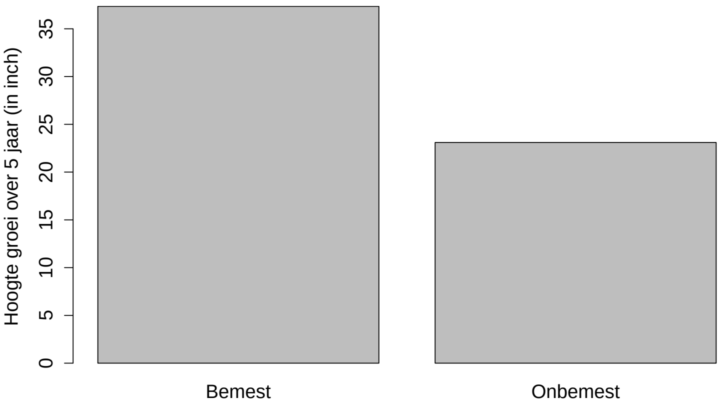
Statistische gegevensanalyse
Project

Titouan Vervack
Bachelor of Science in Informatics

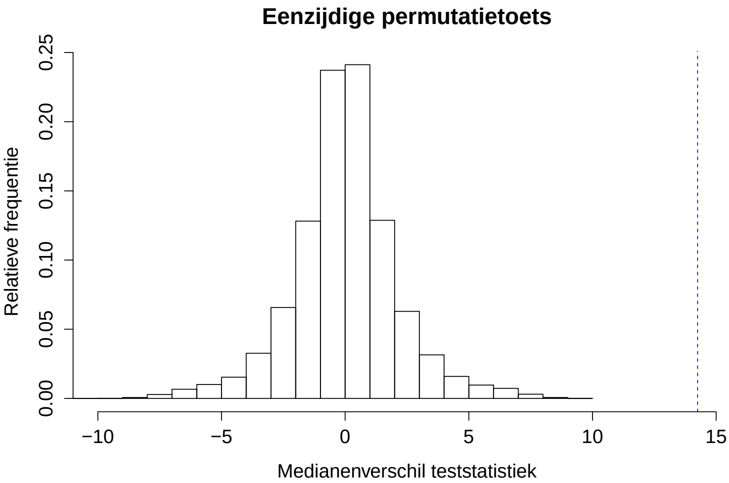
31 mei 2014

Opgave 1

a



Figuur 1: Hoogte groei met bemesting of zonder bemesting

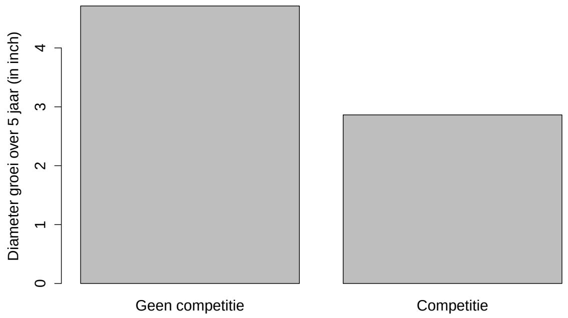


Figuur 2: Permutatieverdeling voor de hoogte groei van (on)bemeste bomen

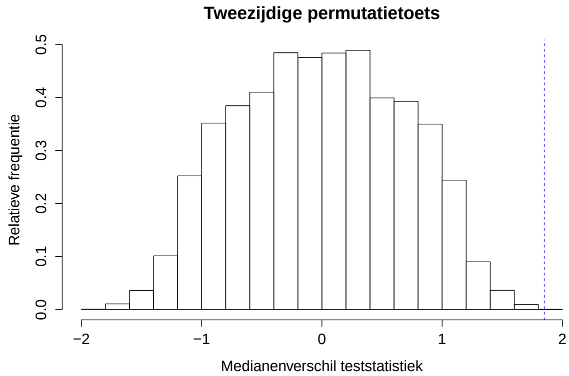
Om de groei in hoogte van de bomen te bekijken trekken we **Height0** af van **Height5**. Aangezien gevraagd wordt wat de invloed is van bemesting op deze hoogte, gebruiken we dit in combinatie met **Fertilizer**. Om een beeld te krijgen van de data maken we er histogrammen van, we kiezen histogrammen omdat de hoogtes continue variabelen zijn. We kunnen zien dat de data niet symmetrisch is en pieken bevat. We kiezen daarom om de mediaan te gebruiken in plaats van het gemiddelde. In figuur 1 zien we de staafdiagram die deze medianen voorstelt, we zien dat bemesting een positief effect heeft op de hoogte groei van bomen. Om zeker te zijn dat dit effect wordt veroorzaakt door de bemesting voeren we een permutatietoets uit, dit is een eenzijdige toets aangezien gevraagd wordt of de bomen sneller groeien bij bemesting. H_0 stelt dat de groei bij geen of wel bemesting dezelfde is. H_A stelt dat de groei bij bemesting positief is. De p-waarde, verkregen uit de permutatietoets, is 10^{-5}. Dit is een overtuigend bewijs tegen de nulhypothese waardoor we deze mogen verwerpen ten voordele van de alternatieve hypothese. Bemesting heeft dus een positief effect op de hoogte groei van bomen. De permutatieverdeling op basis van $M = 99999$ is te zien in figuur 2.

b

Bij deze opgave gebruiken we het verschil tussen **Diameter5** en **Diameter0**, we gebruiken ook **Competition**. We maken opnieuw histogrammen en beslissen om dezelfde reden terug de mediaan te gebruiken. De staafdiagram die de medianen vergelijkt is te zien in figuur 3a. Geen competitie blijkt een positief effect te hebben op de diameter van de bomen. We voeren ook hier een permutatietoets uit, dit is een tweezijdige aangezien het gevraagde effect niet gespecificeerd is. Om deze reden vermenigvuldigen we onze p-waarde met een factor 2. H_0 stelt dat de groei bij geen of wel competitie gelijk is. H_A stelt dat de groei bij geen of wel competitie niet gelijk is. De nulhypothese wordt verworpen door een terug lage p-waarde ten voordele van de alternatieve hypothese. Competitie verwijderen is dus positief voor de diameter van de bomen. De permutatieverdeling op basis van $M = 99999$ is te zien in figuur 3b.



(a) Diameter groei met competitie of zonder competitie

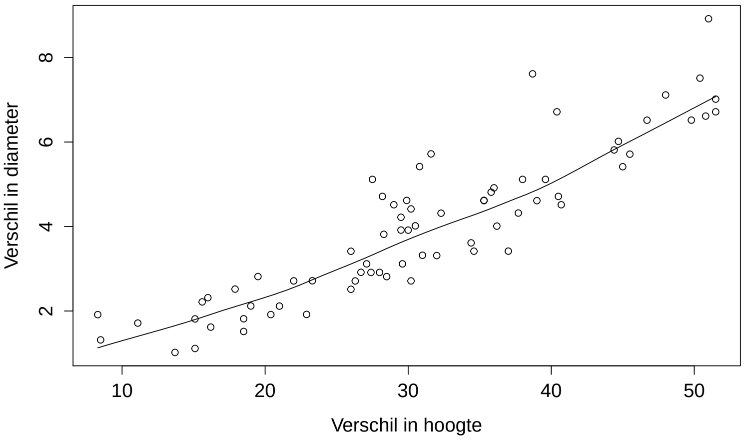


(b) Permutatieverdeling voor de diameter groei bij bomen met en zonder competitie

Figuur 3: Barplot en permutatieverdeling van vraag 1b

c

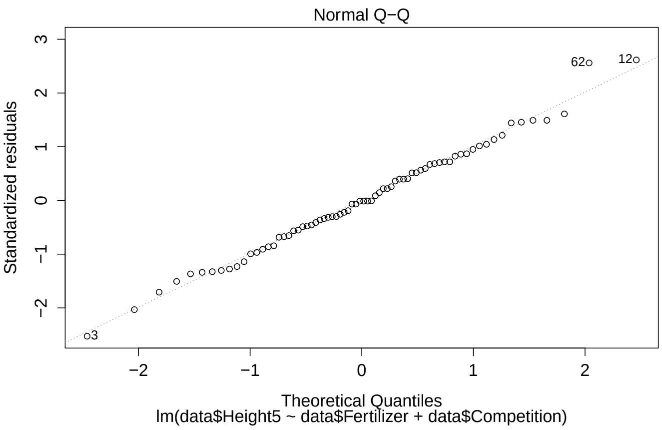
Hier gaan we de associatie tussen de verschillen uit opgaves **a** en **b** bekijken. Aangezien deze beide continue variabelen zijn drukken we de associatie uit aan de hand van de covariantie, deze is 17.82. We kunnen ook gemakkelijk de correlatie berekenen, deze is 0.90. Aan de hand van deze twee waarden kunnen we beslissen dat de associatie groot is. Aangezien we echter niet weten hoe representatief onze steekproef is kunnen we de associatie niet precies bepalen. We kunnen dit ook visualiseren in een scatterplot met een smoother curve, deze is te zien in figuur 4. Deze curve lijkt op een rechte, wat wil zeggen dat er een lineair verband is tussen de toename van de lengte en deze van de diameter. In combinatie met de hoge waarde voor de correlatie kunnen we zeggen dat er een sterk linear verband is tussen de twee variabelen. Dit is echter data voor onze steekproef en niet voor de populatie.



Figuur 4: Lineair verband tussen toename in

d

We stellen hiervoor een lineair regressie model op. Uit de summary van het model zien we dat het niet bemesten een negatieve invloed (schatting van -14.772) heeft op de groei en dat het afwezig zijn van competitie een positieve invloed (schatting van 10.917) heeft.



Figuur 5: Symmetrische verdeling van het model

We zien op de Q-Q plot (figuur 5) dat er bijna geen afwijking van normaliteit is. In combinatie met de determinatie coëfficiënt, die 0.66 is, kunnen we zeggen dat het model dus redelijk kwaliteitsvol is. De p-waarde is extreem laag, waardoor we kunnen beslissen dat de schattingen zeker goed zijn.

e

Hierbij trekken we gewoon **Height0** van **Height5** af. We bekomen terug dat niet bemesten negatief is en dat de afwezigheid van competitie positief is voor de groei van de bomen. De determinatie coëfficiënt is 0.67, net iets hoger dus het model is net iets beter.

f

We zien een schatting voor de gemiddelde hoogtes voor de vier variabelen in de tabel hieronder.

	Bemest	Onbemest
Competitie	33.06	18.35
Geen competitie	43.52	28.81

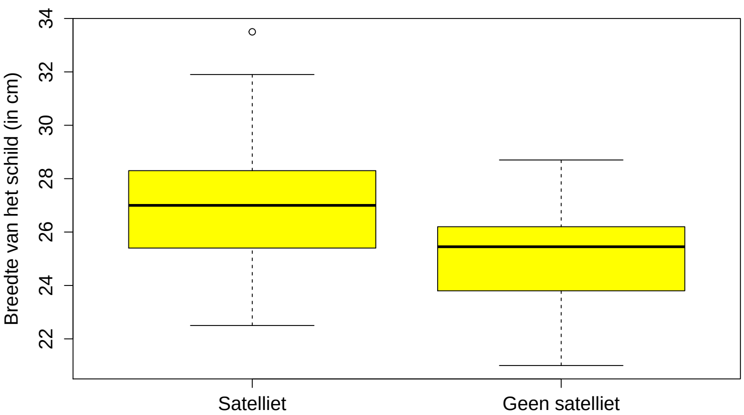
Opgave 2

a

Bij deze opgave is het belangrijk op te merken dat er gevraagd wordt naar een percentage binnen de populatie, het is dus niet voldoende het percentage uit te rekenen voor de steekproef. Om dit percentage te berekenen maken we gebruik van een 95%betrouwbaarheidsinterval, we doen dit aan de hand van de formule in de cursus. We bekomen dat dit interval $[0.5676, 0.7092]$ is. We kunnen dus met 95% zekerheid stellen dat 56.76% tot 70.92% van de vrouwelijke degenkrabben in de de populatie over minstens één satelliet beschikt tijdens de paringsperiode. Aangezien het percentage aan vrouwelijke krabben met een satelliet in onze steekproef 64.16% is, is dit een representatieve steekproef.

b

De aanwezigheid van satellieten hangt inderdaad af van de grootte van het schild. Dit is duidelijk te zien in de boxplot op figuur 6. Om te weten in hoeverre deze verschilt doen we de t-test van Welch. Deze heeft als H_0 dat beide gemiddeldes (de gemiddelde van de schildgrootte van vrouwtjes met en zonder satellieten) gelijk zijn. H_A stelt dan dat de gemiddeldes niet gelijk zijn. De p-waarde is $9.495 * 10^{-9}$, dit levert een overtuigend bewijs tegen de nulhypothese.



Figuur 6: Verschil in grootte van het schild bij vrouwtjes met en zonder satelliet

Hierdoor kunnen we de nulhypothese verwerpen ten voordele van de alternatieve hypothese. Het betrouwbaarheidsinterval dat we bekomen met de t-test is $[1.19, 2.33]$. Vrouwtjes met satellieten hebben dus een schild dat 1.19 tot 2.33 cm groter is dan vrouwtjes zonder satellieten.

c

Voor dit model op te stellen gebruiken we een lineaire en een kwadratische discriminant analyse. We vergelijken dan welke de beste is. We gebruiken ook leave-one-out cross-validation. Aan de hand van de tabellen kunnen we zien dat er bij de lineaire analyse met cross-validation een predictiefout is van 29.48% en bij de kwadratische met cross-validation een van 31.79%. Voor de lineaire zonder cross-validation vinden we een schijnbare predictiefout van 28.90% en voor de kwadratische zonder cross-validation 29.48%. De predictiefouten bij de schattingen met cross-validation liggen hoger dan die zonder, ze overfitten dus de training dataset wat leidt tot een instabielere classificatieregel die minder betrouwbare predicties oplevert. Op basis van de leave-one-out cross-validation schattingen van de predictiefout kunnen we besluiten dat de lineaire discriminant analyse stabielere resultaten oplevert dan de kwadratische.

	FALSE	TRUE
FALSE	27	15
TRUE	35	96

Tabel 1: Misclassificatietabel voor lineaire discriminant analyse

	FALSE	TRUE
FALSE	30	19
TRUE	32	92

Tabel 2: Misclassificatietabel voor lineaire discriminant analyse

	FALSE	TRUE
FALSE	26	15
TRUE	36	96

Tabel 3: Leave-one-out cross-validation misclassificatietabel voor lineaire discriminant analyse

	FALSE	TRUE
FALSE	28	21
TRUE	34	90

Tabel 4: Leave-one-out cross-validation misclassificatietabel voor kwadratische discriminant analyse

Appendix

Hieronder vindt u de gebruikte R code met extra informatie in commentaar.

```
# 1)
################################################################################
data <- read.csv("ZwarteSpar.csv", header = T)
################################################################################


# a)
##########################################################################
# Bereken de groei over de laatste 5 jaar
dataF = subset(data, Fertilizer == "F")
FHeightDiff = dataF$Height5 - dataF$Height0
dataNF = subset(data, Fertilizer == "NF")
NFHeightDiff = dataNF$Height5 - dataNF$Height0

# Aangezien de groei in lengte een continue variabele is maken
# we hier een histogram van
hist(FHeightDiff, xlab = "Groei (in inch)", ylab = "Frequentie")
hist(NFHeightDiff, xlab = "Groei (in inch)", ylab = "Frequentie")
# Deze zijn niet symmetrisch en hebben pieken dus is het beter
# om hier de mediaan te gebruiken dan het gemiddelde

medF = median(FHeightDiff)
medNF = median(NFHeightDiff)

barplot(c(medF, medNF),
        names.arg = c("Bemest", "Onbemest"),
        ylab = "Hoogte groei over 5 jaar (in inch)", xpd = FALSE)
# We zien in de staafdiagram dat bemeste bomen een grotere groei
# vertonen dan onbemeste bomen. Het verschil is groot, meer dan
# 10 inch. Dit betekend dat de bomen sneller groeien als ze bemest
# worden.

# Om zeker te zijn dat dit niet beinvloed wordt door een derde
# variabele doen we een permutatietoets
# Dit is een eenzijdige toets aangezien we enkel geinteresseerd
# zijn in de toename

M <- 99999
meddiffs <- numeric()
v <- c(FHeightDiff, NFHeightDiff)
for(i in 1:M){
  Fperm <- sample(1:(length(FHeightDiff) + length(NFHeightDiff)),
                  length(FHeightDiff))
  meddiffs[i] <- median(v[Fperm]) - median(v[-Fperm])
}
hist(meddiffs, prob = T, xlab = "Medianenverschil teststatistiek",
     ylab = "Relatieve frequentie", main = "Eenzijdige permutatietoets",
     xlim = c(-10, 15))

abline(v = medF-medNF, col="blue", lty = 2)
# Bereken de p waarde
p<-(sum(meddiffs >= (medF-medNF))+1)/(M+1)
p
# Deze is 10^-5, de steekproef levert dus een overtuigend bewijs tegen
```

```
# de nulhypothese van een gelijke hoogte groei
##############################################################################

# b)
##############################################################################
# Bereken de groei over de laatste 5 jaar
dataC = subset(data, Competition == "C")
CDiagDiff = dataC$Diameter5 - dataC$Diameter0
dataNC = subset(data, Competition == "NC")
NCDiagDiff = dataNC$Diameter5 - dataNC$Diameter0

# Aangezien de groei in lengte een continue variabele is maken
# we hier een histogram van
hist(CDiagDiff, xlab = "Diameter groei (in inch)", ylab = "Frequentie")
hist(NCDiagDiff, xlab = "Diameter groei (in inch)", ylab = "Frequentie")
# We zien dat de piek veel verkleint eens er geen competitie is wat de
# verdeling consistenter maakt.

medC = median(CDiagDiff)
medNC = median(NCDiagDiff)

barplot(c(medNC, medC),
        names.arg = c("Geen competitie", "Competitie"),
        ylab = "Diameter groei over 5 jaar (in inch)", xpd = FALSE)
# We zien dat er zonder competitie een veel groter groei in diameter
# is, meer dan 1 inch. Dit wil zeggen dat bomen zonder competitie
# sneller groeien in diameter.

# Om zeker te zijn dat dit niet beinvloed wordt door een derde
# variabele doen we een permutatietoets
# Dit is een tweezijdige toets aangezien er niet gespecifieerd
# is wat het effect moet zijn
M <- 99999
meddiffs <- double()
v <- c(NCDiagDiff, CDiagDiff)
for(i in 1:M){
  Cperm <- sample(1:(length(CDiagDiff) + length(NCDiagDiff)), length(CDiagDiff))
  meddiffs[i] <- median(v[Cperm]) - median(v[-Cperm])
}
hist(meddiffs, prob = T, xlab = "Medianenverschil teststatistiek",
     ylab = "Relatieve frequentie", main = "Tweezijdige permutatietoets")

abline(v = medNC-medC, col = 'blue', lty = 2)
# Bereken de p waarde, vermenigvuldig met 2 aangezien het
# een tweezijdige permutatietoets is
p<-2*((sum(meddiffs >= (medNC-medC))+1)/(M+1))
p

# Hoeweel de p-waarde deze keer iets hoger ligt is dit nog altijd
# een lage waarde. Dit zorgt er dus voor dat de steekproef een
# overtuigend bewijs vormt tegen de nulhypothese van een gelijke
# toename in diameter
##############################################################################

# c)
##############################################################################
# Bereken de toenames
heightDiff = data$Height5 - data$Height0
diagDiff = data$Diameter5 - data$Diameter0
```

```
# Dit zijn continue variabelen dus gebruiken we covariantie om de
# associatie te modelleren
cov(heightDiff, diagDiff)
# Bereken toch de correlatie om makkelijker te vergelijken
cor(heightDiff, diagDiff)
# Plot dit in een scatterplot met een smoother
scatter.smooth(heightDiff, diagDiff, xlab = "Verschil in hoogte",
     ylab = "Verschil in diameter")
# We zien dat de smoother lijkt op een rechte, er is dus een lineair
# verband tussen de toename van de lengte en deze van de diameter.
# Aangezien de correlatie 0.90 is, is dit ook een sterk verband
##########################################################################

# d)
##########################################################################
# Stel het model op, Height5 hangt af van Fertilizer en Competition
model <- lm(data$Height5 ~ data$Fertilizer + data$Competition)
# QQplot p179
plot(model)
summary(model)
# We zien dat niet bemesten een negatief effect heeft en dat de
# afwezigheid van competitie een positief effect heeft op de groei.
# We zien ook dat de p-waarde extreem laag is (< 2.2e-16),
# de schattingen zijn dus zeker goed
# De determinatie co ffici nt is 0.66 dus ook het model is redelijk
# goed.
##########################################################################

# e)
##########################################################################
# Breng de hoogte bij start in rekening
model <- lm(data$Height5 - data$Height0 ~ data$Fertilizer + data$Competition)
plot(model)
summary(model)
# De determinatie co ffici nt is hier net iets hoger wat dit model
# beter maakt.
##########################################################################

# f)
##########################################################################
# Print de inschatting voor de gemiddeldes voor de vier combinaties
print(paste("Fertilized + Competition: ", model$coeff %*% c(1,F,F)))
print(paste("Fertilized + No competition: ", model$coeff %*% c(1,F,T)))
print(paste("Not fertilized + Competition: ", model$coef %*% c(1,T,F)))
print(paste("Not fertilized + No competition: ", model$coeff %*% c(1,T,T)))
##########################################################################




# 2)
##########################################################################
# Clear the environment
closeAllConnections()
rm(list=ls())
# Read the new data
data <- read.csv("Krabben.csv", header = T)
##########################################################################

```

```
# a)
################################################################################
# Splits de data
dataS = subset(data, satell == "TRUE")
dataNS = subset(data, satell == "FALSE")
# Bepaal het percent vrouwelijke krabben dat een satelliet heeft
# binnen onze steekproef
dataSPercent = length(dataS$satell)/length(data$satell)
# Bepaal de grenzen van het 95% betrouwbaarheidsinterval
noemer = length(data$satell) + 4
tellerp = (length(dataS$satell) + 2)/noemer
bovengrens = tellerp + 1.96 * sqrt((tellerp*(1-tellerp))/noemer)
ondergrens = tellerp - 1.96 * sqrt((tellerp*(1-tellerp))/noemer)
ondergrens; bovengrens
################################################################################

# b)
################################################################################
# Visualiseer de data in histogrammen
hist(dataS$width)
hist(dataNS$width)
# De data is niet symmetrisch dus we gebruiken de mediaan

barplot(c(median(dataS$width), median(dataNS$width)),
        names.arg = c("Satelliet", "Geen satelliet"),
        ylab = "Breedte van het schild (in cm)", xpd = FALSE)
# De schildgrootte ligt iets hoger bij de vrouwtjes die over een
# satelliet beschikken. Het verschil is echter heel klein, we wensen
# meer informatie. Deze kan geleverd worden door een boxplot

boxplot(dataS$width, dataNS$width, col="yellow",
        ylab = "Breedte van het schild (in cm)",
        names = c("Satelliet", "Geen satelliet"))
# De schildgrootte heeft dus duidelijk een effect op het al dan niet
# beschikken over een satelliet of niet.

# Nu willen we weten hoezeer de breedte van het schild verschilt
# We voeren de t-test van Welch uit
t.test(dataS$width, dataNS$width)
################################################################################

# c)
################################################################################
# Importeer de bibliotheek die nodig is voor discriminant analyses
library(MASS)
# Voer lineaire en kwadratische discriminant analyses uit
lda = lda(satell ~ ., data)
qda = qda(satell ~ ., data)
# Schatten
plda = predict(lda)
pqda = predict(qda)

# Toon de misclassificatietabellen
lt = table(plda$class, data$satell)
qt = table(pqda$class, data$satell)

(lt[2] + lt[3]) / sum(lt)
# Bereken de kwadratische predictiefout
(qt[2] + qt[3]) / sum(qt)
```

```r

# Doe een lineaire discriminant analyse op de satelliet data, met
# leave-one-out cross-validation
lda = lda(satell ~ ., data, CV=TRUE)
# Doe een kwadratische discriminant analyse op de satelliet data, met
# leave-one-out cross-validation
qda = qda(satell ~ ., data, CV=TRUE)

# Bereken de leave-one-out cross-validation misclassificatietabel voor
# lineaire discriminant analyse
lt = table(lda$class, data$satell)
# Bereken de leave-one-out cross-validation misclassificatietabel voor
# kwadratische discriminant analyse
qt = table(qda$class, data$satell)
lt
qt

# Bereken de lineaire predictiefout
(lt[2] + lt[3]) / sum(lt)
# Bereken de kwadratische predictiefout
(qt[2] + qt[3]) / sum(qt)
################################################################################
```

Oplossingen.R