

Colleges statistiek

Blok 2.6



2017-2018

LH

# College 01 (week 1) 15-11-2017

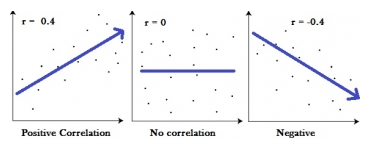
*Correlatie* = De statistische samenhang tussen 2 datasets. De sterkte van de

correlatie wordt uitgedrukt in de correlatiecoëfficiënt. Je gebruikt correlatie dus om te kijken of er een lineair verband is tussen 2 variabele.

Kunnen gebruikt worden om verbanden aan te tonen. Een correlatie hoeft echter niet per sé een verband te hebben.

*Correlatiecoëfficiënt* = De maat voor de correlatie tussen 2 variabele. Waarde zit

tussen -1 en +1. Dicht bij de 0 betekent dat er geen lineiare samenhang is. Hoe verder de waarde van de 0 afzit, hoe sterker de samenhang. Als je de waarde van x weet kun je y voorspellen.



R= Plusgetal = erbij = toename = lijn omhoog

R=0 = voor iedere x zelfde y = lijn constant

R=Mingetal = eraf = afname = lijn omlaag

* Of je een Cc-waarde veel vindt hangt

erg af van het vakgebied.

* Als je bepaalde data weglaat moet je

duidelijk aangeven waarom, daarnaast kan het effect van een uitbijter erg groot zijn.

Met dit soort verbanden moet je minstens 3 groepen hebben om een verband te kunnen aantonen.

**Correlatietest in R**

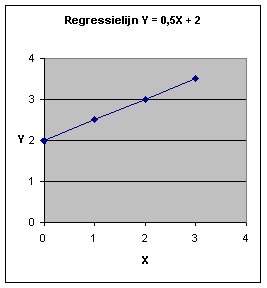
In R kun je correlatie testen met de functie cor(…, ….) of cor.test(. …,....). Je kijkt hiermee of er een lineair verband is tussen 2 variabelen. Met ?cor.test kan er in R gekeken worden welke opties er zijn en kan er meer informatie over de cor.test opgevraagd worden.

Het doel is zo dicht mogelijk bij -1 of +1 te komen want dan is er een lineair verband.

We gebruiken alleen de methode Pearson. Dit is tevens de default waarde van cor.test dus deze hoeft niet per sé extra toegevoegd te worden aan de functie.

VB: cor.test(X,Y, method=”pearson”) of cor.test(X,Y)

# College 02 (week 2) 22-11-2017

Verschil correlatie en regressie

*Correlatie* = lineaire samenhang (stijgende of dalende trend)

**ŷ**

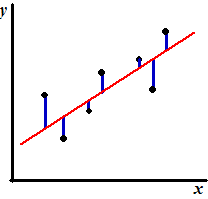
*Regressie* = welke lijn sluit het beste aan bij punten als de

correlatie groot genoeg is. Dus een samenhang die niet per sé lineair is.

*ŷ* = Schattingswaarde van y (zie onder voor formule)

**kleinste kwadraat methode**

Met deze methode wil je de kleinst mogelijke som van de residuen.



Methode: niet

wel

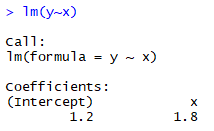
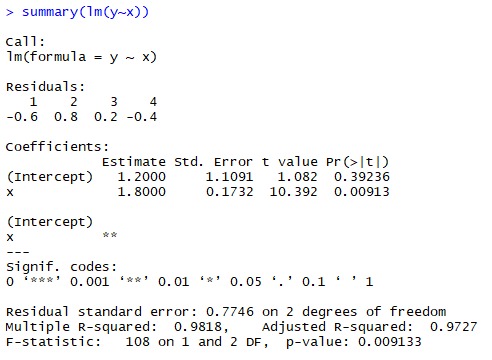
Je weet van een waarde niet of die boven of onder de lijn zit, daarom kwadraateer je de waarde. Mingetallen en plusgetallen komen na het kwadrateren namelijk op hetzelfde getal uit. De getallen na het kwadrateren zijn altijd positief.

*Residue* = Afstand daadwerkelijke punt tot trendlijn.

**Richtingscoëfficiënt**

*Richtingscoëfficiënt =*

Voer in R in: Of voor meer informatie:

Bij summary staan onder andere het aantal vrijheidsgraden (n-2) en de p-waarde genoemd. Als P-waarde klein is voer je de regressie terecht uit. Bij de residu is de Ŷ en Y waarde te vinden. De Multiple R squared is er ook te vinden.

Bij residuals kun je zien of bepaalde waarde boven of onder de trendlijn zitten. Negatieve waardes geven aan dat Ŷ boven Y ligt en positieve getallen geven aan dat Y juist onder Ŷ ligt.

Van de tweede rij getallen kun je de waarde bij de Ŷ waarde optellen. Hiermee krijg je de Y waarde.

Formule:

***Ŷ*** *= Bij veel sterren dan lijkt lijn op punten.*

# College 03 (week 3) 29-11-2017

Herhaling

Relatiecoëfficiënt moet tussen -1 en +1 zitten. Dicht bij de 0 wijst erop dat er GEEN lineair verband is. Hoe verder van de 0 hoe meer verband er is.

Voorbeeld hypotheses opstellen:

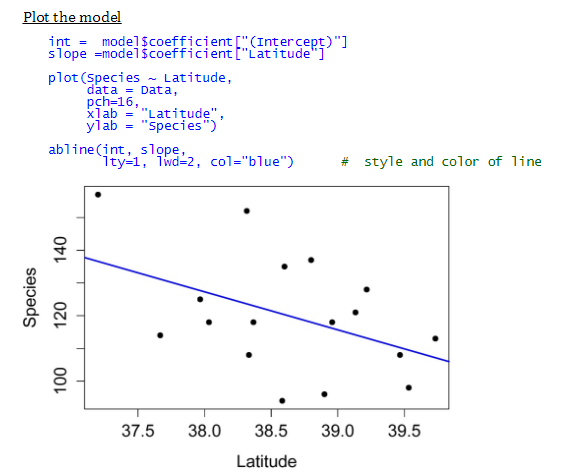
H0: Er is geen correlatie.

HA: Er is wel correlatie.

Vervolgens let je op p-waarde om te zien of hypothese stand houd.

Met cor.test() kun je de coeefiencten inzien maar met summary(lm(y~x)) krijg je meer informatie. Zo kun j ebijvoorbeeld de p-waarde achterhalen.

Met plot(y~x) kun je waardes visaliseren in een dotgrafiek. Plot kun je weer verschillende dingen meegeven (bijvoorbeeld titel, asnamen, dichte dots, lijn tekenen)



Plot 🡪 plotten dotgrafiek

Pch 🡪 dichte dots

Lwd 🡪 lijndikte

x/ylab 🡪 benoemen assen

abline 🡪 trendlijn trekken

BV:

Plot(y~x) 🡪 plotten dotgrafiek tussen y en x

Abline(lm(y~x)) 🡪 trendlijn van y en x

<https://rcompanion.org/rcompanion/e_01.html>

**Multiple R-Squared**

*Multiple R-Squared* = Het percentage van y dat verklaard wordt door x. Hoe dichter bij de 1

hoe beter verband er is. Als je onderzoekt of x invloed heeft op y kun je zien bij welk percentage x daadwerkelijk invloed heeft op y.

Als R2 = 1 🡪 perfect verband.

R2 = Maat voor de samenhang tussen x en y.

Residual Standaard deviation (error) 🡪 Gemiddelde afwijking van trendlijn.

*Confounding variabeles* = storende elementen/variabele.

Verband is iets oorzaak-gevolg. Samenhang hoeft niet te betekenen dat er correlatie is. Om oorzaak-gevolg vast te stellen heb je een experiment nodig met maar 1 variabele. Het effect van die variabele is zo vast te stellen

Bij een experiment:

|  |  |
| --- | --- |
| Wie of wat onderzoek je? | Observational unit |
| Gevolg variabele | Response variabele |
| Oorzaak variabele | Explanatory variabele |

BV: Je onderzoekt bij varkens of roken invloed heeft op zuurstofinname. Observational unit is varkens. De explanatory variabele is wel/niet roken en respons variabele is zuurstofinnaneme.

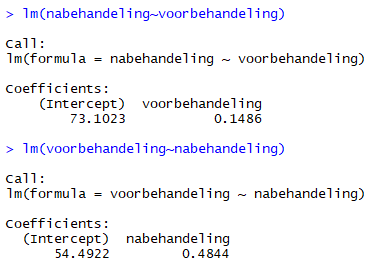
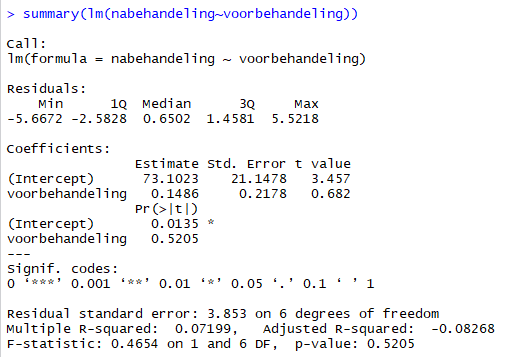
**Samengevat**

*Correlatie* = Is er samenhang tussen variabelen?

* Dit kun je onderzoeken met cor(x, y)
  + Rekent de correlatiecoëfficiënt uit.
    - Is een getal tussen de -1 en +1. Dichtbij 0 wijst of geen samenhang. Hoe verder van de 0 af hoe meer samenhang.
* Kan ook cor.test(x, y) gebruiken
  + Zegt meer dan cor en geeft bijvoorbeeld ook de p-waarde.

*Regressie* = Vertelt hoe de lineaire samenhang is met behulp van de formule: .

* Met lm(y~x) kun je de richtingscoeffiecient (A) en de constante (B) te weten komen,
* Met summary(lm(y~x)) kan dit ook maar daarnaast is er nog addidtionele informatie zoals de p0waarde, vrijheidsgraden, residueafstand etc.0

Links: Verschil tussen lm(y~x) en lm(x~y). Voor juist gebruik moet eerste optie lm(y~x) gebruikt worden.

Rechts: De optie summary(lm(y~x). Geeft meer info dan lm(y~x) maar richtingscoëfficiënt en constante zijn nog steeds te vinden.

**TER HERINNERING:**

Om te testen of de H0 hypothese verworpen kan worden of niet moet je gebruik maken van de p-waarde en de α-waarde. Je vergelijkt de p-waarde met de waarde die je voor α (alfa) hebt gekregen.

*p-waarde* = Waarde waarmee je kunt zien hoe waarschijnlijk je uitkomst is als H0 waar is.

* Hoe kleiner de p-waarde, hoe onwaarschijnlijker de H0 hypothese.
  + Als P-waarde < of gelijk is dan de gegeven α 🡪 H0 verworpen.
  + Als p-waarde > is dan de gegeven α 🡪 H0 niet verworpen.

# College 04 (week 4) 06-12-2017

Waarom hypotheses?

Soms is verschil tussen 2 steekproeven alleen niet genoeg. Als spreiding in uitkomsten in steekproeven klein is kan een verschil wel wijzen op een lineair verband maar vooral bij een grote spreiding is dit niet altijd duidelijk.

Naast verschil heb je dus ook nog een maatstaf voor spreiding nodig.

**Randomisatietest**

Hoe groot is de kans dat je een bepaald gemiddeldeverschil (of nog groter) hebt tussen groepen. Het gaat in op de kans dat waardes “toevallig” zo hoog uitvallen.

* Voordeel randomisatie: Hoeft hierbij geen aannames te maken (ipdv berekenen).

*Methode:*

Je neemt per groep het gemiddelde. Tussen de 2 groepen bereken je het verschil in de gemiddelde. Daarna neem je het totaal aantal van de onderzochte objecten (n = alle observational units van alle groepen bij elkaar). Hiervan neem je de faculteit. Dit deel je door de faculteit van het aantal van 1 groep vermenigvuldigd met totaalaantal - groepaantal faculteit. Wat hieruit komt zijn het aantal mogelijkheden. Deze mogelijkheden zullen gegeven worden (aangezien ze met rekenmachine niet gegeven kunnen worden en met R/Excel duurt dit te lang). In de mogelijkheden kun je de waardes tellen die een net zo groot of groter verschil geven als in jouw resultaten. Het aantal van de resultaten die gelijk of hoger uitvallen dan het verschil in gemiddelde die je in het begin zelf hebt uitgerekend tel je bij elkaar op. Dit aantal deel je door het totale aantal mogelijkheden. De waarde die hieruit komt is de P-waarde.

Dus: (voorbeeld stencil)

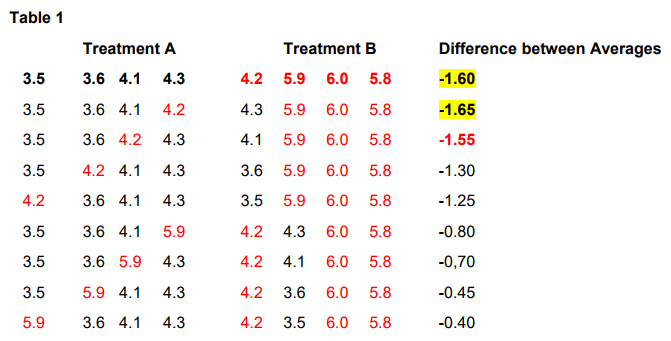
Groep A: 3.5, 3.6, 4.1, 4.3 🡪 gemiddelde = 3,9

Groep B: 4.2, 5.9, 6.0, 5.8 🡪 gemiddelde = 5,5

5,5 - 3,9 = 1,6

H0: Behandeling van groep A heeft hetzelfde effect als de behandeling van groep B (dus: Groep A = Groep B).

HA: Behandeling van groep A heeft niet hetzelfde effect als de behandeling van groep B (dus: Groep A ≠ Groep B).



Van de 8 units in totaal (2x4) pak je de units van 1 groep (4). Vervolgens voer je de formule uit met de faculteiten. Er blijven 70 mogelijkheden over.

In tabel 1 is een deel van de mogelijkhedentabel te zien. Geel gemarkeerd zijn de mogelijkheden met een gemiddeldeverschil dat gelijk of groter is dan die uit de eigen berekening kwam (dus: groter of gelijk aan 1,6).

Over de hele tabel gekeken zijn dit 4 mogelijkheden.

70 mogelijkheden

De totale tabel 1 zal 4 mogelijkheden teruggeven met gemiddeldeverschillen die 1,6 of groter zijn.

🡪 p-waarde

In R kun je verschillende pakketten openen die randomisatie kunnen uitrekenen. Hoeven we (dit jaar) nog niet te kunnen.

Wanneer t.test gebruiken?

* Je waardes moeten normaal verdeeld zijn.
* Je samples moeten onafhankelijk van elkaar zijn.
  + Waardes mogen welk gekoppeld zijn aan waardes aan andere samples maar binnen het sample moeten ze onafhankelijk zijn.
* Je mag alleen t.test gebruiken als je gemiddelde van de 2 steekproeven hebben.
* Als iets een paired t.test is dan zijn de groepsgrootes altijd gelijk.

In R zijn verschillende mogelijkheden om te om te kijken of iets normaal verdeelt zijn.

**Manieren om te zien of iets normaal verdeelt is (soorten testen)**

Shapiro.test

Shipiro.test(x)🡪 De waarde geeft de kans dat het normaal verdeelt kan zijn.

Robinson stelt de acceptabele p-waarde boven 0.1. Bij toets krijg je acceptabele p-waarde echter gegeven.

Hoe hoger hoe beter, dus als uitkomst boven gestelde p-waarde zit is de reeks normaal verdeelt.

Meer info 🡪 ?Shapiro.test

Qqnorm

Qqnorm(x) 🡪 plot moet in een rechte lijn liggen dan normaal verdeelt.

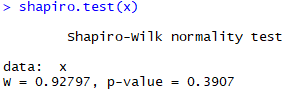
Nadeel: Objectiever

* Wanneer net wel/net niet rechte lijn
* Verschilt ook per persoon wat die acceptabel/recht vindt.

*Normale verdeling =* Alle getallen zijn gekozen uit een lijst die normaal verdeelt is.

*Uniforme verdeling* = Alle getallen hebben de kans om tussen een bepaalde waarde te

zitten.



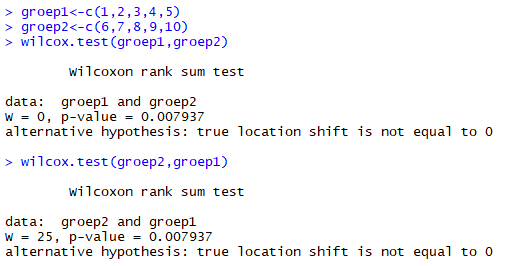
# College 05 (week 5) 13-12-2017

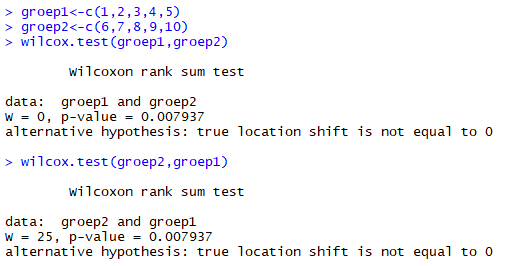
**Wilcoxon-Mann-Whitney**

Gebruik je als je van 2 populaties wil testen of ze dezelfde verdeling hebben.

*In R*

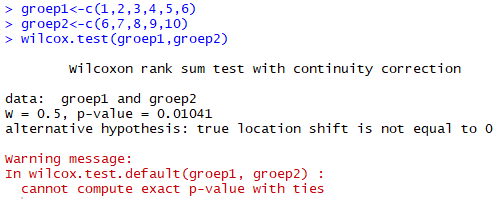
Je maakt van de 2 steekproeven vectoren (groep1 en groep2). Door middel van wilcox.test(…,…) kun je de groepen vervolgens vergelijken. De W geeft de toetsingsgrootheid terug en de p-waarde geeft een getal dat bepaald of H0 verworpen kan worden. R geeft aan dat het de Wilcoxon Rank Sum Test is maar het is de Wilcoxon-Mann-Whitney test.





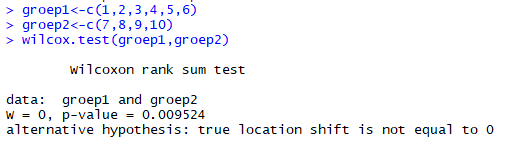
Volgorde van vectoren maakt niet uit.

Er kunnen echter geen dubbele waarden in de groepen zitten anders krijg je een foutmelding. Heb je deze wel kan je het getal een klein beetje veranderen (bv 5 wordt 5,01). P-waarde blijft dan ongeveer gelijk



Vectoren hoeven niet evenveel elementen te bevatten.

Normaal ook paired = TRUE of FALSE ingeven maar bij deze is het de default waarde FALSE en daarom niet erbij gezet.



*Voorwaarden:*

* Geen dubbele getallen.
* Onafhankelijke groepen.
* Alleen bij 2 groepen (2 sampled).

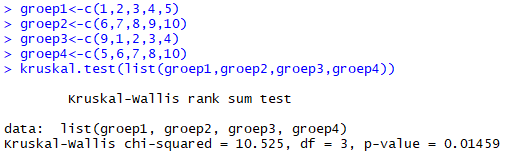
Voorbeeld met dataset:

Wilcox.test(Ozone~Month, data=airquality, subset = Month %in% c(5,9)

* Vergelijk Ozone waardes van Month 5 en 9 uit dataset airquality.

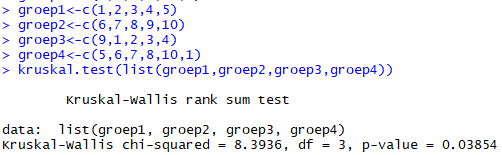
**Kruskal & Wallistest**

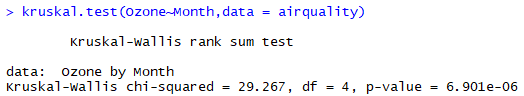
Om te zien of 3+ populaties dezelfde verdeling hebben.



Van de groepen maak je vectoren. Vervolgens vergelijk je ze met de kruskal.rest(list(…,…,…,etc). Met de p-waarde kun je zien of H0 verworpen moet worden.

Vectoren hoeven niet even groot te zijn.

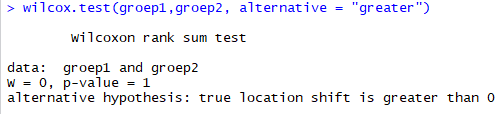




Je kan de kruskal-wallis test ook met een dataset uitvoeren.

# College 06 (week 6) 20-12-2017

Wilcox.test eenzijdig



Wilcox.test(x,y, alternative = “greater”)

🡺 Eerste variabele is groter dan

tweede.

* Volgorde variabele hier wel belangrijk. Welke wil je groter hebben dan andere. Die moet als eerste.

*Toetsingsgrootheid* = Een getal dat je uitrekent op de uitkomsten van je een steekproef.

Grootte van steekproef heeft invloed op p-waarde 🡪 hoe groter je proef,

hoe betrouwbaarder.

* Berekenen toetsingsgrootheid niet gevraagd op tentamen (misschien wel bij herkansing).

Knopen worden niet meegenomen in berekening (of speciaal behandeld in R). Hierdoor klopt p-waarde niet. Steekproeven met veel knopen (dubbele waarden) zijn daarom minder betrouwbaar. Gegeven p-waarde is dan niet exacte p-waarde maar ligt er waarschijnlijk wel dichtbij.

Gepaarde waarnemingen (The Sign test en binom.test)

Je hebt 2 rijen met waardes. Als eerste rij een hogere waarde heeft dan komt er een plusje bij. Heeft de eerste rij een lagere waarde dan komt er een min.

De rij die beter moet zijn dan de andere doe je als rij 1 stellen (bij greater). Vervolgens kijk je of er significant meer plusjes zijn dan minnetjes. Je telt aantal plusjes op en met binom.test kun je dan kijken of het aantal veel is.

Binom.test(9,11, alternative = greater

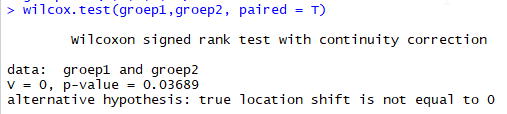
* 9 of meer (of meer door greater) plusjes van 11 objecten.
* Rij 1 had dus 9 waardes die groter waren dan die van rij 2.

**Wilcoxun Signed Rank Test**

Hiermee kijk je bij gepaarde test of 1 groep duidelijk andere waarden heeft dan andere groep.

Verschil WSRT en WMW test?

* WSRT is gepaard.
* WMW is niet gepaard.
* In te stellen door paired = T voor gepaarde test.



*Voorwaarden*

* Constante variabele
* Dat verdeling ongeveer dezelfde vorm hebben.

Wanneer Signed test en Signed Rank test?

Bij WSRT moet je echt de verschillen kunnen aanwijzigen.

# College 07 (week 7) 10-01-2018

**Chi kwadraat toets**

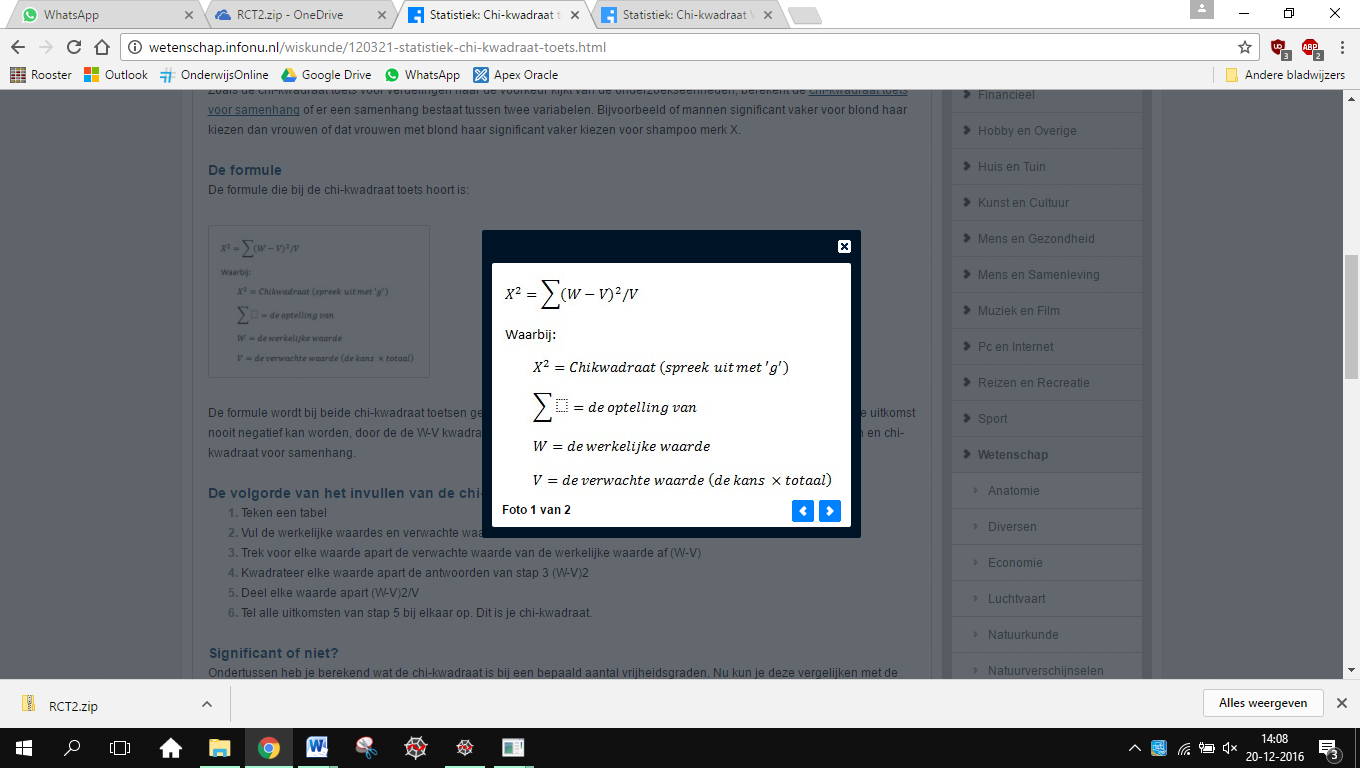
Hypothese toetsen: H0: = gelijke verdeling

Alternatieve hypothese: Ha ≠ niet gelijke verdeling

Vb.

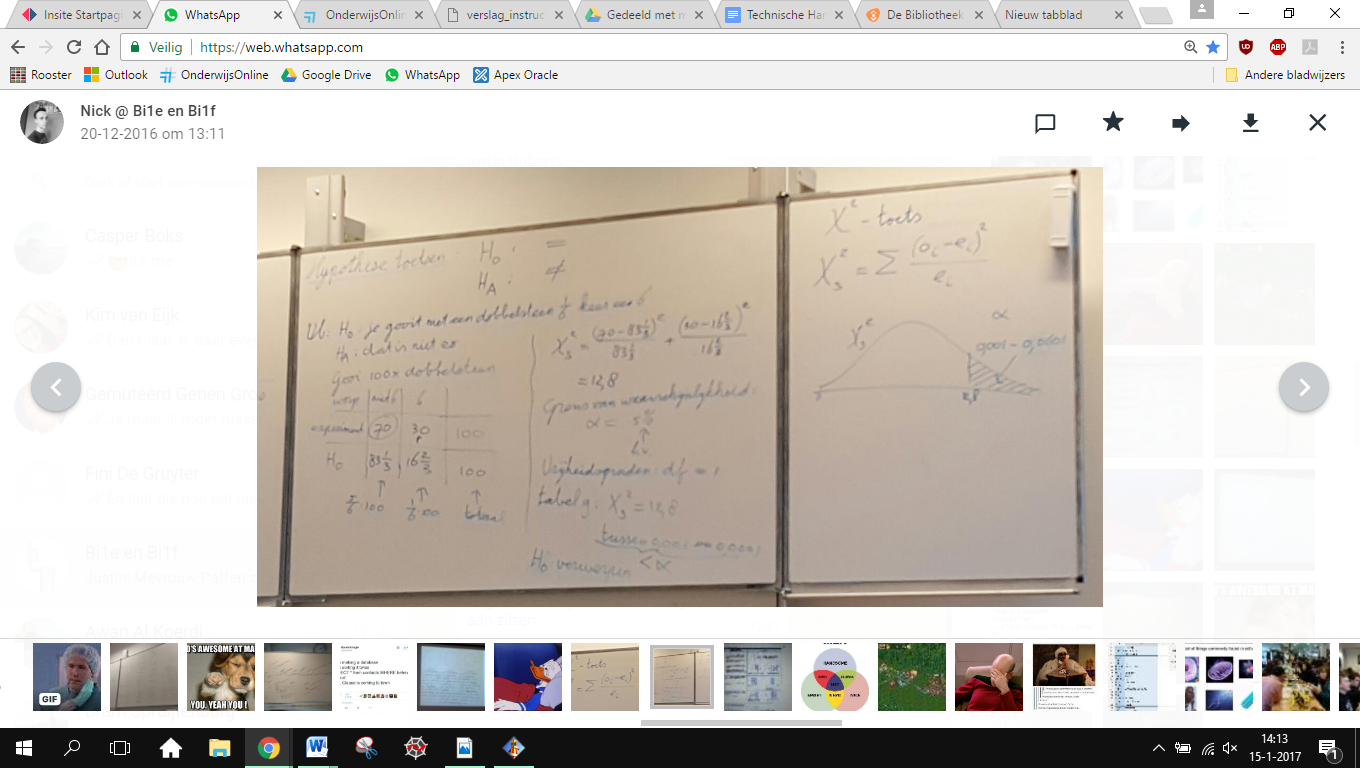
Herten in ongelijkmatig verbrand bos. H0 stelt dat herten gelijk verdeelt blijven over het bos. Ha stelt dat ze niet gelijk verdeelt blijven over het bos.

Chikwadraattoets (Goodness-Of-Fit)



* Voor ieder vakje in de tabel bereken je de chiwaarde (de X2 dus per vakje). Deze tel je op en dan krijg je een chiwaarde. Deze waarde wordt onthouden om te zien of het binnen de toevalkans ligt. Die toevalkans staat in een tabel en wordt bepaald door de waarde die je voor α hebt gekozen en de vrijheidsgraad.

X2 –toets (chikwadraad)

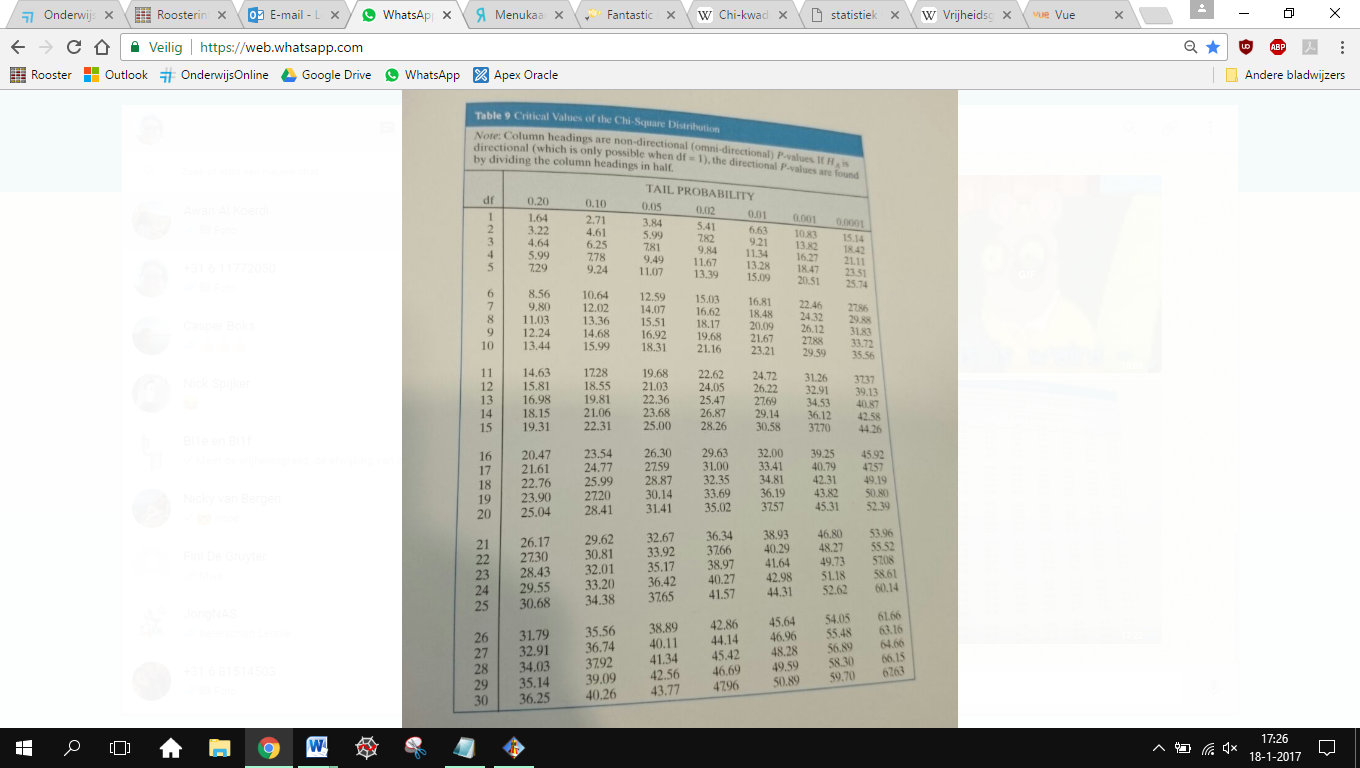


Vrijheidsgraad (df)

Je hebt een tabel en de vrijheidsgraad bereken je door: aantal categorieën - 1.

De vrijheidsgraad heb je nodig om te zien of de chikwadraatwaarde teveel afwijkt voor de H0 hypothese.

Is de waarde van de chiwaardetoets kleiner dan die in de tabel dan valt die binnen de marge en wordt H0 dus niet verworpen. Valt die erbuiten dan is Ha de juiste hypothese.

Links een voorbeeld van een vrijheidsgradentabel.

* Voor tentamen hoeven we niet chikwadraad uit

te kunnen voeren maar moeten we hem wel herkennen en weten wanneer je hem zou toepassen.

*Wanneer chikwadraat?*

Als je een aantal categorieën hebt en bij elke categorie kan je een verwachte waarde invoeren en is er een werkelijke waarde. Op deze manier kun je kijken of er samenhang is tussen 2 (of meer) variabele.

# OVERIGE INFORMATIE

Links

* <http://www.statslectures.com/index.php/r-stats-videos-tutorials/bivariate-analysis/4-9-correlation>
* <http://www.statslectures.com/index.php/r-stats-videos-tutorials/linear-regression/simple-linear-regression-in-r>
* <http://www.statslectures.com/index.php/r-stats-videos-tutorials/linear-regression/checking-assumptions-in-linear-regression>