Opdrachten Statistiek 3 BIN (22 juni 2020)

6/22/2020

Typ je naam hier: Emile Apol

Typ je studentnummer hier:

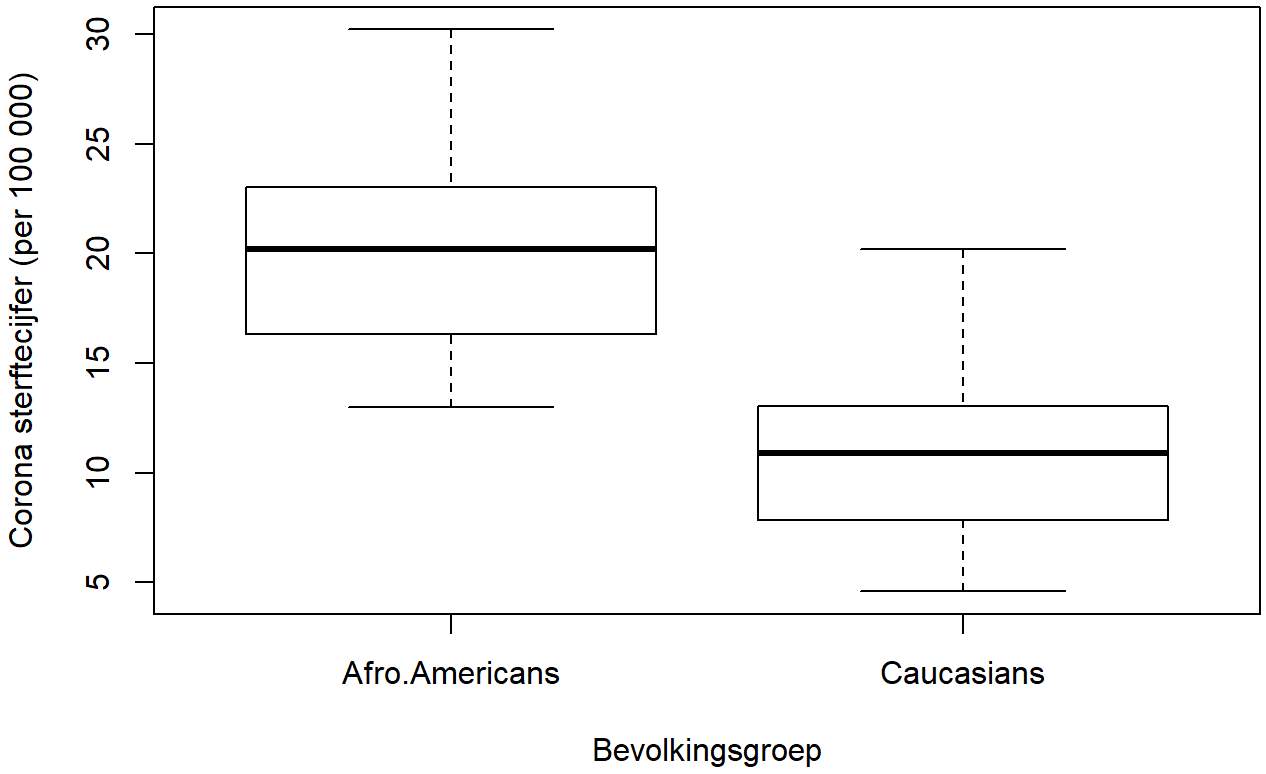
```
myPath <- "H:/Hanze/ILST/vakken/kwartaal 07/Statistiek 3 BIN/tentamens/Emile Apol/2020_06_22 opdrachten/_datasets/
_data/"
```

Opgave 1 - Sterftecijfer Covid-19 voor twee bevolkingsgroepen in de VS

Het SARS-CoV-2 virus woedt nog steeds rond. Onder andere in de VS maakt het veel slachtoffers. Als snel kwam in het nieuws dat Afro-Americans onevenredig meer werden getroffen door Corona-19 dan blanken (Caucasians). Voor een aantal staten in de VS is het corona sterftecijfer per 100 000 inwoners voor zowel Afro-Americans als Caucasians bepaald. De gegevens staan in de file Opgave_1.txt. Dit is een tab-separated ascii file met header.

a. Lees de data in, en maak in R een goede boxplot van het corona sterftecijfer als functie van bevolkingsgroep. Let op de as-titels en eenheden!

```
# setwd(myPath)
# dir()
myData <- read.table(paste0(myPath,"Opgave_1_1.txt"), header = T, sep = "\t")
myData
boxplot(myData, xlab = "Bevolkingsgroep", ylab = "Corona sterftecijfer (per 100 000)")
```




```
##                Afro.Americans Caucasians
## Alabama                  24.6       10.9
## Ohio                     30.2       20.2
## Nevada                   21.4       13.9
## Florida                  17.5       12.2
## South.Carolina           20.2        7.3
## North.Carolina           15.1        8.4
## Tennessee                13.0        4.6
```

b. Leg uit welke statistische toets geschikt is om te onderzoeken of er een verschil is in corona sterftecijfer tussen Afro-Americans en Caucasians.

y = continu, er zijn 2 gemiddelden, gepaarde metingen, dus een gepaarde t-toets.

c. Is er een verschil in corona sterftecijfer tussen Afro-Americans en Caucasians? Toets met $\alpha = 0.05$.

```
res <- t.test(myData$Afro.Americans, myData$Caucasians, paired = T)
p.value <- res$p.value

cat("p-waarde = ",p.value)
```

```
## p-waarde =  0.0002448007
```

Resultaat: p-value = 0.0002 < 0.05, dus H1: Een significant verschil.

d. Wat is de effectsterkte? Is het verschil in C-waarde tussen koeien en schapen een klein, matig of groot verschil?

```
y.1 <- mean(myData$Afro.Americans)
y.2 <- mean(myData$Caucasians)
s2.1 <- var(myData$Afro.Americans)
s2.2 <- var(myData$Caucasians)
d.av <- abs(y.1 - y.2)/sqrt((s2.1 + s2.2)/2)

cat("Effectsterkte: d.av = ",d.av)
```

```
## Effectsterkte: d.av =  1.678187
```

Resultaat: d.av = 1.67, dus een heel groot verschil!

Opgave 2 - Ziekenhuisopnames Covid-19 patiënten in Nederland

Ook Nederland is getroffen door de Covid-19 uitbraak. Een deel van de geinfecteerde personen moet worden opgenomen voor behandeling in het ziekenhuis. Gelukkig valt het aantal opnames in het noorden van ons land mee, maar er wonen hier ook minder mensen. In de file Opgave_2.txt staat per regio (Noord, Zuid, Oost en West) het aantal corona ziekenhuisopnames per 100 000 inwoners. Dit is een tab-separated ascii file met header, in “wide” format, met voor elke provincie een waarde.

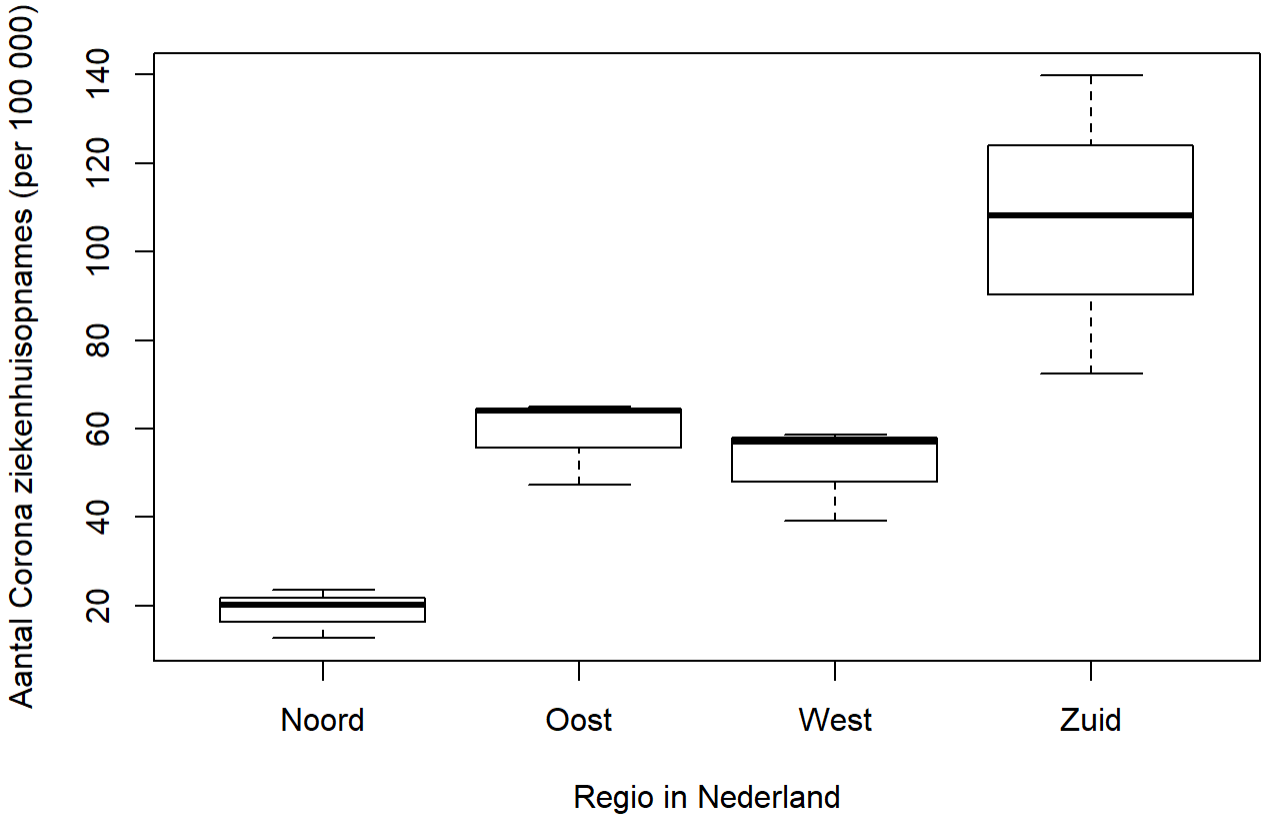
a. Lees de data in R in, en zet het dataframe om naar een dataframe in “long” format.

```
# setwd(myPath)
# dir()
myData <- read.table(paste0(myPath, "Opgave_2_1.txt"), header = T, sep = "\t")
head(myData)

LongData <- stack(myData)
colnames(LongData) <- c("Opnames", "Regio")
head(LongData)

levels(LongData$Regio)
levels(LongData$Regio) <- c("Noord", "Oost", "West", "Zuid")
head(LongData)
```

```
##       N     O     W      Z
## 1 12.67 47.39 39.16 139.77
## 2 20.07 64.10 58.68 108.10
## 3 23.57 64.97 57.02  72.39
##   Opnames Regio
## 1   12.67     N
## 2   20.07     N
## 3   23.57     N
## 4   47.39     O
## 5   64.10     O
## 6   64.97     O
## [1] "N" "O" "W" "Z"
##   Opnames Regio
## 1   12.67 Noord
## 2   20.07 Noord
## 3   23.57 Noord
## 4   47.39  Oost
## 5   64.10  Oost
## 6   64.97  Oost
```

b. Maak een goede boxplot van opnames per 100 000 als functie van regio. Let op de as-titels en eenheden!

```
boxplot(Opnames ~ Regio, data = LongData, xlab = "Regio in Nederland", ylab = "Aantal Corona ziekenhuisopnames (pe
r 100 000)")
```



c. Welke statistische analyse is geschikt om te onderzoeken of het aantal corona ziekenhuisopnames per 100 000 inwoners verschilt per regio? Leg kort uit.

Er zijn 5 gemiddelden, dus een 1-way ANOVA.

d. Is er een verschil in het aantal corona ziekenhuisopnames per 100 000 inwoners tussen de regio's? Toets met $\alpha = 0.05$.

```
( res <- summary(aov(Opnames ~ Regio, data = LongData)) )

p.value <- res[[1]]$Pr[1]
cat("\np-waarde = ",p.value,"\n")
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Regio        3  11860    3953   11.44 0.0029 **
## Residuals    8   2765     346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## p-waarde =  0.002897108
```

p-waarde = 0.0029 << 0.05, dus de regio's in Nederland verschillen significant in ziekenhuisopnames (per 100 000 inwoners).

e. Hoe groot is de effectsterkte? Heeft de regio een klein, matig of sterk effect op het aantal corona ziekenhuisopnames per 100 000 inwoners?

```
eta2 <- res[[1]]$Sum[1]/sum(res[[1]]$Sum)
cat("Effectsterkte: eta2 = ",eta2,"\n")
```

```
## Effectsterkte: eta2 =  0.8109238
```

Eta2 = 0.81 >> 0.14, dus een heel sterk effect!

f. Voer een geschikte post-hoc test uit om te zien welke regio's hetzelfde aantal opnames per 100 000 inwoners geven. Toets met $\alpha = 0.05$.

```
source("Homogeneous_subsets_v1.r")
```

```
## Sourced: Homogeneous_subsets_v1.r
```

```
postHocHomSubsets(LongData$Opnames, LongData$Regio, p.adjust.method = "tukey", alpha = 0.05)
```

```
##
##
##  Tukey-Kramer post-hoc test:
##
##             Noord       Oost       West        Zuid
## Noord 1.000000000 0.11078676 0.21280829 0.001823446
## Oost  0.110786756 1.00000000 0.96268030 0.053251421
## West  0.212808290 0.96268030 1.00000000 0.027473755
## Zuid  0.001823446 0.05325142 0.02747375 1.000000000
##
##
## Homogeneous subsets:
##
##       subset.1 subset.2
## Noord    18.77
## West     51.62
## Oost     58.82  58.8200
## Zuid           106.7533
##
## (Means per level per subset)
```

Dus {N,W} < {O} < {Z}

```
postHocHomSubsets(LongData$Opnames, LongData$Regio, p.adjust.method = "fdr", alpha = 0.05)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  y and g
##
##      Noord  Oost   West
## Oost 0.0447 -      -
## West 0.0749 0.6480 -
## Zuid 0.0024 0.0269 0.0200
##
## P value adjustment method: fdr
##
##
## Homogeneous subsets:
##
##       subset.1 subset.2 subset.3
## Noord    18.77
## West     51.62    51.62
## Oost              58.82
## Zuid                    106.7533
##
## (Means per level per subset)
```

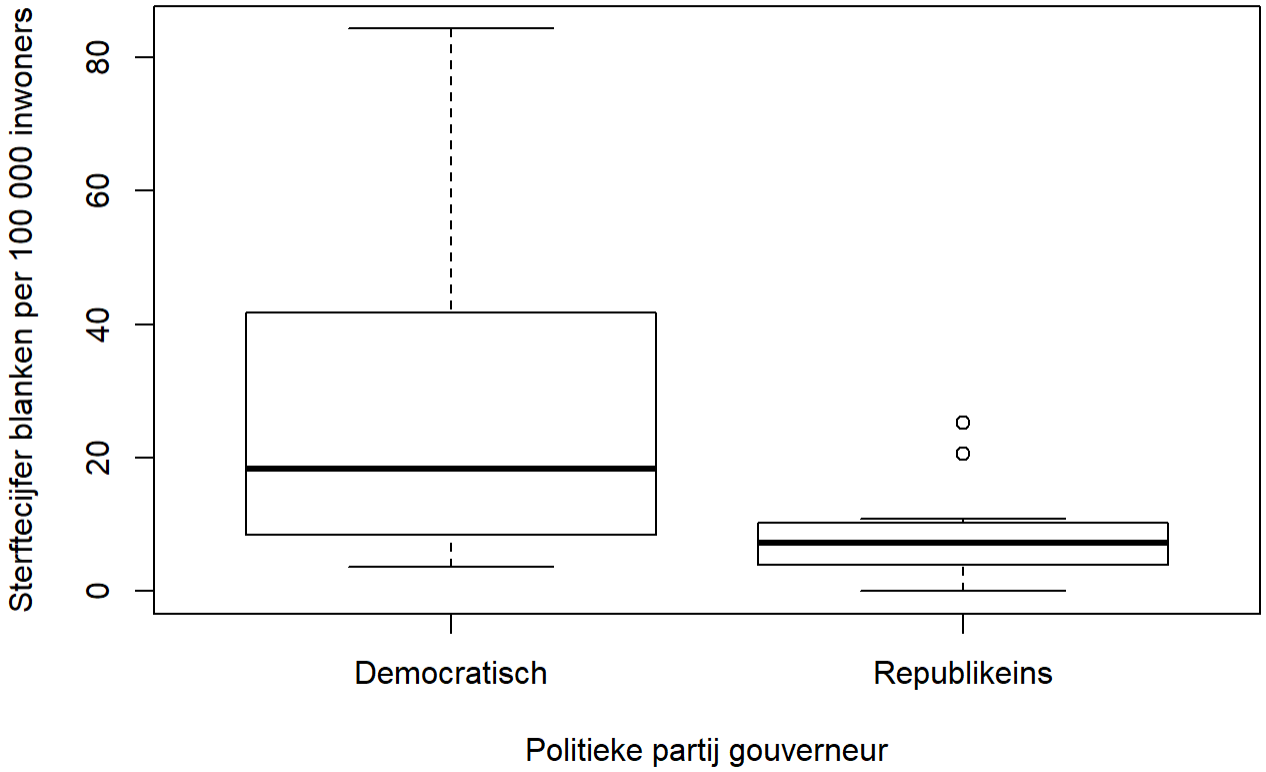
Dus {N} < {W, O} < {Z}

Opgave 3 - Sterftecijfer Covid-19 in de VS: republikeinen vs. democraten

Het SARS-CoV-2 virus woedt nog steeds rond. Onder andere in de VS maakt het veel slachtoffers. De republikeinse regering o.l.v. Donald Trump ontkende eerst het gevaar van het virus. Democratische gouverneurs van verschillende staten kondigden echter al snel allerlei maatregelen aan om verspreiding van het virus te stoppen. Republikeinse gouverneurs volgden pas later maar beëindigden de maatregelen ook weer eerder. Democratische staten zijn vooral de stedelijke gebieden aan de oost- en westkust, republikeinse staten zijn vooral het dunbevolktere platteland in het midden. Voor 26 van de 51 random gekozen staten in de VS is het corona sterftecijfer van *blanken* per 100 000 inwoners bepaald (peildatum 9 juni 2020), alsmede de partij van de gouverneur: D = democraat of R = republikein. De gegevens staan in de file Opgave_3.txt. Dit is een tab-separated ascii file met header.

a. Lees de data in, en maak in R een goede boxplot van het corona sterftecijfer van blanken als functie van de partij van de gouverneur. Let op de as-titels en eenheden!

```
# setwd(myPath)
# dir()
myData <- read.table(paste0(myPath, "Opgave_3_1.txt"), header = T, sep = "\t")
myData
levels(myData$Gouverneur) <- c("Democratisch", "Republikeins")
boxplot(Sterfte ~ Gouverneur, data = myData, xlab = "Politieke partij gouverneur", ylab = "Sterftecijfer blanken p
er 100 000 inwoners")
```



```
##                      Sterfte Gouverneur
## Alabama                 10.9          R
## Colorado                25.1          D
## Delaware                41.8          D
## District.of.Columbia    20.8          D
## Illinois                33.9          D
## Indiana                 25.2          R
## Kansas                   6.5          D
## Louisiana               43.4          D
## Maine                    7.1          D
## Minnesota               15.8          D
## Mississippi             20.6          R
## Missouri                 9.7          R
## New.York                84.3          D
## North.Carolina           8.4          D
## North.Dakota             2.0          R
## Oregon                   3.6          D
## Rhode.Island            74.3          D
## South.Carolina           7.3          R
## Tennessee                4.6          R
## Texas                    5.1          R
## Vermont                  8.8          R
## Virginia                14.7          D
## West.Virginia            3.4          R
## Wisconsin                9.1          D
## Wyoming                  0.0          R
```

b. Leg uit welke statistische toets geschikt is om te onderzoeken of er een verschil is in corona sterftecijfer tussen democratische en republikeinse staten.

Een Welch t-toets, want y = continu, en 2 gemiddelden en niet gepaarde data.

c. Is er een verschil in corona sterftecijfer tussen democratische en republikeinse staten? Toets met $\alpha = 0.05$.

```
( res <- t.test(Sterfte ~ Gouverneur, data = myData) )
p.value <- res$p.value

cat("p-waarde = ",p.value,"\n")
```

```
##
##  Welch Two Sample t-test
##
## data:  Sterfte by Gouverneur
## t = 2.6312, df = 15.964, p-value = 0.01818
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.669627 34.127776
## sample estimates:
## mean in group Democratisch mean in group Republikeins
##                  27.771429                   8.872727
##
## p-waarde =  0.01817608
```

p-waarde = 0.018 < 0.05, dus H1: Er is een significant verschil in sterftecijfer tussen democratische en republikeinse staten.

d. Wat is de effectsterkte? Is het verschil in corona sterftecijfer tussen democratische en republikeinse staten een klein, matig of groot verschil?

```
# myData$Gouverneur
( n.s <- tapply(myData$Gouverneur,myData$Gouverneur,length) )
t <- res$statistic
d.s <- t * sqrt(1/n.s[1] + 1/n.s[2])

cat("Effectsterkte: d.s = ",d.s,"\n")
```

```
## Democratisch Republikeins
##           14           11
## Effectsterkte: d.s =  1.06014
```

d.s groter dan 0.8, dus een groot verschil tussen democratische en republikeinse staten.

e. Toont deze analyse eenduidig aan dat er een significant verschil zit tussen de resultaten van de democratische en de republikeinse aanpak van de coronacrisis? Licht je antwoord kort toe, en geef ook een mogelijke andere verklaring voor de eventueel gevonden verschillen in sterftecijfer.

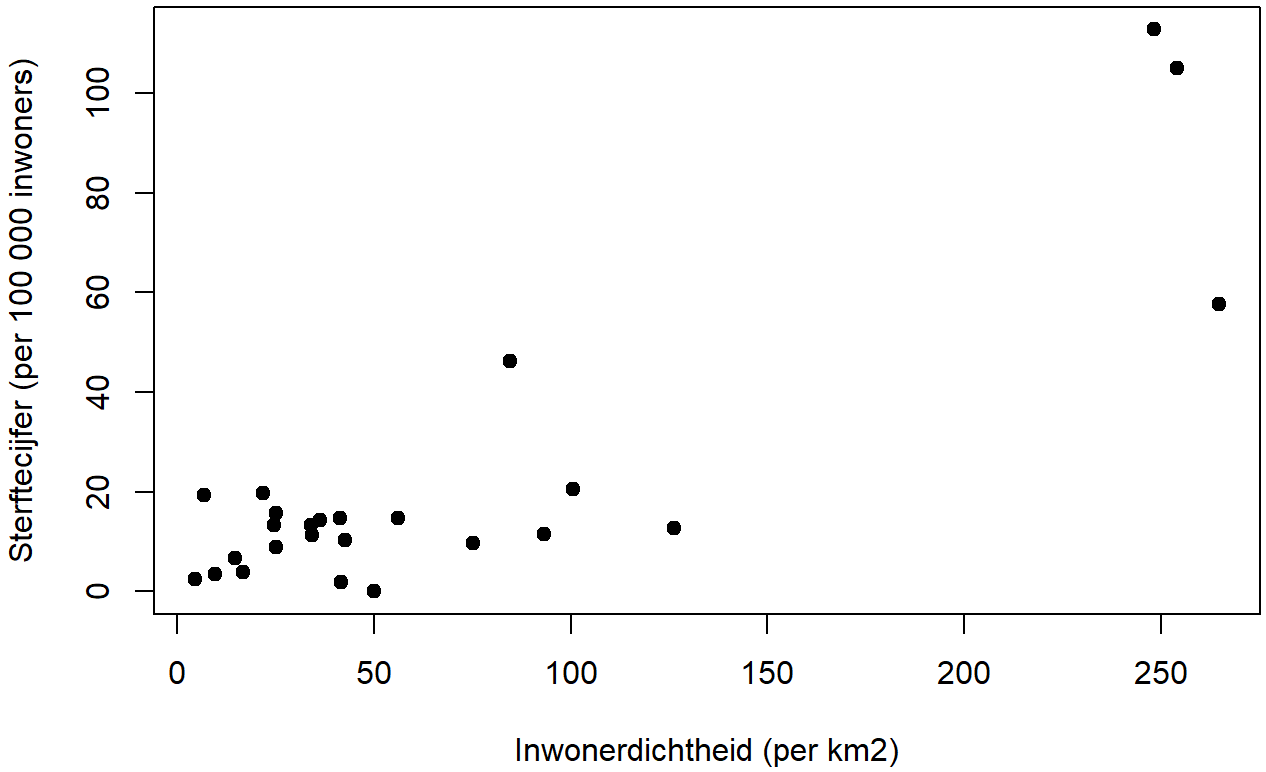
Nee, want er zijn ook andere factoren per staat die een rol kunnen spelen: - ouderdom (leeftijdsopbouw) van de bevolking - bevolkingsdichtheid - welvaart van de bevolking

Opgave 4 - Corona sterftecijfer in de VS: effect van bevolkingsdichtheid?

Net als vele landen wordt de VS hard getroffen door de corona uitbraak. Aangezien het virus zich van persoon naar persoon verspreidt kun je je afvragen of het corona sterftecijfer (per 100 000 inwoners) van een staat binnen de VS afhankelijk is van de bevolkingsdichtheid (in aantal inwoners per km2). De gegevens voor 25 van de 52 staten (op peildatum 9 juni 2020) staan in de file Opgave_4.txt. Dit is een tab-separated ascii file met header.

a. Lees de data in R in, en maak een goede scatterplot van het corona sterftecijfer als functie van de bevolkingsdichtheid. Let op de as-titels en eenheden!

```
# setwd(myPath)
# dir()
myData <- read.table(paste0(myPath, "Opgave_4_1.txt"), header = T, sep = "\t")
myData

plot(Sterftecijfer ~ Inwonerdichtheid, data = myData, pch = 19,
     ylab = "Sterftecijfer (per 100 000 inwoners)",
     xlab = "Inwonerdichtheid (per km2)")
```


Sterftecijfer (per 100 000 inwoners)
0
20
40
60
80
100
0
50
100
150
200
250
Inwonerdichtheid (per km2)

```
##             State Sterftecijfer Inwonerdichtheid
## 1         Alabama          14.2             36.11
## 2         Arizona          13.3             24.65
## 3      California          11.4             93.20
## 4     Connecticut         112.8            248.33
## 5         Florida          12.6            126.11
## 6          Hawaii           0.0             50.01
## 7        Illinois          46.2             84.48
## 8            Iowa          19.8             21.65
## 9        Kentucky          10.2             42.69
## 10          Maine           6.7             14.67
## 11  Massachusetts         105.0            254.23
## 12      Minnesota          15.7             25.05
## 13       Missouri          13.3             33.99
## 14       Nebraska           3.4              9.66
## 15  New.Hampshire          14.7             56.15
## 16     New.Mexico          19.3              6.66
## 17 North.Carolina           9.7             75.24
## 18           Ohio          20.6            100.68
## 19         Oregon           3.8             16.55
## 20   Rhode.Island          57.7            264.77
## 21   South.Dakota           2.4              4.43
## 22          Texas           1.9             41.68
## 23        Vermont           8.8             25.05
## 24     Washington          14.7             41.24
## 25      Wisconsin          11.2             34.32
```

Om te onderzoeken of er een (statistisch) verband is tussen sterftecijfer en bevolkingsdichtheid kun je in R lineaire regressie uitvoeren met het commando summary(lm(y ~ x)) waarbij y het sterftecijfer is, en x de bevolkingsdichtheid. Je fit dan een lijn $y = a + b \cdot x$ door de grafiek; a een b zijn de coëfficiënten van het regressiemodel.

b. Voer lineaire regressie uit op de data. Wat is de waarde (“estimate”) van de as-afsnede (“intercept”) a, en wat is de waarde van de helling b?

```
res <- summary(lm(Sterftecijfer ~ Inwonerdichtheid, data = myData))
res
a <- res$coefficients[1,1]
b <- res$coefficients[2,1]
cat("\n\nCoefficienten:\na = ",a,", b = ",b)
```

```
##
## Call:
## lm(formula = Sterftecijfer ~ Inwonerdichtheid, data = myData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.220 -11.054   1.285   5.916  32.257
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.67802    4.08480  -0.166     0.87
## Inwonerdichtheid  0.32707    0.03994   8.188 2.87e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.03 on 23 degrees of freedom
## Multiple R-squared:  0.7446, Adjusted R-squared:  0.7335
## F-statistic: 67.05 on 1 and 23 DF,  p-value: 2.871e-08
##
##
##
## Coefficienten:
## a =  -0.6780224 , b =  0.3270678
```

De uitvoer van lineaire regressie geeft meer dan alleen maar de coëfficiënten a en b: het geeft ook de standaardfouten in a en b (“Std. Error”) en voert per coëfficiënt ook een t-toets uit met als hypotheses:

- H0: de parameter is 0;
- H1: de parameter is niet 0

De uitvoer van lineaire regressie geeft de t-waarde van deze toets (“t value”) en de p-waarde van de toets (“Pr(>|t|)”).

c. Is de waarde van de helling b significant anders dan 0? Toets met $\alpha = 0.05$.

```
p.value.b <- res$coefficients[2, 4]
cat("p-waarde: ",p.value.b)
```

```
## p-waarde:  2.870509e-08
```

De p-waarde is 2.87e-08 << 0.05, dus is b significant anders dan de waarde “0”,

d. Is er een significant verband tussen sterftecijfer en bevolkingsdichtheid?

Ja, omdat de coefficient b significant anders is dan 0, is er een significant verband tussen sterftecijfer en bevolkingsdichtheid.

De effectsterkte van regressie wordt gegeven door de waarde van R^2 (“Multiple R-squared”): dit is de fractie van de variatie in de y-waarden die wordt “verklaard” door het regressiemodel (dus: de lijn). De interpretatie is gelijk aan die van η^2 bij een 1-way ANOVA.

e. Wat is de effectsterkte van de bevolkingsdichtheid op het corona sterftecijfer? Heeft de bevolkingsdichtheid een klein, matig of sterk effect op het sterftecijfer?

```
R2 <- res$r.squared
cat("R2 = ",R2)
```

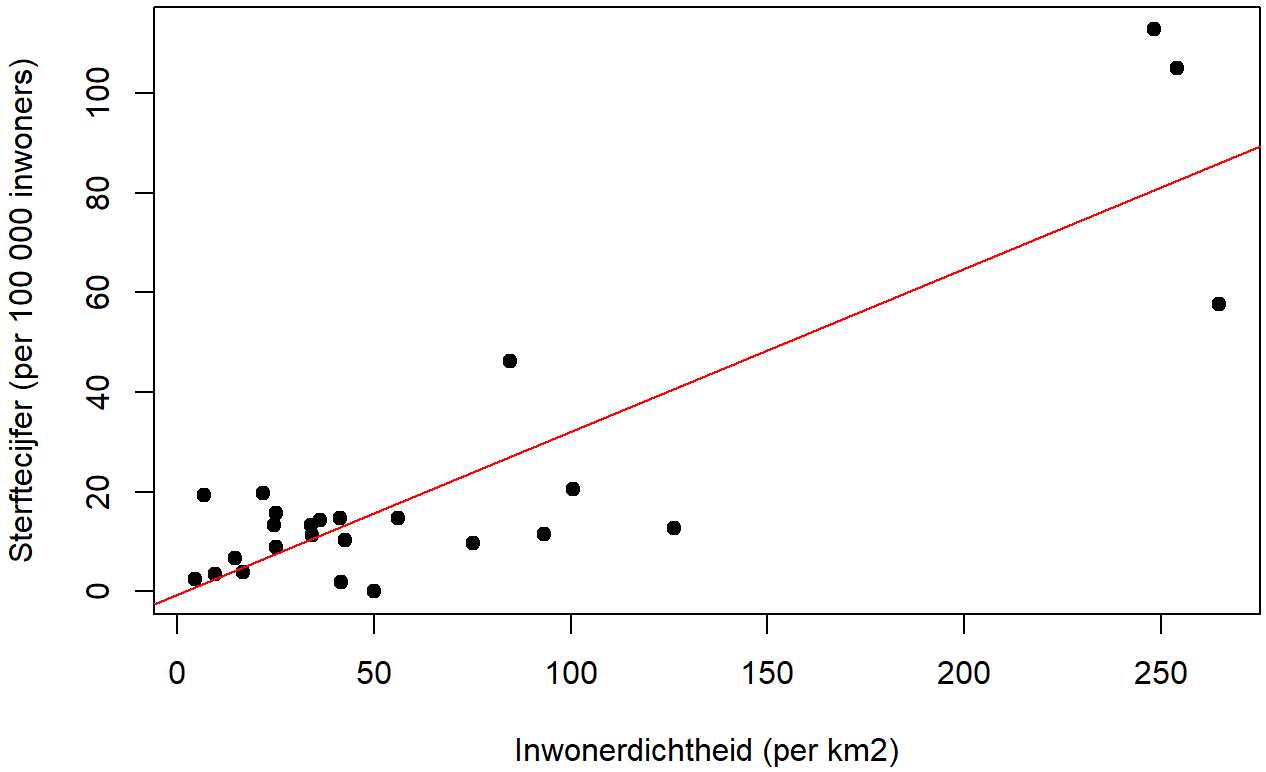
```
## R2 =  0.7445846
```

De waarde van R2 >> 0.14, dus een heel erg sterk effect!

f. Plot de gefitte regressielijn $y = a + b \cdot x$ in de grafiek van opgave **a.** Hint: gebruik de functie abline.

```
plot(Sterftecijfer ~ Inwonerdichtheid, data = myData, pch = 19,
     ylab = "Sterftecijfer (per 100 000 inwoners)",
     xlab = "Inwonerdichtheid (per km2)")
abline(res, col = "red")
```

```
## Warning in abline(res, col = "red"): only using the first two of 8 regression
## coefficients
```



g. Wat zouden andere verklarende factoren voor het sterftecijfer kunnen zijn? Noem twee suggesties.

Bijvoorbeeld:

- Percentage Afro-Americans
- Percentage ouderen
- Levenstandaard
- Luchtkwaliteit

Opgave 5 - Clusteren van voedingsmiddelen

a. Lees deze file in R in als dataframe. Wat is het gemiddelde aantal besmettingen, herstelden en doden per 100 000 inwoners? En wat is de gemiddelde fijnstofconcentratie?

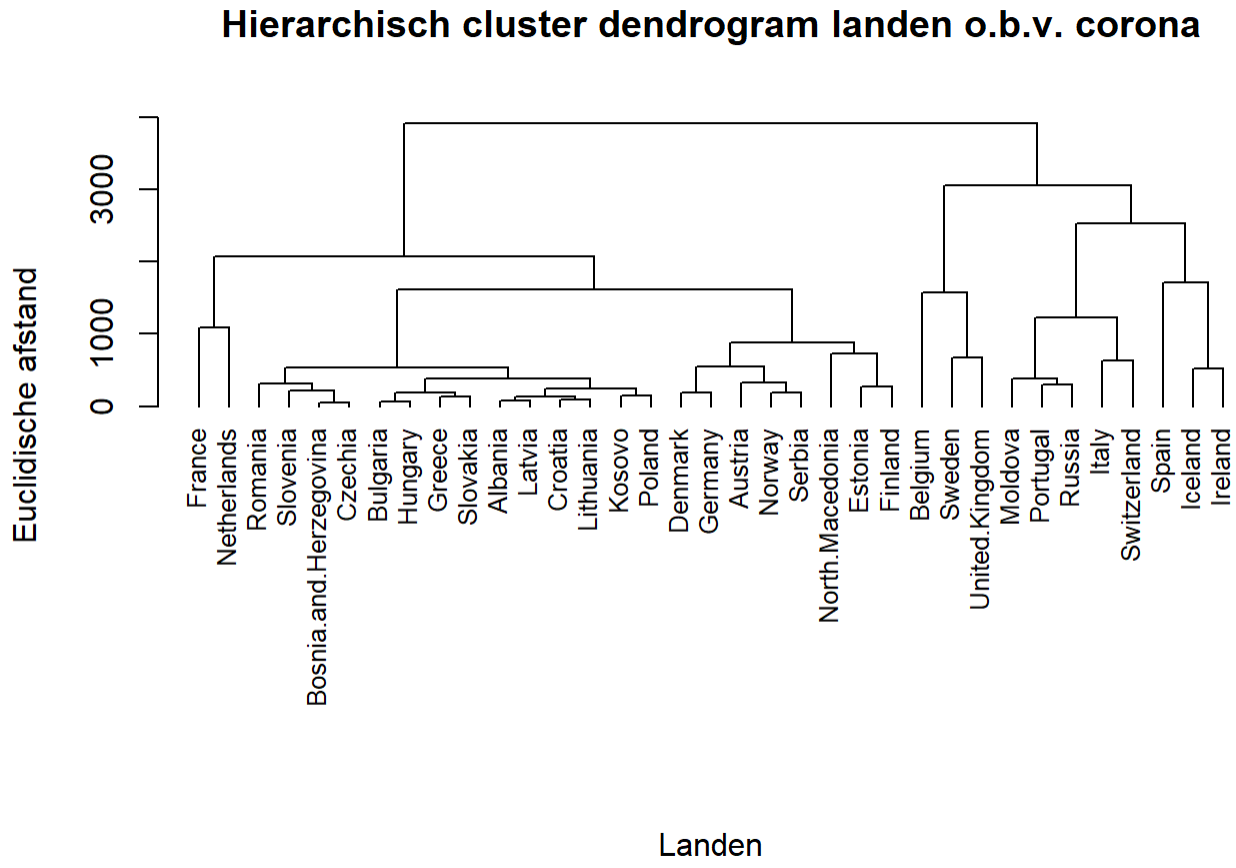
```
# setwd(myPath)
# dir()
myData <- read.table(paste0(myPath, "Opgave_5_1.txt"), header = T, sep = "\t")
# myData

corona <- c("Confirmed", "Recovered", "Dead")
data2 <- myData[, corona]
rownames(data2) <- rownames(myData)
head(data2)

Confirmed.av <- mean(data2$Confirmed)
Recovered.av <- mean(data2$Recovered)
Dead.av <- mean(data2$Dead)

cat("\n\nResultaten per categorie:\n")
cat("Confirmed = ",Confirmed.av," per 100 000 inwoners\n")
cat("Recovered = ",Recovered.av," per 100 000 inwoners\n")
cat("Dead      = ",Dead.av," per 100 000 inwoners\n")
```

```
##                        Confirmed Recovered  Dead
## Albania                    554.7     368.1  12.6
## Austria                   1936.8    1816.0  76.6
## Belgium                   5261.7    1454.2 845.8
## Bosnia.and.Herzegovina     914.6     650.4  49.6
## Bulgaria                   477.3     254.9  25.1
## Croatia                    551.2     523.3  26.2
##
##
## Resultaten per categorie:
## Confirmed =  2230.523  per 100 000 inwoners
## Recovered =  1334.343  per 100 000 inwoners
## Dead      =  166.6457  per 100 000 inwoners
```

b. Voer hiërarchische clustering uit op de landen op basis van de kolommen Confirmed, Recovered en Dead. Gebruik euclidische afstanden en

“average” linkage. Maak een dendrogram van het resultaat.

```
dMat <- dist(data2, method = "euclidean")
hcl <- hclust(dMat, method = "average")
plot(hcl, hang = -1, xlab = "Landen", ylab = "Euclidische afstand", main = "Hierarchisch cluster dendrogram landen
o.b.v. corona", sub = "", cex = 0.8)
```



c. Splits de cluster in 2 subclusters m.b.v. de functie cutree. Welke landen horen bij subcluster 1 en welke landen horen bij subcluster 2?

```
hsubclusters <- cutree(hcl, k = 2)
landen <- rownames(data2)

cat("\nHierarchisch clusteren:\n")
cat("\nEerste subcluster: ",landen[hsubclusters == 1],"\n")
cat("\nTweede subcluster: ",landen[hsubclusters == 2],"\n")
```

```
##
## Hierarchisch clusteren:
##
## Eerste subcluster:  Albania Austria Bosnia.and.Herzegovina Bulgaria Croatia Czechia Denmark Estonia Finland Fra
nce Germany Greece Hungary Kosovo Latvia Lithuania Netherlands North.Macedonia Norway Poland Romania Serbia Slovak
ia Slovenia
##
## Tweede subcluster:  Belgium Iceland Ireland Italy Moldova Portugal Russia Spain Sweden Switzerland United.Kingd
om
```

d. Toets of er een significant verschil zit in fijnstofconcentratie tussen de landen van subcluster 1 en van subcluster 2. Toets met $\alpha = 0.05$.

```
( res <- t.test(myData$PM2.5 ~ hsubclusters) )
p.value <- res$p.value

cat("p-waarde = ",p.value,"\n")
```

```
##
##  Welch Two Sample t-test
##
## data:  myData$PM2.5 by hsubclusters
## t = 2.8527, df = 31.596, p-value = 0.007582
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.625980 9.760005
## sample estimates:
## mean in group 1 mean in group 2
##        16.85208        11.15909
##
## p-waarde =  0.007582092
```

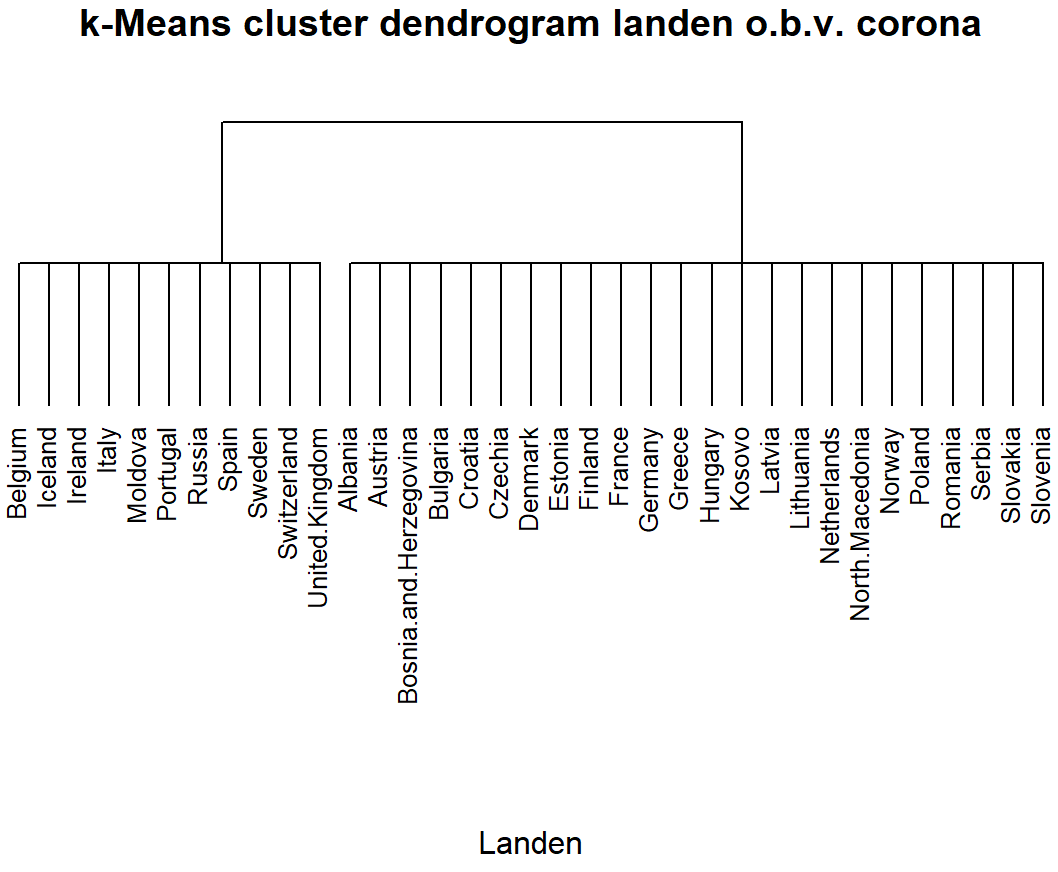
p-waarde = 0.007 << 0.05, dus significant anders!

e. Voer k-means clustering uit op de landen op basis van de kolommen Confirmed, Recovered en Dead. Gebruik $k = 2$ clusters en gebruik 5 random startwaarden voor de clustercentra. Maak m.b.v. de speciale functie makeKMeansDendrogram (zie map “R functies” op Blackboard) een dendrogram van het resultaat.

```
source("Dendrograms_v8.R")
```

```
## Sourced: Dendrograms_v8.r
```

```
kcl <- kmeans(data2, centers = 2, nstart = 5)
hkcl <- makeKMeansDendrogram(kcl)

plot(hkcl, hang = -1, xlab = "Landen", ylab = "", main = "k-Means cluster dendrogram landen o.b.v. corona", sub =
"", cex = 0.8, axes = F)
```



f. Zijn de twee k-means clustering subclusters hetzelfde als de twee hierarchische subclusters bij **c.**?

```
cat("\nHierarchisch clusteren:\n")
cat("\nEerste subcluster: ",landen[hsubclusters == 1],"\n")
cat("\nTweede subcluster: ",landen[hsubclusters == 2],"\n")

ksubclusters <- kcl$cluster
cat("\nk-Means clusteren:\n")
cat("\nEerste subcluster: ",landen[ksubclusters == 1],"\n")
cat("\nTweede subcluster: ",landen[ksubclusters == 2],"\n")
```

```
##
## Hierarchisch clusteren:
##
## Eerste subcluster:  Albania Austria Bosnia.and.Herzegovina Bulgaria Croatia Czechia Denmark Estonia Finland Fra
nce Germany Greece Hungary Kosovo Latvia Lithuania Netherlands North.Macedonia Norway Poland Romania Serbia Slovak
ia Slovenia
##
## Tweede subcluster:  Belgium Iceland Ireland Italy Moldova Portugal Russia Spain Sweden Switzerland United.Kingd
om
##
## k-Means clusteren:
##
## Eerste subcluster:  Belgium Iceland Ireland Italy Moldova Portugal Russia Spain Sweden Switzerland United.Kingd
om
##
## Tweede subcluster:  Albania Austria Bosnia.and.Herzegovina Bulgaria Croatia Czechia Denmark Estonia Finland Fra
nce Germany Greece Hungary Kosovo Latvia Lithuania Netherlands North.Macedonia Norway Poland Romania Serbia Slovak
ia Slovenia
```

Ja, ze zijn hetzelfde!

Opgave 6 - Analyse op DEG's (1)

a. Lees de microarray data in R in als dataframe. Bekijk de structuur van de data.

```
# setwd(myPath)
# dir()
myData <- read.table(paste0(myPath, "Opgave_6_1.txt"), header = T, sep = "\t")
# myData

head(myData)
```

```
##               R.1       R.2       R.3       R.4       R.5      R.6      G.1
## gen.001 2.6806704 1.3170036 2.0978309 2.1557271 2.3943296 1.796606 3.630958
## gen.002 4.3841791 3.6422650 3.9353895 3.7801228 4.4460060 3.940015 5.801292
## gen.003 0.1562163 0.6886828 0.7984332 0.6949271 0.8563135 1.308906 4.845057
## gen.004 4.9002565 3.7722591 4.1441646 3.9869362 4.0910950 4.558654 4.452592
## gen.005 3.5667736 3.9165948 4.2382296 4.3002958 4.1408580 3.673826 7.616799
## gen.006 3.5506823 4.2886492 3.6605072 3.5681640 3.9425575 4.360557 4.697073
##              G.2      G.3      G.4      G.5      G.6
## gen.001 3.739944 3.445893 3.643647 3.876452 3.452070
## gen.002 6.008549 5.576546 4.342391 5.665738 5.502788
## gen.003 2.807616 4.258744 4.024509 4.072796 4.689868
## gen.004 4.069242 4.420451 4.061025 4.390478 3.809750
## gen.005 6.796648 6.603631 7.090200 6.462617 6.764509
## gen.006 4.160956 3.622111 3.269113 4.222366 4.322069
```

b. Welke statistische toets is geschikt om direct dit dataframe te analyseren op de aanwezigheid van DEG's (differentially expressed genes)?

Een gepaarde t-toets, want gepaard vanwege R en G metingen op hetzelfde MA.

c. Maak een eigen functie die de toets van **b.** uitvoert per regel van het dataframe met microarray data.

```
myTTest <- function(x){
  g <- factor(rep(c(1, 2), each = 6))
  return(t.test(x ~ g, paired = T)$p.value)
}
```

d. Voer deze functie uit per gen (= regel van het dataframe). Hoeveel en welke genen komen zonder multiple testing correctie differentieel tot expressie (d.w.z. zijn DEG's)? Toets met $\alpha = 0.05$.

```
alpha <- 0.05
genes <- rownames(myData)

pVals <- apply(myData, 1, myTTest)

n <- sum(pVals < alpha)
cat("Aantal DEG's zonder correctie: ",n,"\n")
DEGs <- genes[pVals < alpha]
cat("DEG's: ",DEGs,"\n")
```

```
## Aantal DEG's zonder correctie:  11
## DEG's:  gen.001 gen.002 gen.003 gen.005 gen.008 gen.013 gen.021 gen.034 gen.039 gen.055 gen.089
```

e. Pas nu FDR multiple testing correctie toe op de p-waarden. Hoeveel en welke genen komen na correctie differentieel tot expressie (d.w.z. zijn werkelijk DEG's)? Toets met $\alpha = 0.05$.

```
pVals.FDR <- p.adjust(pVals, method = "fdr")
n <- sum(pVals.FDR < alpha)
cat("Aantal DEG's zonder correctie: ",n,"\n")
DEGs <- genes[pVals.FDR < alpha]
cat("DEG's: ",DEGs,"\n")
```

```
## Aantal DEG's zonder correctie:  9
## DEG's:  gen.001 gen.002 gen.003 gen.005 gen.013 gen.021 gen.034 gen.055 gen.089
```

f. Hoeveel vals positieve uitslagen verwacht je zonder multiple testing correctie?

```
G <- nrow(myData)
cat("Verwacht aantal vals positieven: ",G*alpha,"\n")
```

```
## Verwacht aantal vals positieven:  5
```

g. Maak van de bij **e.** gevonden DEG's een heatmap. Cluster zowel de genen als de samples o.b.v. euclidische afstand en average linkage. Gebruik een geschikt kleurenpalet.

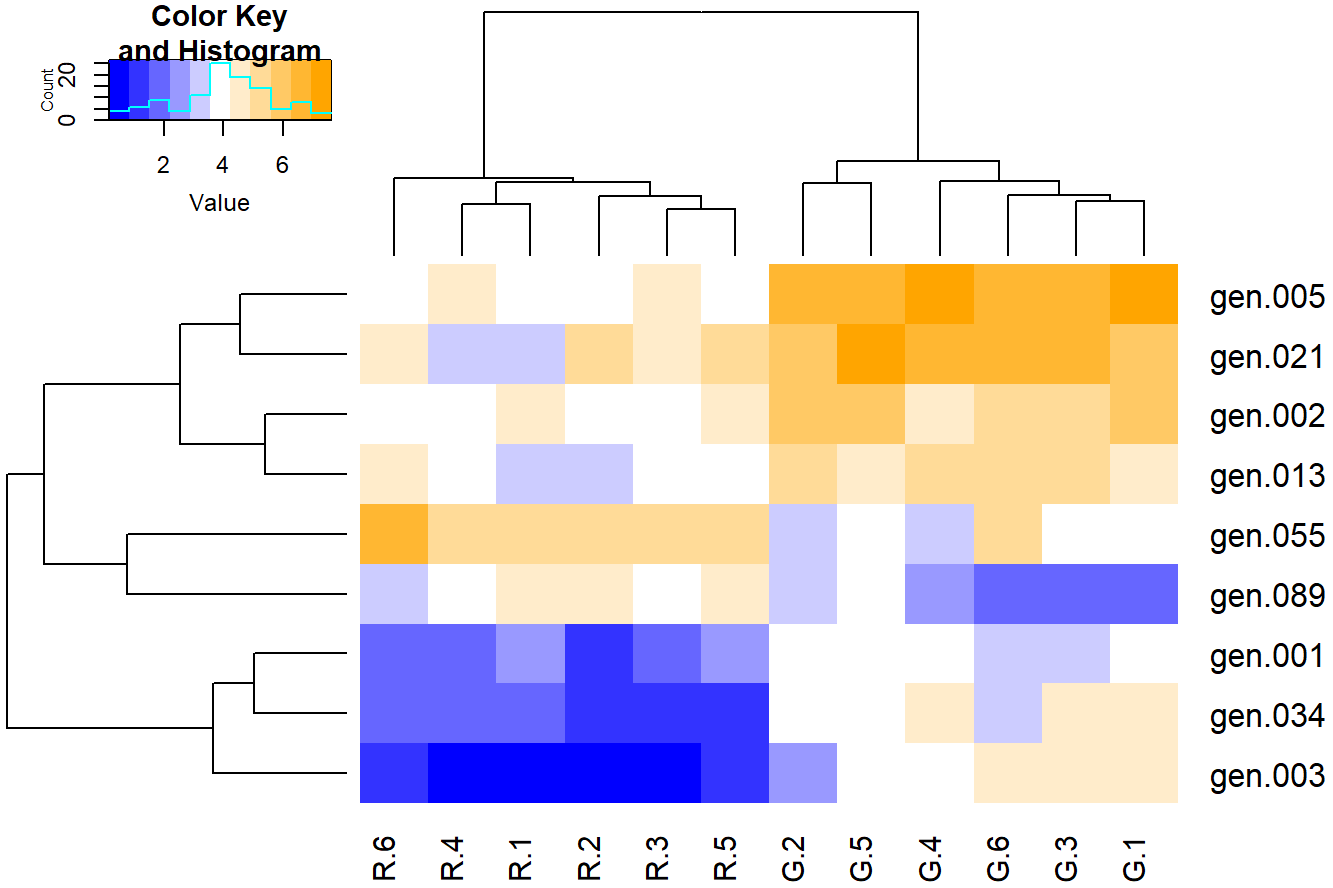
```
# kleurpalet:
MA.color <- function(n=11){
  colorRampPalette(c("blue", "white", "orange"), space = "rgb")(n)
}

# Definieer zelf eigen distance en linkage functies:

myDist <- function(x){
  return(dist(x, method = "euclidean"))
}

myHClust <- function(x){
  return(hclust(x, method = "average"))
}

# netste manier: via heatmap.2 in gplots

library("gplots")
```

```
## Warning: package 'gplots' was built under R version 3.6.3
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
heatmap.2(as.matrix(myData[DEGs, ]), distfun=myDist, hclustfun = myHClust,
          symbreaks = F, trace = "none", scale = "none", col = MA.color())
```


Color Key
and Histogram
Count
0
20
2
4
6
Value
gen.005
gen.021
gen.002
gen.013
gen.055
gen.089
gen.001
gen.034
gen.003
R.6
R.4
R.1
R.2
R.3
R.5
G.2
G.5
G.4
G.6
G.3
G.1