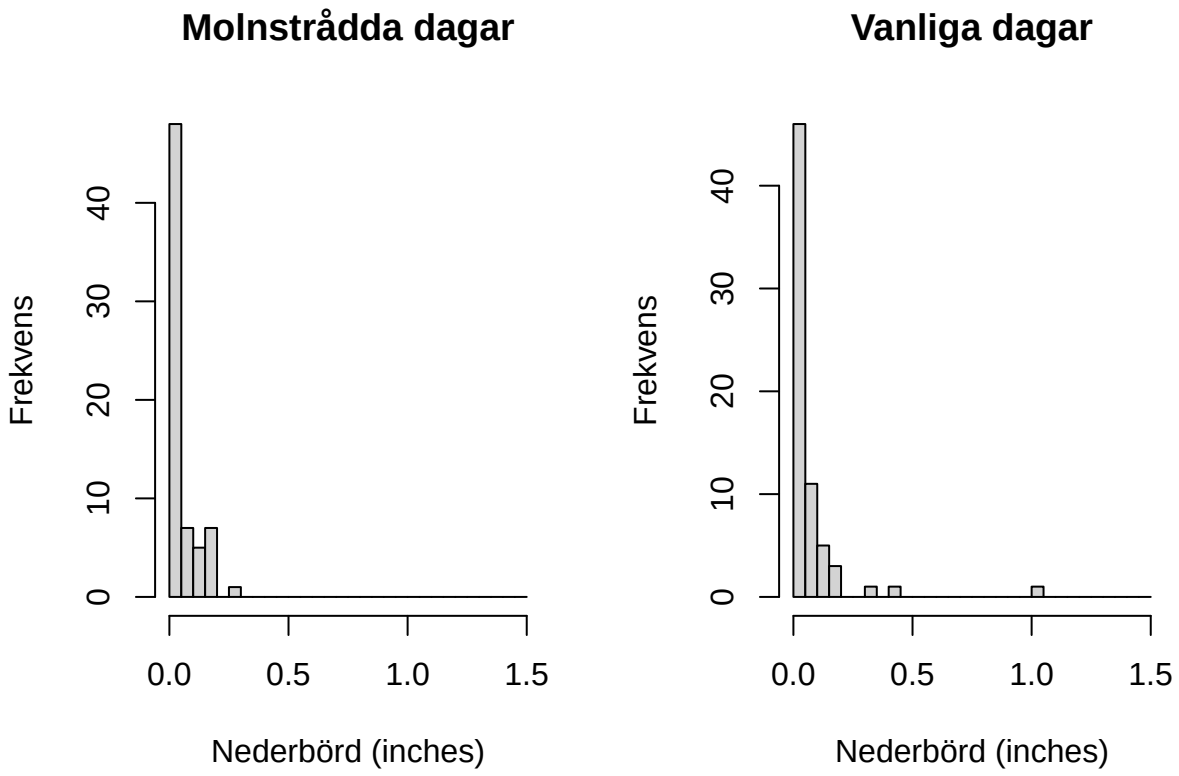
Vi börjar med att undersöka om båda stickproven följer samma fördelning, och i så fall, vilken. Vi anar att histogramen kommer duga för detta, så vi plottar dem sida vid sida (tillsammans hjälps de åt).

```
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 data, oberoende av om molnsådd genomförs eller inte, är exponentialfördelad. Detta är inte så konstigt, i och med 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 differensen av nederbörd mellan dagarna i tvådagarsperioderna 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 säkerhets skull.

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

hist(difference, breaks = seq(from = -1, to = 1, by = 0.1),
     main = "Histogram av differensen",
     xlab = "Differens av nederbörd (inches)",
     ylab = "Frekvens")
```



Figur 1: Histogram av datan i Arizona.csv, som svarar mot frekvensen i nederbörd (mätt i inches), för dagar som där molnsådd har genomförts (vänster), och inte (höger).

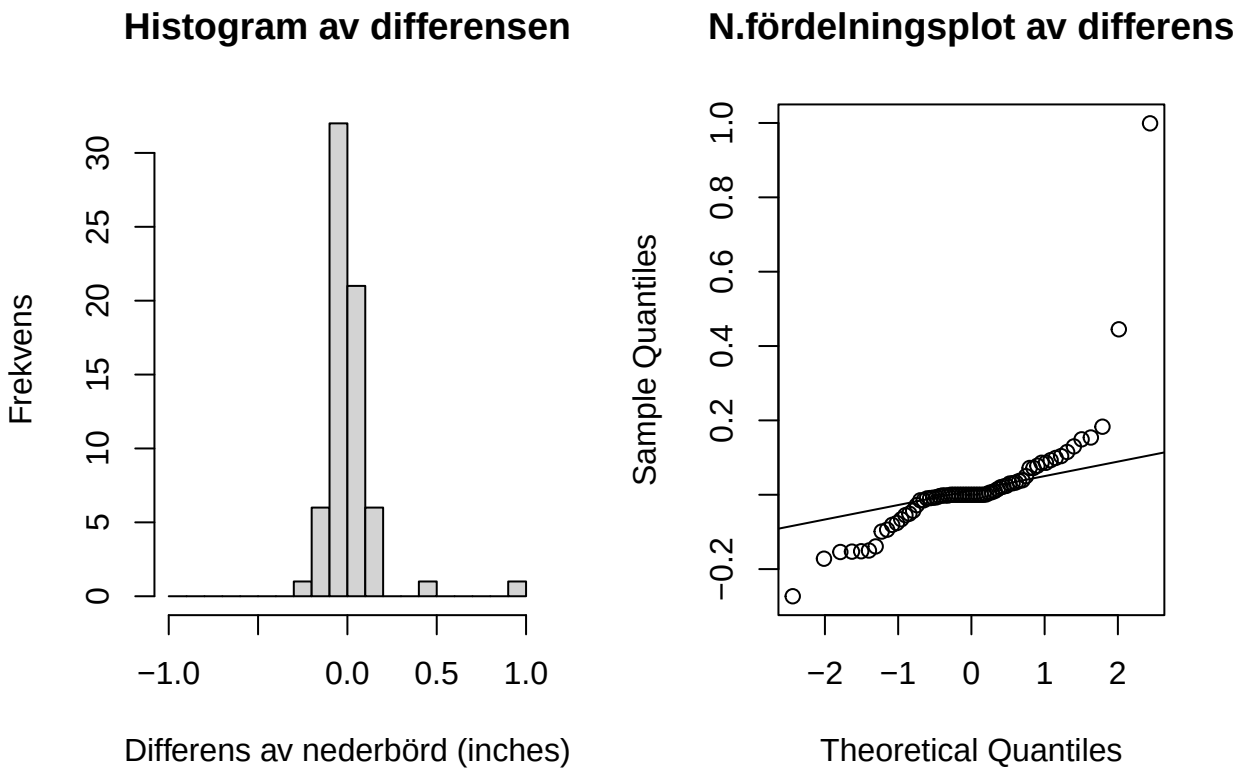
```
qqnorm(difference, main = "N.fördelningsplot av differens")
qqline(difference)
```

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

Eftersom differensen mellan tvådagarsperioderna är oberoende, och fördelningen behåller en viss symmetri mellan dagarna (en differens ligger långt ut till höger men det är acceptabelt), så tror vi att vägen framåt här är att utföra ett Wilcoxon teckenrangtest där vi testar om medianen av fördelningen differansen ligger på talet 0. Alternativa hypotesen blir då att detta ej är fallet, och vi testar detta med en konfidensgrad av 95. Vi väljer en tvåsidig mothypotes för att vara mer konservativa, det vill säga minska risken för att felaktigt råka förkasta nollhypotesen. För att utföra testet använder vi oss av Rs inbyggda funktion `wilcox.test()`.

```
wilcox.test(difference,
            alternative = c("two.sided"), # Mer säkerhet är bättre än mindre.
                                          # Man kan tänka sig att molnstössel är
                                          # kostsamt att implementera.
            mu = 0,
            conf.int = TRUE,
            conf.level = 0.95)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
```



Figur 2: Histogram och normalfördelningsplott av differansen i nederbörd av de parvist kopplade dagarna i experimentet som utfördes.

```
## 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 vårt $p_{obs} = 0.5107 > 0.05$, 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: Molnsådd i Oregon

2.1 Problemformulering

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 man 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 en stor del av data var koncentrerad kring att det inte var någon nederbörd alls (se Figur 1).

Anledningen till att man 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 i stället

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 delade man upp data i detta experiment efter vilket typ av område det var man undersökte. Vi kommer att undersöka två av dessa områdestyper: den ena 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"[2].

Givet denna uppsättning vill vi ta reda på om data som samlades in har något att säga om effektiviteten av molnströssling. Man kan säkert anta att mängd och varians i nederbörd ser olika ut beroende på vilket typ av området vi undersöker, därför kan det vara fördelaktigt att inte slå ihop data, utan att vi kan betraktar områdena var för sig, där variansen kan antas vara mer lik (men ännu okänd).

2.2 Hämtning och undersökning av data

Vi läser in data för Oregon och skapar tre vektorer utifrån kolonnerna i inläst dataframe:

```
oregon <- read.csv("oregon.csv", header = FALSE) # Läser in fil

trial <- oregon$V1 # 1 anger att molnsådd inte genomfördes, 2 anger att molnsådd genomfördes
typ1 <- oregon$V2 # Nederbörd i områden av typ 1 (stora områden i vindriktningen)
typ2 <- oregon$V3 # Nederbörd i områden av typ 2 (områden särskilt känsliga för molnsådd)
```

Vi är intresserade av de två olika typerna av områden, och vill filtrera på om molnsådd genomfördes eller ej:

```
nonseed_typ1 <- typ1[trial == 1] # Område av typ 1 (stora områden i vindriktningen), dagar då molnsådd
seed_typ1 <- typ1[trial == 2] # Område av typ 1 (stora områden i vindriktningen), dagar då molnsådd gen

nonseed_typ2 <- typ2[trial == 1] # Område av typ 2 (särskilt känsliga områden), dagar då molnsådd INTE
seed_typ2 <- typ2[trial == 2] # Område av typ 2 (särskilt känsliga områden), dagar då molnsådd genomför
```

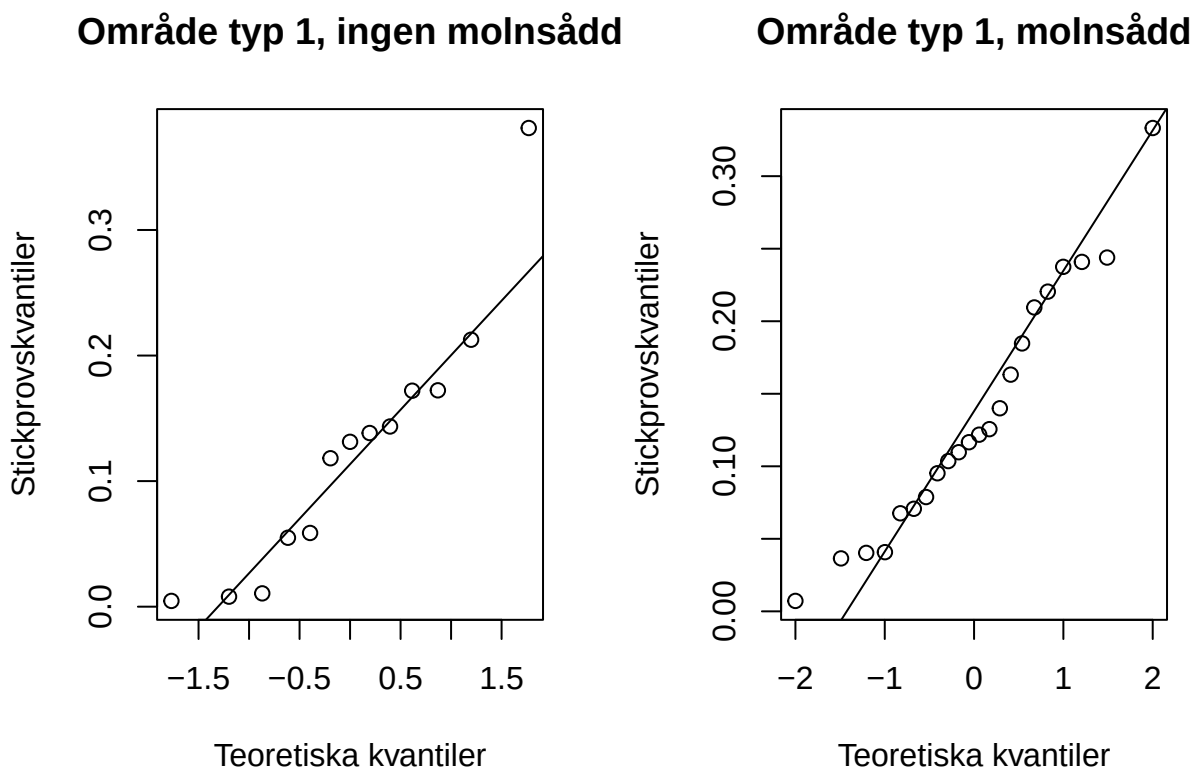
Vi har nu fyra stickprov, vilka uppdelade på område utgör två utfall av en situation med *två oberoende stickprov*. Variansen är okänd, men eftersom vi jämför samma typ av område, har vi inga skäl att tro att variansen mellan stickproven skiljer sig åt nämnvärt, så vi kan anta att $\sigma_{seed} \approx \sigma_{nonseed}$ för respektive område.

Vi börjar med att illustrera våra data i några grafer, för att få en uppfattning om den:

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

qqnorm(nonseed_typ1,
       main = "Område typ 1, ingen molnsådd",
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler")
qqline(nonseed_typ1)

qqnorm(seed_typ1,
       main = "Område typ 1, molnsådd",
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler")
qqline(seed_typ1)
```

Det vi ser i figur 3 är att data från område typ 1 följer en rät linje i normalfördelningsplotten. Mot bakgrund av våra resultat i föregående laboration känner vi oss nöjda med dessa resultat, och tillräckligt säkra för att gå vidare med ett hypotestest för *skillnader i väntevärde mellan två oberoende normalfördelade stickprov*. Se vidare i del 2.2.1.

[2]Enligt labbinstruktionerna på sida 5.

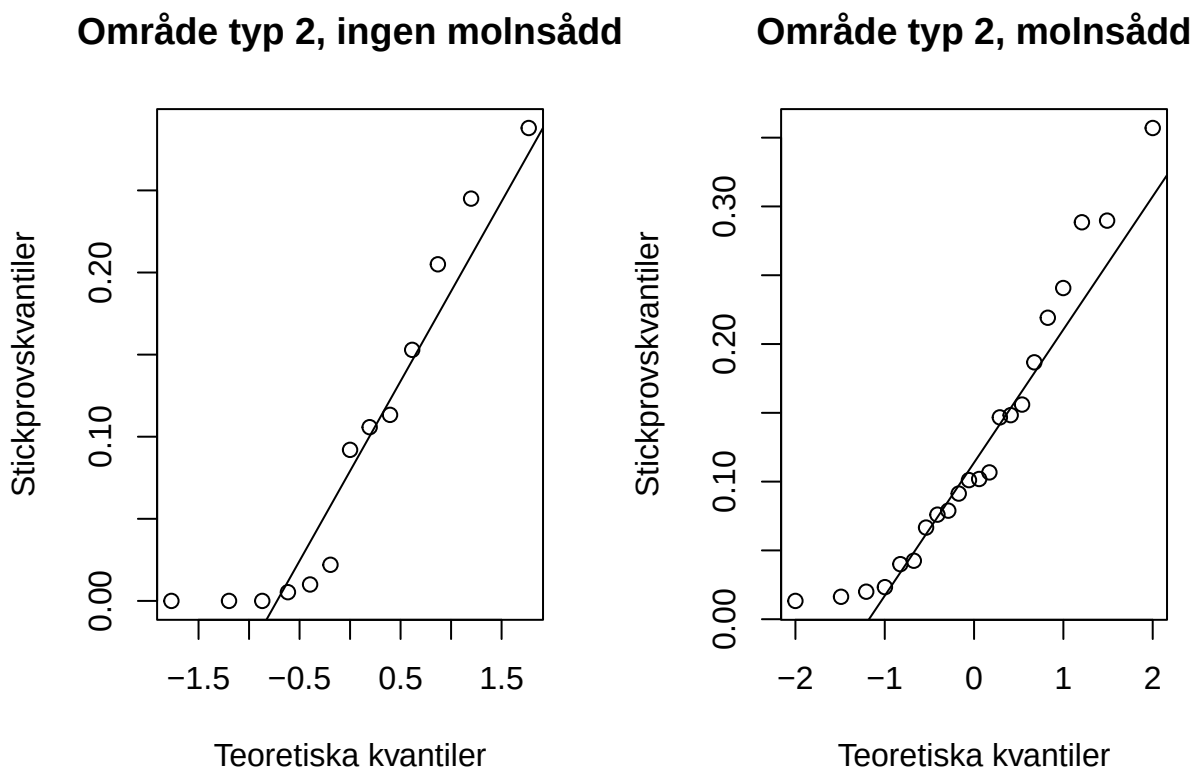


Figur 3: Data från område av typ 1 (stora områden i vindriktningen) presenterat i två normalfördelningsplottar. Ingen molnsådd i figuren till vänster, molnsådd i figuren till höger.

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

qqnorm(nonseed_typ2,
       main = "Område typ 2, ingen molnsådd",
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler")
qqline(nonseed_typ2)

qqnorm(seed_typ2,
       main = "Område typ 2, molnsådd",
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler")
qqline(seed_typ2)
```

För data från område av typ 2 anar vi en S-formad kurva, som vi kan se i figur 4, vilket får oss att vilja undersöka data i histogram, för att få en bättre uppfattning om fördelnigen och därmed kunna fatta beslut om hypotestest.

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

hist(nonseed_typ2, breaks = seq(from=0, to=0.4, by=0.03),
     main = "Område typ 2, ingen molnsådd",
     xlab = "Nederbörd",
     ylab = "Frekvens",
     prob = TRUE)
x <-  seq(from = 0, to = 0.5, length.out = 100)
```



Figur 4: Data från område av typ 2 (mindre områden särskilt känsliga för molnsådd) presenterat i två normalfördelningsplottar. Ingen molnsådd i figuren till vänster, molnsådd i figuren till höger.

```
lines(x, dnorm(x,0.2,0.1))

hist(seed_typ2, breaks = seq(from=0, to=0.4, by=0.03),
     main = "Område typ 2, molnsådd",
     xlab = "Nederbörd",
     ylab = "Frekvens",
     prob = TRUE)
x <-  seq(from = 0, to = 0.5, length.out = 100)
lines(x, dnorm(x,0.2,0.1))
```

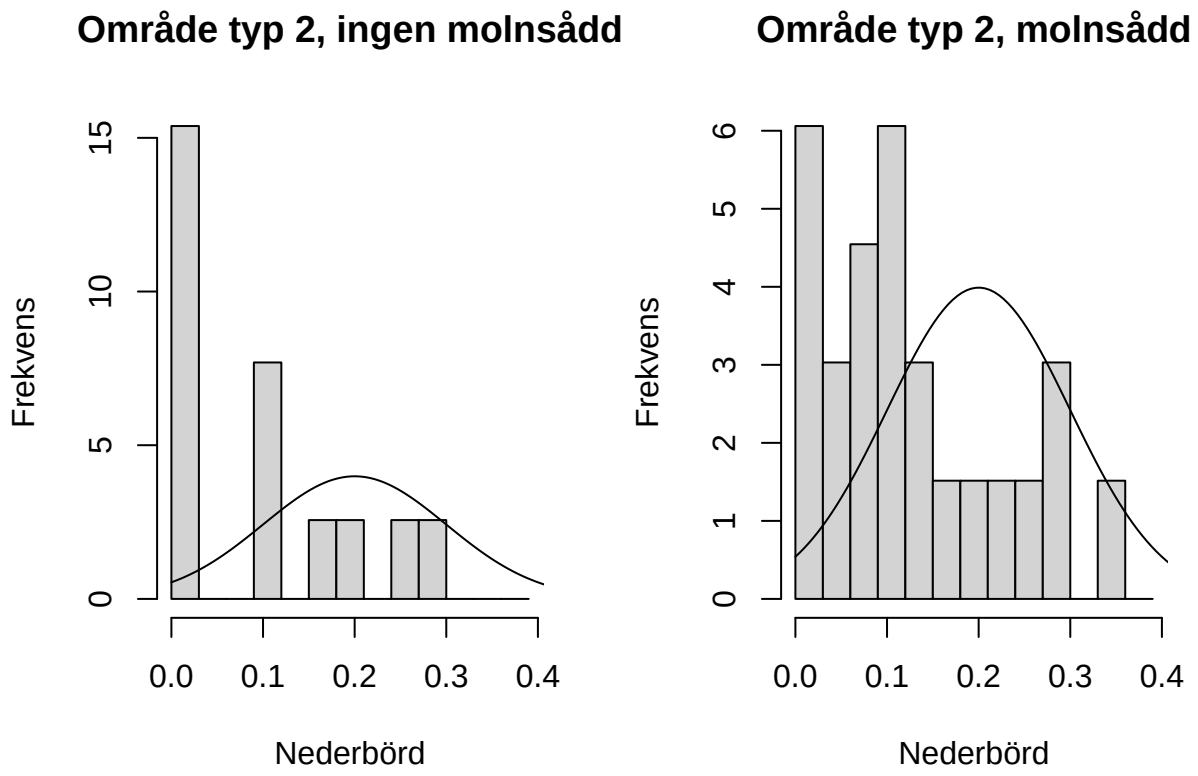
Mycket riktigt ser vi i figur 5 en stor avvikelse mellan den teoretiska normalfördelningen (linjen) och vår data (staplarna). Vi känner oss inte trygga med att göra antagande om att data är normalfördelade, och kommer för område 2 därför att utföra ett icke-parametriskt hypotestest, vilket kan utföras utan antagande om fördelnigen.

2.2.1 Test för område typ 1

Vår nollhypotes är att det inte föreligger någon skillnad i väntevärde mellan stickproven, alltså $H_0 : \mu_{seed} - \mu_{nonseed} = 0$. Vi formulerar en tvåsidig mothypotes (för att vara mer konservativa), $H_1 : \mu_{seed} - \mu_{nonseed} \neq 0$. Vi utför ett t-test med hjälp av R:s inbyggda funktion:

```
t.test(nonseed_typ1, seed_typ1, alternative = "two.sided", mu = 0, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
```



Figur 5: Data från område av typ 2 (mindre områden särskilt känsliga för molnsådd) presenterat i två histogram. Ingen molnsådd i figuren till vänster, molnsådd i figuren till höger.

```
## data:  nonseed_typ1 and seed_typ1
## t = -0.36159, df = 21.309, p-value = 0.7212
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08264564  0.05814424
## sample estimates:
## mean of x mean of y
## 0.1235538 0.1358045
```

Utifrån t-testet får vi ett 95 % konfidensintervall för $\mu_{seed} - \mu_{nonseed}$: $I_\mu = (-0.083, 0.058)$. Vi konstaterar att värdet 0 ligger i konfidensintervallet och således *behåller vi* H_0 .

2.2.2 Test för område typ 2

Vi vill genomföra ett icke-parametriskt hypotestest och kommer att utföra *Wilcoxons tvåstickprovstest*, vilket fungerar bra i vår situation där vi inte vill anta något om fördelningarna och stickproven är olika stora. För detta använder vi R:s inbyggda funktion `wilcox.test`.

Vår nollhypotes är att det inte finns någon skillnad i väntevärde mellan dagar med molnsådd och dagar utan, $H_0 : \mu_{seed} - \mu_{nonseed} = 0$, och vår mothypotes formuleras tvåsidigt (för att vara mer konservativa) som att det *finns* en skillnad i väntevärde, $H_1 : \mu_{seed} - \mu_{nonseed} \neq 0$.

```
wilcox.test(nonseed_typ2, seed_typ2,
            alternative = c("two.sided"),
            mu = 0,
            conf.int = TRUE,
            conf.level = 0.95)
```

```
## Warning in wilcox.test.default(nonseed_typ2, seed_typ2, alternative =
## c("two.sided"), : cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(nonseed_typ2, seed_typ2, alternative =
## c("two.sided"), : cannot compute exact confidence intervals with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  nonseed_typ2 and seed_typ2
## W = 108, p-value = 0.2387
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.09663490  0.03735689
## sample estimates:
## difference in location
##            -0.03420231
```

Utifrån testet får vi ett 95 % konfidensintervall för $\mu_{seed} - \mu_{nonseed}$ enligt $I_\mu = (-0.096, 0.037)$. Vi konstaterar att värdet 0 ligger i konfidensintervallet och vi får ett p-värde på $0.2387 > 0.05$. Vi behåller H_0.

Kommentar: Vi får en varning av R, vilken beror på att vi har flera identiska värden. Detta gör att vi får en approximation. För skojs skull gör vi en körning med samma test som för stickproven från område av typ 1, det vill säga, vi tänker oss att stickproven ändå är *tillräckligt* normalfördelade.

```
t.test(nonseed_typ2, seed_typ2, alternative = "two.sided", mu = 0, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  nonseed_typ2 and seed_typ2
## t = -0.92228, df = 24.825, p-value = 0.3653
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10485474  0.04000719
## sample estimates:
##  mean of x  mean of y
## 0.09533077 0.12775455
```

Detta test ger oss ett 95 % konfidensintervall på $I_\mu = (-0.105, 0.040)$, alltså ett bredare konfidensintervall än med Wilcoxons tvåstickprovstest. Vårt erhållna p-värde är $0.3653 > 0.05$. Även här konstaterar vi att $0 \in I_\mu$. Vi behåller H_0.

Diskussion och slutsats

Frågan vi ville ha svar på är huruvida molnsådd Ökar nederbörden. För att svara på den frågan utförde vi ett hypotestest för skillnader i väntevärden för oberoende stickprov, och Wilcoxons tvåstickprovstest. Inget av testen gav oss p-värden som var tillräckligt små för att vi skulle förkasta nollhypotesen. Alltså är vår slutsats att vi *inte* kan styrka att molnsådd ökar mängden nederbörd.