UNIVERSITEIT GENT

Statistische gegevensanalyse

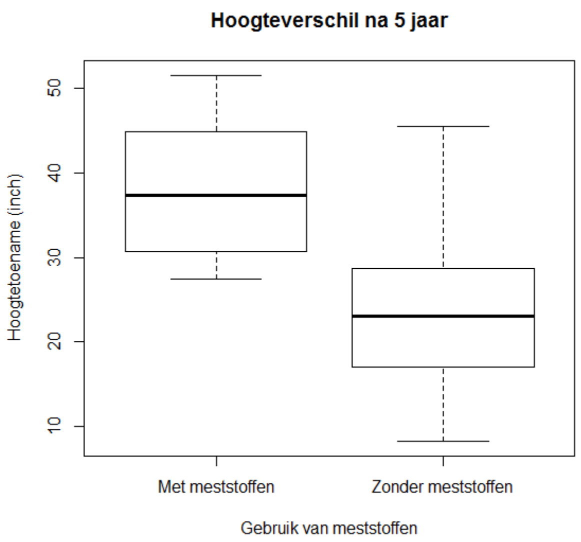
Project

Bart Middag
3de bachelor informatica
Academiejaar 2013-2014

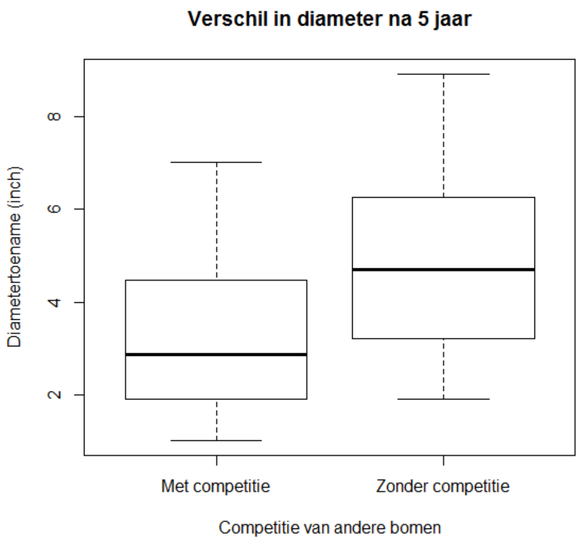
1. DE ZWARTE SPAR

a) Het effect van gebruik van meststoffen op de hoogte:

We bepalen de groei van de bomen met en zonder gebruik van meststoffen. Het resultaat wordt getoond in Figuur 1. In deze steekproef is er dus duidelijk een verschil tussen de groei met gebruik van meststoffen en zonder. Om met zekerheid te kunnen zeggen dat dit geen toeval is, bepalen we aan de hand van een permutatietest dus de kans dat het verschil even groot of groter is. Voor geen enkele van de 10000 permutaties was het verschil zo groot als in de originele steekproef. De kans om een dergelijk verschil te bekomen is dan ook $\frac{1}{10001}$. We kunnen met 95% zekerheid stellen dat bomen bij gebruik van meststoffen gemiddeld tussen 11.60278 en 17.88889 inch hoger zullen worden dan zonder meststoffen.



Figuur 1: Boxplot van de groei van de zwarte spar in verband met het gebruik van meststoffen.



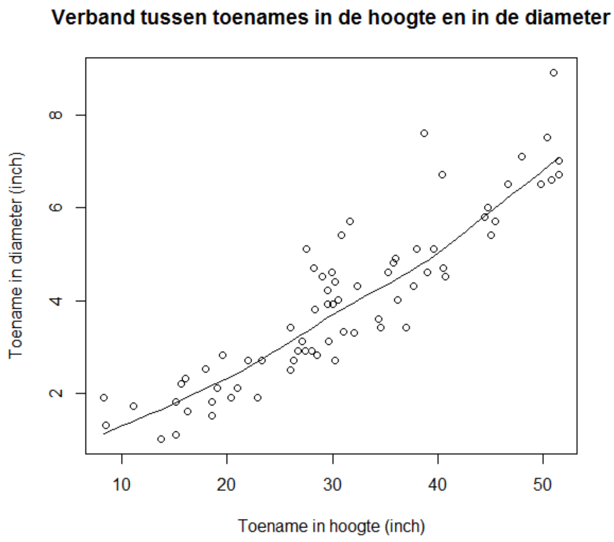
Figuur 2: Boxplot van de groei van de diameter van de zwarte spar in verband met competitie van andere bomen in de omgeving.

b) Het effect van de omgeving op de diameter:

Analoog met de vorige opgave, bepalen we de diameter van de bomen met en zonder competitie van andere bomen in de omgeving. Het resultaat in Figuur 2 vertoont een duidelijk negatief effect van bomen in de omgeving. Aan de hand van een t-test bekomen we dat we met 95% zekerheid kunnen stellen dat bomen met competitie van andere bomen in de omgeving tussen 0.81824 en 2.33575 inch minder dik zullen worden dan bomen zonder competitie.

c) Het verband tussen de toename van de hoogte en van de diameter:

We bepalen het verband tussen de toename van de hoogte en van de diameter aan de hand van Pearsons product-moment correlatiecoëfficiënt. Deze wordt geschat op 0.90208. Er is dus bijna een perfect lineair verband tussen de toename van de hoogte en van de diameter. We kunnen met 95% zekerheid stellen dat dit ligt tussen 0.84753 en 0.93777. Dit verband wordt duidelijk getoond op Figuur 3.



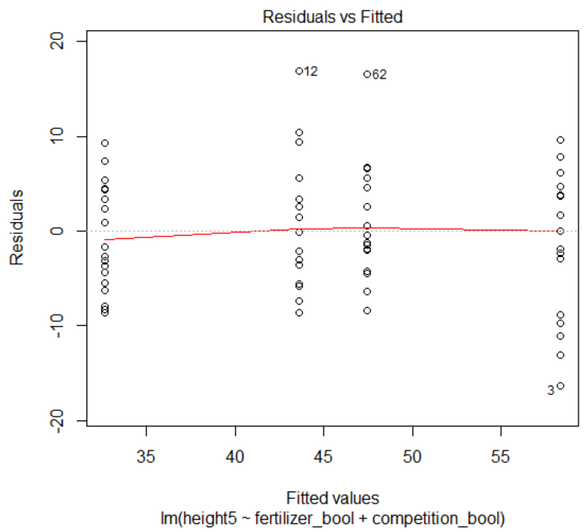
Figuur 3: Scatterplot van het verband tussen toename in de hoogte en toename in de diameter van de zwarte spar.

d) Model van de hoogte na 5 jaar a.d.h.v. meststofgebruik en competitie:

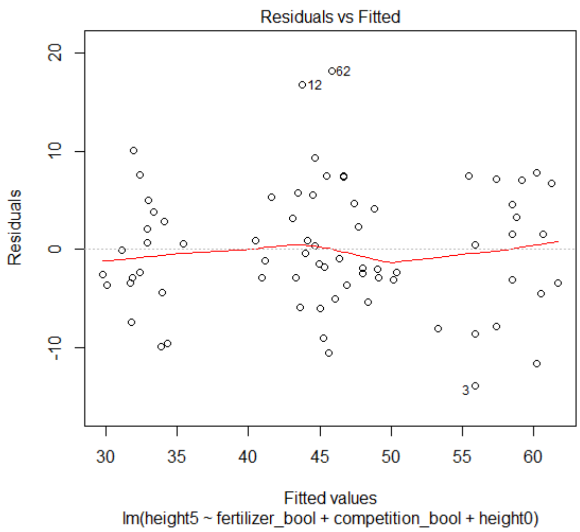
Na toepassing van lineaire regressie, bekomen we het volgende model:

$$w = 43.578 + 14.772 * x - 10.917 * y$$

Hierbij staat x voor het gebruik van meststoffen en y voor de aanwezigheid van andere bomen in de omgeving. De p-waarde die bij de nulhypothese hoort die stelt dat er geen invloed is van meststoffen en/of competitie, is kleiner dan $2.2 * 10^{-16}$. Zoals we onderzocht hebben in opgave a en b, hebben beide variabelen dus invloed op de hoogte. De determinatiecoëfficiënt van dit model is 0.6586. De standaardfout van het residu is 6.611. Op Figuur 4 zien we de residuen tegenover de gefitte waarden.



Figuur 4: Residuen tegenover de gefitte waarden van een model dat de beginhoogte niet in rekening brengt.



Figuur 5: Residuen tegenover de gefitte waarden van een model dat de beginhoogte in rekening brengt.

e) Model van de hoogte na 5 jaar a.d.h.v. meststofgebruik, competitie en beginhoogte:
Na toepassing van lineaire regressie, bekomen we het volgende model:

$$w = 30.8296 + 14.7195 * x - 10.5235 * y + 0.8631 * z$$

Hierbij staat x voor het gebruik van meststoffen, y voor de aanwezigheid van andere bomen in de omgeving en z voor de beginhoogte. De p-waarde die bij de nulhypothese hoort die stelt dat er geen invloed is van meststoffen en/of competitie en/of de beginhoogte, is kleiner dan $2.2 * 10^{-16}$. Alle variabelen hebben invloed op de hoogte na 5 jaar. De determinatiecoëfficiënt van dit model is 0.6786. De standaardfout van het residu is 6.414. In vergelijking met het vorige model ligt de fout lager en de determinatiecoëfficiënt hoger: dit model is dus beter. Op Figuur 5 zien we de residuen tegenover de gefitte waarden.

f) Schattingen voor de gemiddelde hoogte van de spar na 5 jaar:
Aan de hand van het laatste model voeren we de schattingen uit. De resultaten zijn weergegeven in Tabel 1.

Geen meststoffen	Geen competitie	31.69274 inch
Meststoffen	Geen competitie	46.41222 inch
Geen meststoffen	Competitie	21.16929 inch
Meststoffen	Competitie	35.88876 inch

Tabel 1: Schattingen voor de gemiddelde hoogte van de zwarte spar na 5 jaar

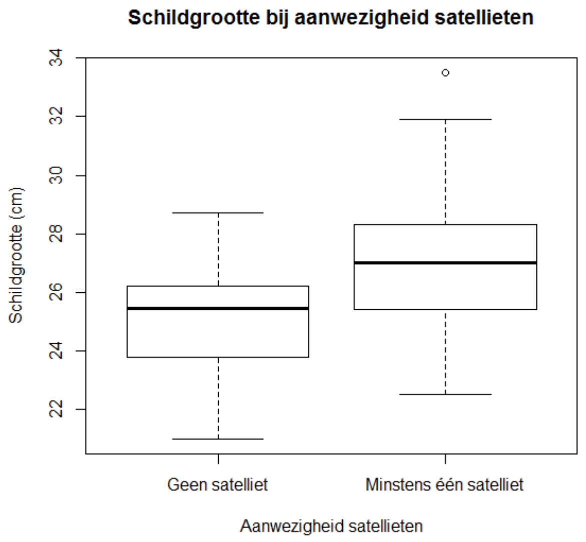
2. DEGENKRABBEN

a) Het percentage vrouwelijke degenkrabben met minstens één satelliet:
Op basis van de steekproef schatten we dat het percentage vrouwelijke degenkrabben met minstens één satelliet gelijk is aan 64.16185%.

b) Het effect van de schildgrootte op de aanwezigheid van satellieten:
Analoog met de opgaves 1a en 1b, bepalen we de schildgrootte van de krabben met en zonder de aanwezigheid van satellieten. Aan de hand van Figuur 6 zien we dat er meer satellieten aanwezig zullen zijn bij een krab met een groter schild. Met een t-test bekomen we dat we met 95% zekerheid kunnen stellen dat krabben met satellieten een schild zullen hebben dat gemiddeld tussen 1.18848 cm en 2.33227 cm groter is dan het schild van krabben zonder satellieten.

c) Model voor de aanwezigheid van satellieten:
Om dit model op te stellen, gebruiken we lineaire en kwadratische discriminantanalyse. Het schijnbaar foutenpercentage bij lineaire discriminantanalyse is 33.07692% en dit bij kwadratische discriminantanalyse is 29.23077%. We beschikken over relatief weinig data, dus gebruiken we kruisvalidatie gebruikt om deze percentages te berekenen. De bijhorende misclassificatietabellen zijn weergegeven in Tabel 2 en Tabel 3. We zien dat

kwadratische discriminantanalyse een kleiner foutenpercentage oplevert, dus we gebruiken dit model. We zien dat dit model een predictiefout van 37.2093% heeft.



Figuur 6: Boxplot van de schildgrootte in verband met de aanwezigheid van satellieten.

LDA	Geen satelliet	≥1 satelliet
Geen satelliet	16.15385%	13.84615%
≥1 satelliet	19.23077%	50.76923%

Tabel 2: Misclassificatietabel bij lineaire discriminantanalyse. De echte klassen worden weergegeven in de kolommen, de voorspelde klassen in de rijen.

QDA	Geen satelliet	≥1 satelliet
Geen satelliet	19.23077%	13.07692%
≥1 satelliet	16.15385%	51.53846%

Tabel 3: Misclassificatietabel bij kwadratische discriminantanalyse. De echte klassen worden weergegeven in de kolommen, de voorspelde klassen in de rijen.

3. APPENDIX: R-CODE

In het verslag werden enkel de resultaten gerapporteerd: de code zelf is iets ingewikkelder en er werden bij enkele opgaves verschillende controles uitgevoerd om te weten welke oplossingsmethode gebruikt moest worden. De werkwijze staat in onderstaande code beschreven. U vindt de code ook terug in het bestand `Project.R`.

```r
##########################################
# Project: Statistische Gegevensanalyse #
# Naam: Bart Middag                     #
# Richting: 3de bachelor informatica    #
# Academiejaar: 2013-2014               #
##########################################

# Voorbereiding project
nsim <- 10000 # aantal simulaties
set.seed(98765) # seed voor reproduceerbare resultaten
setwd("C:/BART/UNIF/Statistische Gegevensanalyse/Project")

############
# OPGAVE 1 #
############

# Voorbereiding opgave 1

zwarte_spar <- read.csv("ZwarteSpar.csv",header = T)
str(zwarte_spar)
summary(zwarte_spar)
colnames(zwarte_spar) <- tolower(colnames(zwarte_spar))
attach(zwarte_spar)

# OPGAVE 1A

# Hoogte met/zonder meststoffen en groei berekenen
f_height0 <- height0[fertilizer == "F"]
f_height5 <- height5[fertilizer == "F"]
nf_height0 <- height0[fertilizer == "NF"]
nf_height5 <- height5[fertilizer == "NF"]
f_growth <- f_height5 - f_height0
nf_growth <- nf_height5 - nf_height0
height_growth <- height5 - height0

# Boxplots
boxplot(height_growth~fertilizer, main="Hoogteverschil na 5 jaar", xlab="Gebruik van meststoffen", ylab="Hoogtetoename (inch)", xaxt="n")
axis(side=1, at=1:2, labels=c("Met meststoffen","Zonder meststoffen"))

# We kijken naar de verdeling van de groei en zien of ze normaal verdeeld is
plot(density(f_growth))
plot(density(nf_growth))
plot(density(height_growth))
qqnorm(f_growth); qqline(f_growth, col = 2)
qqnorm(nf_growth); qqline(nf_growth, col = 2)
qqnorm(height_growth); qqline(height_growth, col = 2)

```

```
# We kijken of de groei normaal verdeeld is - de Shapiro-Wilk test geeft ons duidelijke resultaten.
shapiro.test(f_growth) # Niet normaal verdeeld, de t-test mogen we dus niet gebruiken!
shapiro.test(nf_growth) # Wel normaal verdeeld
shapiro.test(height_growth) # Algemen normaal verdeeld

# We bekijken het verschil in gemiddeldes
mean_diff <- mean(f_growth) - mean(nf_growth)
# We bepalen de kans dat een willekeurige permutatie een resultaat geeft >= het gemiddelde in deze situatie.
means <- numeric()
for(i in 1:nsim) {
  permutation <- sample(height_growth)
  means[i] <- mean(permutation[fertilizer == "F"]) - mean(permutation[fertilizer == "NF"])
}
hist(means)
means[nsim+1] <- mean_diff
means_p <- sum(means >= mean_diff)/(nsim+1)
# Op basis van 10000 permutaties is de kans ongeveer 1/10001.

# We bepalen het betrouwbaarheidsinterval.
differences <- numeric()
for(i in 1:nsim) {
  differences[i] <- mean(sample(f_growth,replace = T)) - mean(sample(nf_growth, replace = T))
}
differences[nsim+1] <- mean_diff
differences <- sort(differences)
interval <- c(differences[0.05*(nsim+1)],differences[(1-0.05)*(nsim+1)])
cat(paste0("We kunnen met 95% zekerheid stellen dat de extra groei zal liggen tussen ", interval[1], " en ", interval[2], " inch."))

# OPGAVE 1B

# Diameter met/zonder competitie en groei berekenen
c_diameter0 <- diameter0[competition == "C"]
c_diameter5 <- diameter5[competition == "C"]
nc_diameter0 <- diameter0[competition == "NC"]
nc_diameter5 <- diameter5[competition == "NC"]
c_growth <- c_diameter5 - c_diameter0
nc_growth <- nc_diameter5 - nc_diameter0
diameter_growth <- diameter5 - diameter0

# Boxplots
boxplot(diameter_growth~competition, main="Verschil in diameter na 5 jaar", xlab="Competitie van andere bomen", ylab="Diametertoename (inch)", xaxt="n")
axis(side=1, at=1:2, labels=c("Met competitie","Zonder competitie"))

# We kijken naar de verdeling van de groei en zien of ze normaal verdeeld is
plot(density(c_growth))
plot(density(nc_growth))
qqnorm(c_growth); qqline(c_growth, col = 2)
qqnorm(nc_growth); qqline(nc_growth, col = 2)
```

```
# We kijken of de groei normaal verdeeld is - de Shapiro-Wilk test geeft ons duidelijke resultaten.
shapiro.test(c_growth) # Wel normaal verdeeld
shapiro.test(nc_growth) # Wel normaal verdeeld

# We mogen de t-test dus gebruiken.
t.test(c_growth,nc_growth)
# Er is dus een negatief effect op de toename in diameter als er competitie is.

# OPGAVE 1C

# We bepalen de associatie tussen de toenames van de hoogte en van de diameter
height_diameter <- cor(height_growth, diameter_growth)
cor.test(height_growth, diameter_growth)

# We plotten het verband
scatter.smooth(height_growth, diameter_growth, main="Verband tussen toenames in de hoogte en in de diameter", xlab="Toename in hoogte (inch)", ylab="Toename in diameter (inch)")

# OPGAVE 1D

# We zetten dit om naar een logische eenheid voor R, zodat R niet NC en NF beschouwt maar F en C.
# Als we zouden werken met een model dat NC en NF beschouwt, moeten we het effect van F en C inverteren en dat is verwarrend.
fertilizer_bool <- as.logical(fertilizer == "F")
competition_bool <- as.logical(competition == "C")

lmfit <- lm(height5~fertilizer_bool+competition_bool)
summary(lmfit)
plot(lmfit)
coef(lmfit)

# OPGAVE 1E

lmfit_height0 <- lm(height5~fertilizer_bool+competition_bool+height0)
summary(lmfit_height0)
plot(lmfit_height0)
coef(lmfit_height0)

# OPGAVE 1F

lmfit_height0$coeff %*% c(1,F,F,T)
lmfit_height0$coeff %*% c(1,T,F,T)
lmfit_height0$coeff %*% c(1,F,T,T)
lmfit_height0$coeff %*% c(1,T,T,T)

############
# OPGAVE 2 #
############

# Voorbereiding opgave 2

krabben <- read.csv("Krabben.csv",header = T)
str(krabben)
summary(krabben)
```

```
# colnames(krabben) <- tolower(colnames(krabben)) # Ze zijn al lowercase.
attach(krabben)
library(MASS)

# OPGAVE 2A

satellites_percent <- (length(satell[satell=='TRUE']) / length(satell))*100
satellites_percent

# OPGAVE 2B

# Grootte van het schild in functie van aanwezigheid van satellieten berekenen
t_width <- width[satell == T]
f_width <- width[satell == F]

# Boxplots
boxplot(width~satell, main="Schildgrootte bij aanwezigheid satellieten", xlab="Aanwezigheid satellieten", ylab="Schildgrootte (cm)", xaxt="n")
axis(side=1, at=1:2, labels=c("Geen satelliet","Minstens één satelliet"))

# We kijken naar de verdeling van beide groepen en zien of ze normaal verdeeld zijn
plot(density(t_width))
plot(density(f_width))
qqnorm(t_width); qqline(t_width, col = 2)
qqnorm(f_width); qqline(f_width, col = 2)

# We kijken of de groepen normaal verdeeld zijn - de Shapiro-Wilk test geeft ons duidelijke resultaten.
shapiro.test(t_width) # Wel normaal verdeeld
shapiro.test(f_width) # Wel normaal verdeeld

# We mogen de t-test dus gebruiken.
t.test(t_width,f_width)
# Er is dus een duidelijk verband tussen de schildgrootte en de aanwezigheid van satellieten.

# OPGAVE 2C

krabben <- krabben[,c(1,3,2)] # kolom 2 en 3 switchen
index <- sample(c(rep("training", 130), rep("test", 43))) # test: training/3
krabben_train <- krabben[index == "training",]
krabben_test <- krabben[index == "test",]

# Trainen van modellen
lda_train <- lda(satell ~ ., data = krabben_train)
qda_train <- qda(satell ~ ., data = krabben_train)

# Validatie (functie uit practicum 10)
K <- 5 #aantal folds

own.cv <- function(x,y, K = 5,method = lda){
  f <- method
  n <- nrow(x)

```

```r
  grid <- rep(1:K, n%/%K+1 )[1:n]
  id <- sample(grid)
  preds <- rep(NA,n)
  for(i in 1:K){
    f.model <- f( x[id != i,],y[id != i])
    preds[id == i]<-  predict(f.model,newdata = x[id == i,])$class
  }
  preds-1
}

# Schijnbaar foutenpercentage lineaire discriminantanalyse
preds.cvlda<-own.cv(krabben_train[,1:2], krabben_train[,3], K = K, method = lda)
sum(krabben_train$satell != preds.cvlda)/nrow(krabben_train) #cross-validation error
table(preds.cvlda,krabben_train$satell)/130

# Schijnbaar foutenpercentage kwadratische discriminantanalyse
preds.cvqda<-own.cv(krabben_train[,1:2],krabben_train[,3],K = K,method = qda)
sum(krabben_train$satell != preds.cvqda)/nrow(krabben_train) #cross-validation error
table(preds.cvqda,krabben_train$satell)/130

# QDA > LDA, maar dit hangt af van de seed.
# Predictiefout
sum(predict(qda_train,newdata = krabben_test)$class != krabben_test$satell)/43
```