Aflevering 1

Lucas Bagge

2021-02-09

Opgave 2.7, MSRR: Analyse af spruce datasættet

Til denne opgave skal vi bruge data som er beskrevet i afsnit 1.10. Data er blevet udarbejdet tilbage i 1990 af en biolog som var interesseret i at undersøge hvad der påvirker væksten af nåletræer.

Data indlæses som følger med funktionen `read_csv()`:

```
df <- read.csv(file = "MatStat-R/data/Spruce.csv")
```

Herhefter bruger jeg `glimse()` til at få et overblik over data.

```
df %>% glimpse()
```

```
## Rows: 72
## Columns: 9
## $ Tree        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ...
## $ Competition <chr> "NC", "NC", "NC", "NC", "NC", "NC", "NC", "NC", "NC", "...
## $ Fertilizer  <chr> "F", "F", "F", "F", "F", "F", "NF", "NF", "NF", "NF", "...
## $ Height0     <dbl> 15.0, 9.0, 12.0, 13.7, 12.0, 12.0, 16.8, 14.6, 16.0, 15...
## $ Height5     <dbl> 60.0, 45.2, 42.0, 49.5, 47.3, 56.4, 43.5, 49.2, 54.0, 4...
## $ Diameter0   <dbl> 1.984375, 1.190625, 1.785937, 1.587500, 1.587500, 1.587...
## $ Diameter5   <dbl> 7.4, 5.2, 5.7, 6.4, 6.2, 7.4, 4.9, 5.4, 7.1, 5.1, 4.1, ...
## $ Ht.change   <dbl> 45.0, 36.2, 30.0, 35.8, 35.3, 44.4, 26.7, 34.6, 38.0, 2...
## $ Di.change   <dbl> 5.415625, 4.009375, 3.914062, 4.812500, 4.612500, 5.812...
```

Vi ser at data indholder 72 rækker og 9 kolonner. Vi har højde og dimensioner fra før undersøgelsen og 5 år efter. To interssante features er `Competition` og `Fertilizer` angiver de faktor der kan påvirke nåletræets vækst. De sidste to features `Ht.change` og `Di.change` beskriver ændringen i højden og diameteren efter 5 år.

a) Numeriske deskriptorer, højdevækst

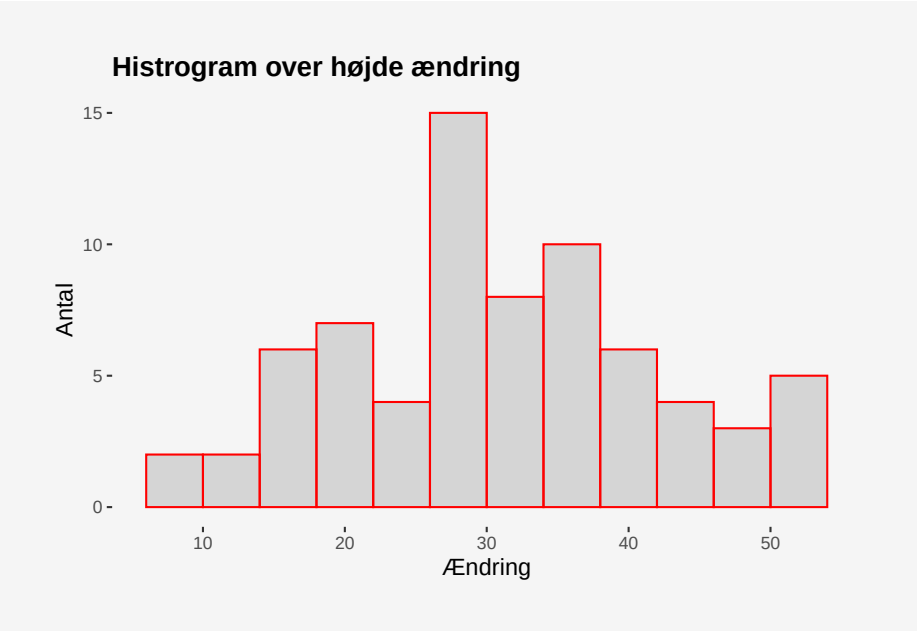
```
df %>%
  summarise(
    min = min(Ht.change),
    ned_fraktil = quantile(Ht.change, c(0.25)),
    mean_ht.change = mean(Ht.change),
    øvre_fraktil = quantile(Ht.change, c(.75)),
    max = max(Ht.change),
    sd = sd(Ht.change)
    )
```

```
##   min ned_fraktil mean_ht.change øvre_fraktil  max       sd
## 1 8.3        23.2       30.93333       38.175 51.5 11.04943
```

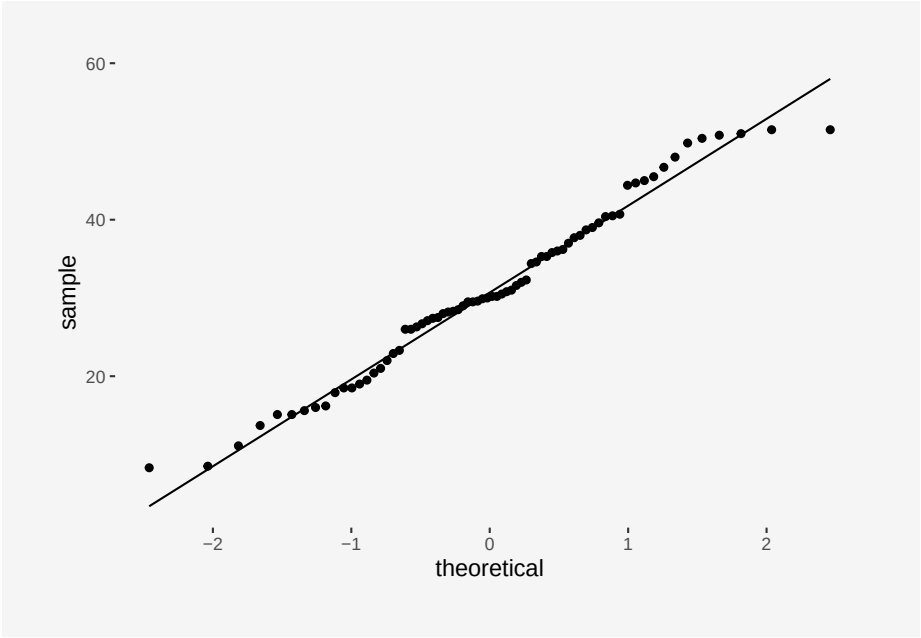
b) Histogram og normalfraktilplot, højdevækst

Jeg laver herforneden to to efterspurgte grafer for et histrogram og normalfraktilplot.

```
df %>%
  ggplot(aes(x = Ht.change)) +
  geom_histogram(binwidth = 4,
                 col = "red",
                 alpha = 0.2) +
  labs(title = "Histrogram over højde ændring",
       x = 'Ændring',
       y = 'Antal') +
my_theme()
```



```
df %>%
  ggplot(aes(sample = Ht.change)) +
  stat_qq() +
  stat_qq_line() +
  my_theme()
```

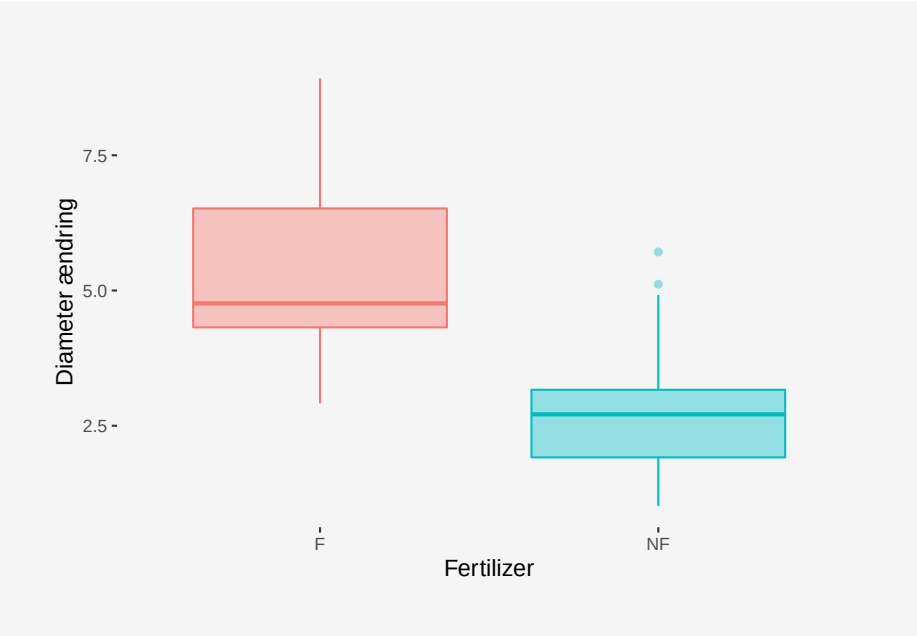


Ud fra histrogrammet og normalfraktilplot kan vi så bedømme om det er normalt fordelt. Man skal tolke normalfraktilplottet således at hvis punkter falder på den rette linje, så er data approximativ normal fordelt. Fra vorea qq plot (normalfraktilplot) så falder de nogenlunde på linjen og vi kan konkluderer at data er approximativ normalt fordelt.

c) Grafisk sammenligning af væksten i diameter, med og uden gødning

I denne opgave skal jeg lavet et `boxplot`, som giver et overblik over de fem summary værdier.

```
df %>%
  ggplot(aes(Fertilizer, Di.change, fill = Fertilizer, color = Fertilizer)) +
  geom_boxplot(alpha = 0.4) +
  labs(y = "Diameter ændring", color = NULL, fill = NULL) +
  my_theme()
```



Ud fra boxplottet kan vi se der tydelig er en forskel i diameteren ud fra om nåletræerne har været `Fertilized` eller ej.

Det ser ud til at de nåle træer i et Fertilized miljø ser en større dimeter ændring. Desuden ser det ud til at nåletræerne i (F) er mere højre skæve end dem som ikke får gødning.

d) Numerisk sammenligning af væksten i diameter, med og uden gødning

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
tapply(df$Di.change, df$Fertilizer, quantile)
```

```
## $F
```

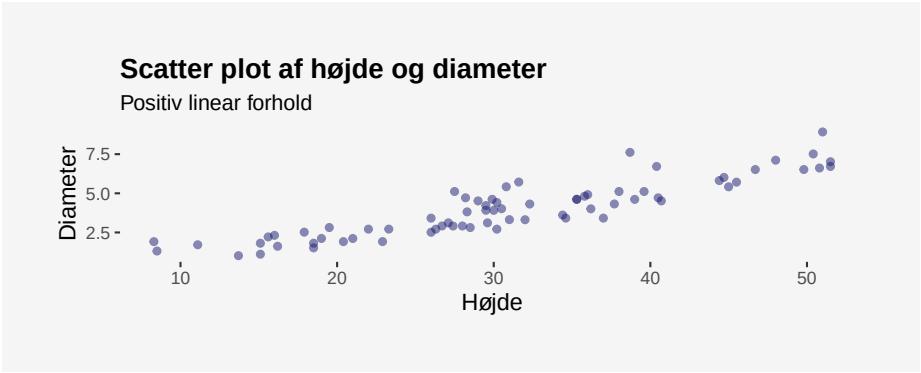
```
##       0%      25%      50%      75%     100%
## 2.912500 4.317969 4.762500 6.517578 8.918750
##
## $NF
##       0%      25%      50%      75%     100%
## 1.018750 1.915234 2.711719 3.164844 5.712500
```

Foroven har jeg brugt funktionen `tapply` til at udregne de numeriske værdier for at sammenligne diameter ændring for de to niveauer.

Uden at udregne de præcise værdier i box plottet i opgave c, hvor jeg jeg vurderer det ud fra en visual inspektion, så ser resultater ud til at være ens. Dermed kommer jeg frem til den samme konklusion.

e) Sammenhæng mellem væksten i højde og væksten i diameter

```
df %>%
  ggplot(aes(Ht.change, Di.change)) +
  geom_point(alpha = 0.5, color = "midnightblue") +
  coord_fixed() +
  labs(y = "Diameter", x = "Højde",
       title = "Scatter plot af højde og diameter",
       subtitle = "Positiv linear forhold") +
  my_theme()
```



Ud fra scatter plottet så ser vi et positiv lineær forhold mellem højde og diameter. Det giver god mening da begge mål er et udtryk for størrelsen af nåle træet og jeg tænker at et højere træ også må have en større diameter.

Det er ofte at man plotter den primær variable på y aksen. I dette tilfælde ved jeg ikke hvilken der har primær interesse, så jeg plotte diamerter på y aksen.