# G0N11C Statistiek & data-analyse
# Project eerste zittijd 2015-2016

**Naam 1:** Mathias Van Herreweghe

**Studierichting 1:** bachelor informatica

**Groepsnummer:** 067

## Gegevens

De gegevens die de basis vormen voor dit project bestaat voor het eerste gedeelte uit de BMI-waarden van vrouwen in Afrika uit 2007 en 2008 . Voor het tweede gedeelte gebruiken we de BMI-waarden van mannen en vrouwen in Malaysia van 1980 tot en met 2008. Verder gebruiken we het significantieniveau $\alpha = 0.05$.

# 1 Opgave 1

Ga na of er een stijging is in het gemiddelde BMI bij vrouwen. Vergelijk hiervoor de BMI-waarden van de twee desbetreffende jaren (2007 en 2008) voor een bepaald continent (Afrika).

## 1.1 Hypothesetest

Allereerst valt het op dat de gegevens gepaarde kwantitatieve variabelen zijn. We zullen dus proberen gebruik maken van de toevalsvariabele $V = X - Y$, met $v_i = x_i - y_i$ met $i = 1, ..., n$. Dit doen we in de veronderstelling dat de verschilvariabele $V$ normaal verdeeld is, we controleren dit bij subsectie 1.2. We nemen X voor de BMI-waarden in 2008 en Y voor de BMI-waarden in 2007, beide worden enkel beschouwd voor het continent Afrika (aangeduid met waarde 4 onder *Continent* in de gegevensset).

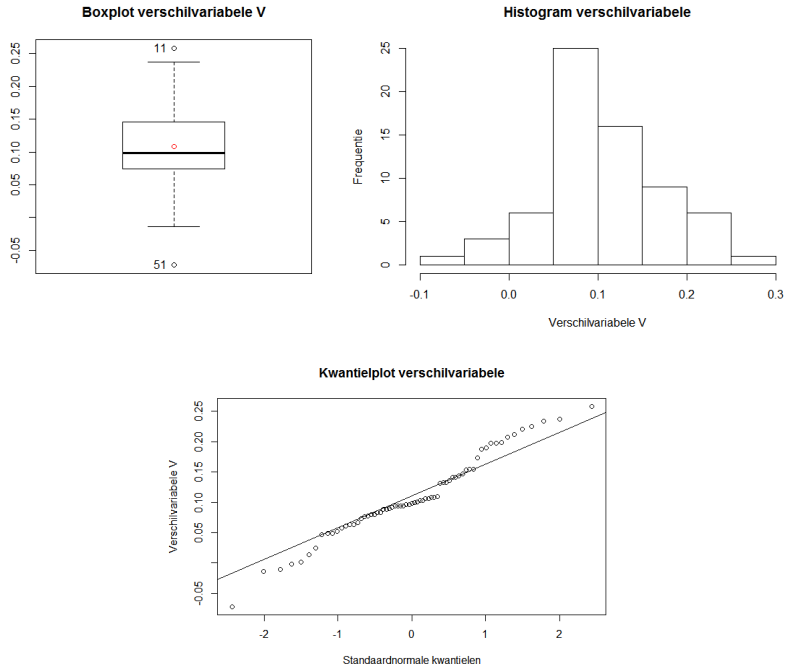
Aangezien we willen nagaan of er een stijging is gebruiken we de rechtséénzijdige test. De hypothesen zijn dan als volgt.

$$\begin{aligned} &\mathrm{H}_0\colon\ \mu_v \leqslant 0 \\ &\mathrm{H}_1\colon\ \mu_v > 0 \end{aligned}$$

## 1.2 Controle voorwaarden

We zullen nu controleren of dat de verschilvariabele $V$ een normale verdeling heeft.

**Grafisch**




Figuur 1: Een boxplot, histogram en kwantielplot van de verschilvariabele V

De boxplot laat zien dat er 2 waarnemingen buiten de snorharen liggen, dit zouden uitschieters kunnen zijn. Verder zijn de snorharen even lang en ligt het gemiddelde (rode cirkel) dicht bij de mediaan, dit wijst op een normale verdeling. Ook merken we op dat de mediaan iets lager ligt dan het midden, dit kan wijzen op een (licht) rechtsscheve verdeling.

Het histogram laat een benaderende klok-curve zien, dit wijst op een normale verdeling. De rechterstaart is een beetje langer, dit kan opnieuw wijzen op een (licht) rechtsscheve verdeling.

Vervolgens laat de kwantielplot benaderend een rechte zien, dit wijst opnieuw in de richting van een normale verdeling. Bij de kwantielplot merken we wederom twee afwijkende waarnemingen op die uitschieters zouden kunnen zijn.

**Formeel**

We gaan testen of de verdeling normaal is aan de hand van de Shapiro-Wilk test. We gebruiken de volgende hypothesen hiervoor:

$H_0$: De gegevens komen uit een normale verdeling.
$H_1$: De gegevens komen niet uit een normale verdeling.

De test-waarde $w = 0.97497$, $P\text{-}waarde = 0.1946$ en aangezien $P\text{-}waarde > \alpha$, kunnen we $H_0$ niet verwerpen en bevestigen we dat de verschilvariabele $V$ normaal verdeeld is onder significantieniveau $\alpha = 0.05$.

We nemen verder aan dat waarneming 11 en 51 redelijk sterk verschillen van de andere waarnemingen, maar niet genoeg om te besluiten dat het uitschieters zijn.

### 1.3 Uitvoering hypothesetesten

Als eerste berekenen we de testwaarde aan de hand van de teststatistiek

$$T = \frac{\overline{V}}{S_v/\sqrt{n}} \rightarrow t = 13.391 \text{ met } S_{\text{v}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \overline{V})^2.$$

Deze testwaarde levert ons de P-waarde

$$P(T > t) = P(T > 13.391) = 2.2 \times 10^{-16} \approx 0 \text{ met } T \sim t_{n-1} = t_{66}.$$

De P-waarde is veel kleiner dan $\alpha$. We kunnen $\text{H}_0$ dus verwerpen en dit bevestigt $\text{H}_1$ op significantieniveau $\alpha = 0.05$. Als we het 100(1-$\alpha$)%-betrouwbaarheidsinterval maken voor $\mu_v = E(X) - E(Y)$ dan krijgen we:

$$\begin{gathered}[\ \overline{v} - t_{n-1,\alpha/2}\tfrac{s_v}{\sqrt{n}}\ ,\ +\infty\ [ \\ \longrightarrow [\ 0.0951823,\ +\infty\ [\end{gathered}$$

Zoals verwacht na het berekenen van de P-waarde, is 0 geen element van dit interval en kan men opnieuw $\text{H}_0$ verwerpen op significantieniveau $\alpha = 0.05$.

### 1.4 Besluit

Aan de hand van P-waarde $2.2 \times 10^{-16}$ kunnen we $\text{H}_0$ verwerpen op significantieniveau $\alpha = 0.05$. We kunnen besluiten dat $\mu_v > 0$, met andere woorden is de gemiddelde BMI-waarde van vrouwen in het continent Afrika significant gestegen tussen 2007 en 2008.

## 2 Opgave 2

Een BMI-waarde is hoog als deze groter is dan de grenswaarde (die verschilt van land tot land), deze is laag als deze kleiner dan is of gelijk aan de grenswaarde is. Malaysia heeft een grenswaarde van *23.1*. Is er een verschil in proportie 'hoge BMI' waarde bij mannen en vrouwen voor het land Malaysia (voor alle beschikbare jaren)?

### 2.1 Hypothesetest

We willen nagaan of er een verschil is tussen de proporties van hoge BMI-waarden bij mannen en vrouwen in Malaysia. Om dit te staven gebruiken we dus de tweezijdige test. De voorwaarde hiervoor is dat de gegevens onafhankelijk van elkaar zijn, dit wordt gecontroleerd in subsectie 2.2. We gebruiken index 1 voor gegevens van de mannen en index 2 voor vrouwen. De hypothesen zijn dan als volgt:

$$\begin{aligned}&\text{H}_0\text{: } p_1 = p_2 \\ &\text{H}_1\text{: } p_1 \neq p_2\end{aligned}$$

## 2.2 Controle voorwaarden

Om de voorwaarde van onafhankelijkheid na te gaan kunnen we volgende hypothesetesten gebruiken:

$H_0$: Het hoog of laag zijn van de BMI-waarde is onafhankelijk van het geslacht
$H_1$: Het hoog of laag zijn van de BMI-waarde is afhankelijk van het geslacht

Bij volledige onafhankelijkheid kunnen we bepaalde waarden verwachten, zo is de

$$\text{verwachte waarde} = \frac{\text{rijtotaal} * \text{kolomtotaal}}{\text{n}}$$

Aan de hand hiervan bekomt men de tabel met de verwachte waarden.

| | Hoog BMI | Laag BMI | |
|---|---|---|---|
| Mannen | 17.5 | 11.5 | 29 |
| Vrouwen | 17.5 | 11.5 | 29 |
| | 35 | 23 | 58 |

Tabel 1: verwachte waarden bij onafhankelijkheid

De geobserveerde waarden zijn zoals volgt.

| | Hoog BMI | Laag BMI | |
|---|---|---|---|
| Mannen | 14 | 15 | 29 |
| Vrouwen | 21 | 8 | 29 |
| | 35 | 23 | 58 |

Tabel 2: geobserveerde waarden

Deze geobserveerde waarden verschillen van de verwachte waarden maar op het eerste zicht lijkt het verschil niet significant.

Om de onafhankelijkheid formeler aan te tonen berekenen we als eerste de testwaarde. De chi-kwadraat($X^2$)-teststatistiek is gedefinieerd als

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(n_{ij} - n f_{i+} f_{+j})^2}{n f_{i+} f_{+j}} \text{ waarbij } X^2 \approx_{H_0} \chi^2_{\text{v}} \text{ met } v = (r-1)(k-1),$$
waarbij $n_{ij}$ de geobserveerde waarde is en $n f_{i+} f_{+j}$ de verwachte waarde.

Dit geeft ons de testwaarde $\chi^2 = 2.5938$. Hiermee berekenen we de P-waarde

$$P(X^2 > \chi^2) = P(X^2 > 2.5938) = 0.1073 \text{ met } X^2 \sim \chi^2_1.$$

De P-waarde is groter dan $\alpha$ en bijgevolg kunnen we $H_0$ niet verwerpen op significantieniveau $\alpha = 0.05$. We nemen dus aan dat de gegevens onafhankelijk zijn van elkaar en dus is aan de voorwaarde voldaan.

Bijkomende voorwaarden om de testwaarde te kunnen gebruiken, die gebruik maakt van een gewogen gemiddelde van $\hat{P}_1$ en $\hat{P}_2$ zijn:

$$n_1\hat{p_1} \geqslant 5,\ n_1(1-\hat{p_1}) \geqslant 5 \rightarrow 14 \geqslant 5, 15 \geqslant 5$$
$$n_2\hat{p_2} \geqslant 5,\ n_2(1-\hat{p_2}) \geqslant 5 \rightarrow 21 \geqslant 5, 8 \geqslant 5$$

Deze voorwaarden zijn ook voldaan.

## 2.3 Uitvoering hypothesetesten

Als eerste berekenen we de testwaarde aan de hand van de teststatistiek

$$Z = \frac{\hat{P_1}-\hat{P_2}}{\sqrt{\hat{P_0}(1-\hat{P_0})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \rightarrow z = -1.878945 \text{ met } Z \approx N(0,1).$$

Aan de hand van deze testwaarde bekomen we de P-waarde

$$2P(Z > |z|) = 2P(Z > 1.878945) = 0.06025$$

De P-waarde is groter dan $\alpha$. We kunnen $H_0$ dus niet verwerpen op significantieniveau $\alpha = 0.05$.

Het 100(1-$\alpha$)%-betrouwbaarheidsinterval wordt gegeven door

$$[\ \hat{p_1} - \hat{p_2} - z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}},\ \hat{p_1} - \hat{p_2} + z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\ ]$$
$$\longrightarrow [\ -0.485383337,\ 0.002624717\ ]$$

Wederom zoals verwacht is 0 een element in dit interval en bijgevolg valt $H_0$ niet te verwerpen op significantieniveau $\alpha = 0.05$.

## 2.4 Besluit

Aan de hand van de gevonden P-waarde van 0.06025 kunnen we de nulhypothese niet verwerpen en behouden we dus $p_1 = p_2$, we kunnen dus besluiten dat er geen significant verschil is in de proporties tussen mannen en vrouwen omtrent hoge BMI-waarden, in Malaysia, gemeten van 1980 tot en met 2008.

# Bijlagen

Listing 1: R script

```
BMIfemale=read.csv(file=file.choose(),header=TRUE,dec=",",sep=";")
BMImale=read.csv(file=file.choose(),header=TRUE,dec=",",sep=";")
AfrikaBMIfemale = BMIfemale[BMIfemale$Continent == 4,]
VerschilAfrika=AfrikaBMIfemale$X2008-AfrikaBMIfemale$X2007
Boxplot(VerschilAfrika,xlab="",ylab="",main="Boxplot verschilvariabele V")
points(1, mean(VerschilAfrika), col = "red")
qqnorm(VerschilAfrika, xlab="Standaardnormale kwantielen", ylab="Verschilvariabele V",
    main="Kwantielplot verschilvariabele")
qqline(VerschilAfrika)
hist(VerschilAfrika, xlab="Verschilvariabele V", ylab="Frequentie", main="Histogram
    verschilvariabele")
shapiro.test(VerschilAfrika)
t.test(AfrikaBMIfemale$X2008, AfrikaBMIfemale$X2007, paired=TRUE, alternative="greater")

MBMIfemale=BMIfemale[BMIfemale$Country == "Malaysia",]
MBMImale=BMImale[BMImale$Country == "Malaysia",]
grenswaarde=23.1
MBMImaleArray=as.numeric(MBMImale[1,])
MBMImaleArray=MBMImale[-c(1,2)]
n1=length(MBMImaleArray)
MBMImaleArray=MBMImaleArray[MBMImaleArray > grenswaarde]
MBMIfemaleArray=as.numeric(MBMIfemale[1,])
MBMIfemaleArray=MBMIfemale[-c(1,2)]
n2=length(MBMIfemaleArray)
MBMIfemaleArray=MBMIfemaleArray[MBMIfemaleArray > grenswaarde]
succesMBMImale=length(MBMImaleArray)
succesMBMIfemale=length(MBMIfemaleArray)
males=c(succesMBMImale,n1-succesMBMImale)
females=c(succesMBMIfemale,n2-succesMBMIfemale)
Ctable=rbind(males,females)
colnames(Ctable)=c("Hoog BMI","Laag BMI")
ChiSq <- chisq.test(Ctable)
ChiSq$expected
ChiSq$observed
ChiSq
prop.test(c(succesMBMImale,succesMBMIfemale),c(n1,n2), correct=FALSE)
p1=succesMBMImale/n1
p2=succesMBMIfemale/n2
p0=(n1*p1+n2*p2)/(n1+n2)
z=(p1-p2)/sqrt(p0*(1-p0)*(1/n1+1/n2))
```

Listing 2: R output

```
[1] 51 11

        Shapiro-Wilk normality test

data:  VerschilAfrika
W = 0.97497, p-value = 0.1946

> t.test(AfrikaBMIfemale$X2008, AfrikaBMIfemale$X2007, paired=TRUE, alternative="greater
    ")

        Paired t-test

data:  AfrikaBMIfemale$X2008 and AfrikaBMIfemale$X2007
t = 13.391, df = 66, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.0951823       Inf
sample estimates:
mean of the differences
              0.1087276

        Hoog BMI Laag BMI
males       17.5     11.5
females     17.5     11.5

        Hoog BMI Laag BMI
males         14       15
females       21        8


        Pearson's Chi-squared test with Yates' continuity correction

data:  Ctable
X-squared = 2.5938, df = 1, p-value = 0.1073


        2-sample test for equality of proportions without continuity correction

data:  c(succesMBMImale, succesMBMIfemale) out of c(n1, n2)
X-squared = 3.5304, df = 1, p-value = 0.06025
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.485383337  0.002624717
sample estimates:
   prop 1    prop 2
0.4827586 0.7241379

[1] -1.878945
```