Aflevering 2

Lucas Bagge

2021-02-16

Opgave 3.15

a) Undersøg data og forklar om det er et matched par.

Indsæt `Groceries` data.

```
data <- read.csv("MatStat-R/data/Groceries.csv")
```

Så vil vi se på selve data:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
glimpse(data)
```

```
## Rows: 30
## Columns: 4
## $ Product <chr> "Kellogg NutriGrain Bars", "Quaker Oats Life Cereal  Origin...
## $ Size    <chr> "8 bars", "18oz", "11.50z", "18oz", "14.3oz ", "13oz", "10o...
## $ Target  <dbl> 2.50, 3.19, 3.19, 2.82, 2.99, 2.64, 3.99, 4.79, 1.49, 3.49,...
## $ Walmart <dbl> 2.78, 6.01, 2.98, 2.68, 2.98, 1.98, 2.50, 4.79, 1.28, 2.98,...
```

Jeg lægger mærke til at data indholder 30 rækker og 4 kolonner. De 4 variabler er:

- `Product`: De forskellig morgenmads produkter
- `Size`: Størrelsen på produktet.
- `Target` Er butikskæden Targets pris for det given produkt.
- `Walmart` Er butikskæden Walmarts pris for det given produkt.

Er dette data således et matched par?

Med udtrykket matched par, så betyder det at forskellige samples har nøjagtig de samme karakteristika, på nær en, som er hvad man gerne vil undersøge.

I vores tilfælde har vi det samme produkt, den samme størrelse, men forskellen ligger i prisen. Derfor er dette data et eksempel på et matched par.

b) Beregn opsummerings statistik for priser for hver butik.

Jeg vil her kun beregne `mean`. Det gør jeg af den simple årsag at de øvelser og eksempler der bliver gennemgået i bogen kun laver den beregning og den kan betragtes som *statistikken over alle statistikker*.

```
library(glue)
```

```
##
## Attaching package: 'glue'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
glue("
     Gennemsnitprisen for Target er {round(mean(data$Target),2)},
     men den for Wallmart er {round(mean(data$Walmart),2)}.

     ")
```

```
## Gennemsnitprisen for Target er 2.76,
## men den for Wallmart er 2.71.
```

Ud fra gennemsnittet ser vi ikke den store forskel i priserne.

c) Lav en permutation test of se om der er forskel i mean priser.

Her jeg gerne se på følgende hypotese

$$H_0 : Pris_{target} = Pris_{walmart}$$

$$H_1 : Pris_{target} \neq Pris_{walmart}$$

Jeg benytter mig af at lave et for løkke magen til det vi har set i bogen og beregne p værdien.

```
mean_target <- mean(data$Target)
mean_walmart <- mean(data$Walmart)

ObsMeanDiff <-   mean_target - mean_walmart
names(ObsMeanDiff) <- NULL

target <- subset(data, select = Target, drop = TRUE)
walmart <- subset(data, select = Walmart, drop = TRUE)
DA <- c(target, walmart)
m <- length(target)
n <- length(walmart)
N <- 10^4 - 1
Diff <- numeric(N)
for (i in 1:N) {
    index <- sample(m + n, m, replace = FALSE)
    Diff[i] <- mean(DA[index]) - mean(DA[-index])
}
pvalue <- ((sum(Diff <= ObsMeanDiff) + 1)/N + 1) * 2
format(pvalue,digits = 3)
```

```
## [1] "3.12"
```

Vi ser at pværdien er 3.1 og vi kan ikke forkaste H0. Det betyder at Så der er ingen bevis for at Target priser er højere end Walmart.

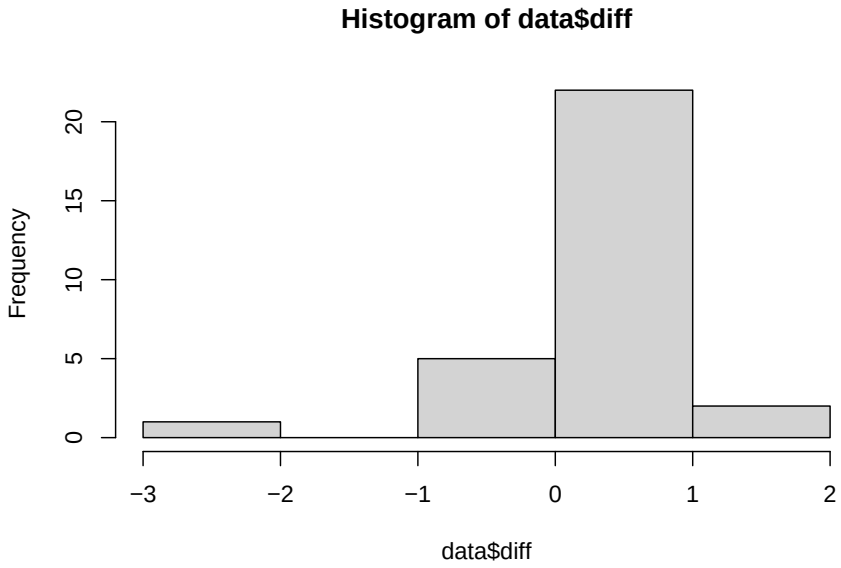
d) Lav et histogram over forskellen i priser og se på Quaker Oats life.

Først laver jeg en ny kolonne `diff`, som beregner pris forskellen mellem walmart og target.

```
data['diff'] <- data$Target - data$Walmart
```

Herefter laver jeg et histogram.

```
hist(data$diff)
```



Ud for histogrammet kan vi vi se at vi har nærmest en normal fordeling, men vi har yderpunktet mellem -2 og -3 som driller ens øjne.

I opgave teksten giver de et hint til at det er produktet **Quaker Oats life** som driller, så lav os hurtig kigge på den:

```
knitr::kable(head(arrange(data, diff), 5))
```

Product	Size	Target	Walmart	diff
Quaker Oats Life Cereal Original	18oz	3.19	6.01	-2.82
Campbell's Chicken Noodle Soup	10.75oz	0.99	1.58	-0.59
Dove Promises Milk Chocolate	8.87oz	3.19	3.50	-0.31
Kellogg NutriGrain Bars	8 bars	2.50	2.78	-0.28
Cheez-it Original Baked	21oz	4.79	4.79	0.00

Vi ser tydeligt at der er en stor forskel på prisen for Quaker Oats i de to kæder og den er meget højere i Walmart.

e) Lav den samme test, men uden den observation

Jeg vil lave samme test, men uden den observatoin for at se om jeg rokker på min konklusion.

```
data <- filter(data, data$Product != "Quaker Oats Life Cereal  Original ")
mean_target <- mean(data$Target)
mean_walmart <- mean(data$Walmart)
```

```r
ObsMeanDiff <-   mean_target - mean_walmart
names(ObsMeanDiff) <- NULL

target <- subset(data, select = Target, drop = TRUE)
walmart <- subset(data, select = Walmart, drop = TRUE)
DA <- c(target, walmart)
m <- length(target)
n <- length(walmart)
N <- 10^4 - 1
Diff <- numeric(N)
for (i in 1:N) {
    index <- sample(m + n, m, replace = FALSE)
    Diff[i] <- mean(DA[index]) - mean(DA[-index])
}
pvalue <- ((sum(Diff <= ObsMeanDiff) + 1)/N + 1) * 2
format(pvalue,digits = 3)
```

```
## [1] "3.29"
```

I den tidligere opgave fik vi en p-værdi på 3.1, men forventligt nok ser vi p-væriden stiger efter vi får fjernet den meget høje observation.

Dog ændre sig ikke på min konklusion og der er ikke den store prisforskel mellem de to kæder.