

## 11 Essentie van de statistiek

Alle theorie die in de voorgaande hoofdstukken is besproken vormt de basis voor dit hoofdstuk. Dit hoofdstuk bespreekt de manieren waarop de theorie in de praktijk wordt gebruikt. De focus ligt hierbij op het vergelijken van variabelen met elkaar. Hierbij kunnen kwalitatieve met kwalitatieve variabelen vergeleken worden, kwalitatieve met kwantitatieve variabelen (en andersom) en kwantitatieve met kwantitatieve variabelen. Deze variabelen worden op verschillende manieren met elkaar vergeleken. Onderdelen hiervoor zijn Student's t-verdelingen,  $X^2$ -verdelingen, F-verdelingen, ANOVA's, lineaire regressies en 7 stappen plannen. Dit hoofdstuk vormt tevens de basis van menig Statistiek tentamen op Universiteiten.

### 11.1 Het vergelijken van variabelen

Zoals al eerder aangegeven kent de statistiek 2 soorten variabelen, kwalitatieve en kwantitatieve. In de tabel hieronder is een overzicht te vinden van de verschillende testen die gebruikt worden bij de vergelijking van variabelen.

	Kwalitatief	Kwantitatief
Kwalitatief	$X^2$ -test	t-test ( $k=2$ ) ANOVA ( $k>2$ )
Kwantitatief		Regressie

De  $k$  bij kwalitatief/kwantitatief staat voor het aantal opties van de kwalitatieve variabele. Bij  $k=2$  is dit dus bijvoorbeeld een variabele waarbij het antwoord "ja" of "nee" kan zijn of "man" of "vrouw". Bij  $k>2$  zal dit bijvoorbeeld een verzameling regio's kunnen zijn. Bijvoorbeeld "Europa", "Noord-Amerika", "Azië", enz.

Voordat hier onder de t-test wordt besproken wordt hier eerst nog even aangegeven welke symbolen waar gebruikt worden.

	Gemiddelde	Variantie	Standaarddeviatie
Populatie	$\mu$	$\sigma^2$	$\sigma$
Steekproef	$\bar{x}$	$s^2$	$s$

De steekproef geeft informatie over de populatie.

### 11.2 t-test

Als men de steekproeven van 2 verschillende populaties bekijkt, kunnen zich 2 zaken voordoen:

1. De steekproeven zijn onafhankelijk van elkaar (Bijvoorbeeld, er is geen relatie tussen "man" en "vrouw".)
2. De steekproeven zijn afhankelijk van elkaar (Bijvoorbeeld, het aantal planten per maand in een bepaald gebied gemeten in 2007 en 2008. Het aantal planten in 2008 is (onder andere) afhankelijk van het aantal planten in 2007.)

#### 11.2.1 Onafhankelijke t-test

Voordat er met de t-test begonnen wordt, zal men eerst moeten bepalen of er een zogenaamde "pre-test" gedaan moet worden. Als de variantie  $\sigma^2$  van de populatie bekend is, is dit niet het geval. Is de variantie  $\sigma^2$  van de populatie niet bekend, dan zal er een pre-test gedaan moeten worden om te bepalen of de variantie van de steekproef  $s_1^2$  gelijk is aan  $s_2^2$ . Kortom, men doet een pre-test om te kijken of

$$s_1^2 = s_2^2$$

of

$$s_1^2 \neq s_2^2$$

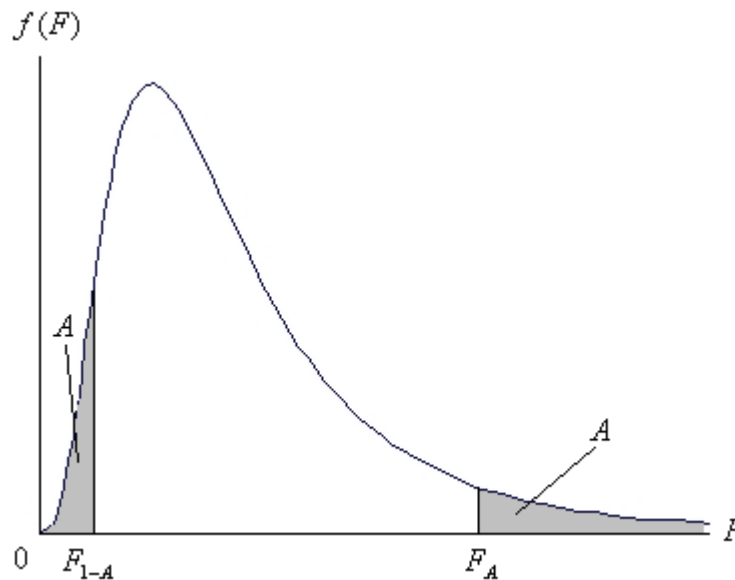
Dit doet men door middel van een tweezijdige F-test. Het feit dat dit een tweezijdige test is geeft aan dat de A uit  $F_{A, v_1, v_2}$  gedeeld wordt door 2.

LET OP! Bij ANOVA en regressie die in latere paragrafen worden besproken zijn de testen altijd eenzijdig.

A (ook wel  $\alpha$ ) is meestal gegeven en  $v_1$  is gelijk aan  $n_1 - 1$  en  $v_2$  is gelijk aan  $n_2 - 1$ . Het symbool  $n$  staat voor de grootte van de steekproef. De kritieke waarden voor deze tweezijdige test berekent men

voor de rechterkant van de curve door het opzoeken van de waarde corresponderend met

$F_{\frac{A}{2}, v_1, v_2}$  en aan de linkerkant van de curve door het berekenen van  $\frac{1}{F_{\frac{A}{2}, v_2, v_1}}$



Figuur 11<sup>a</sup>. F-verdeling

Nu is men in staat door middel van het doorlopen van de al eerder genoemde 7 stappen tot een conclusie te komen. Voor het gemak zijn hieronder de 7 stappen nogmaals opgesomd:

1. Stel  $H_0$  en  $H_1$  vast
2. Stel de toetsingsgrootte vast
3. Stel de kritieke waarden vast → verwerpt men  $H_0$  voor grote waarden, kleine waarden of voor zowel grote als kleine waarden?
4. Stel het significantieniveau  $\alpha$  vast
5. Bepaal het kritieke gebied door middel van het opzoeken van de kritieke waarden van de test
6. Bereken de toetsingsgrootte → valt deze binnen het kritieke gebied, dan wordt  $H_0$  verworpen op significantie niveau  $\alpha$ . Is dit niet het geval, dan wordt  $H_0$  niet verworpen op significantieniveau  $\alpha$ .
7. Geef een duidelijke conclusie voor een breed publiek.

De 7 stappen voor de pre-test van de t-test zouden er dan als volgt uitzien:

1.  $H_0: s_1^2 = s_2^2$   $H_1: s_1^2 \neq s_2^2$
2. toetsingsgrootheid F-test:  $\frac{s_1^2}{s_2^2}$
3. Tweezijdige F-test  $\rightarrow$  verwerp  $H_0$  voor zowel grote als kleine waarden.
4.  $\alpha = A$  is bijvoorbeeld 5% (=0,05)
5. Voorbeeld kritieke waarden:  $F_{\frac{\alpha}{2}, v_1, v_2} = 2$  en  $\frac{1}{F_{\frac{\alpha}{2}, v_2, v_1}} = 0,5$
6. Bereken  $\frac{s_1^2}{s_2^2}$
7. Voorbeeld conclusie: Op een significantieniveau van 5% is er voldoende bewijs om er van uit te gaan dat de varianties niet gelijk zijn. Daarom wordt  $H_0$  verworpen.

Na de pre-test is er bepaald of de varianties van de steekproef wel of niet gelijk zijn. Als de varianties gelijk zijn geldt er een andere procedure voor de t-test dan als de varianties ongelijk zijn.

### 11.2.2 Ongelijke varianties

Na de pre-test wordt er gekeken of de gemiddeldes  $\mu_1$  en  $\mu_2$  aan elkaar gelijk zijn. Als de varianties ongelijk zijn is de volgende toetsingsgrootheid van toepassing:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

waar  $\bar{x}$  staat voor het steekproefgemiddelde en  $\mu_1 - \mu_2$  in de meeste gevallen gelijk is aan 0 en genoteerd wordt als  $\delta_0$ .  $S_p^2$  staat voor de gepoolde variantie en heeft een eigen formule, te weten

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

De formule voor het berekenen van de vrijheidsgraden bij deze t-test is

$$v = n_1 + n_2 - 2$$

Nu kan het 7 stappen schema ook weer toegepast worden:

1.  $H_0: \mu_1 = \mu_2$   $H_1: \mu_1 \neq \mu_2$  (tweezijdig) of  $H_1: \mu_1 > \mu_2$  (eenzijdig) of  $H_1: \mu_1 < \mu_2$  (eenzijdig)

2. toetsinggrootheden: 
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{en} \quad S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

3. Afhankelijk van  $H_1$  is de test eenzijdig of tweezijdig en verwerpt men  $H_0$  voor grote waarden, kleine waarden of voor zowel grote als kleine waarden
4.  $\alpha = A$  is bijvoorbeeld 5% (=0,05)
5. Kritieke waarden worden in de tabel in Appendix A2 gevonden.
6. Bereken  $S_p^2$  en  $t$ .
7. Conclusie: Op een significantieniveau van 5% is...enz

### 11.2.3 Gelijke varianties

In geval van gelijke varianties is er een andere toetsingsgrootheid voor  $t$ , te weten

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In zo'n geval is het aantal vrijheidsgraden gegeven (bijvoorbeeld in SPSS output). De 7 stappen veranderen niet, op stap 2 na waar uiteraard een andere toetsinggrootheid vermeld staat.

In alle bovenstaande 7 stappen schema's is de klassieke methode gehanteerd. Hier kan uiteraard ook de p-waarde methode toegepast worden. In paragraaf 6.5 staat beschreven hoe die methode werkt, maar hier wordt er een korte samenvatting gegeven.

- Verwerp  $H_0$  als  $p > \alpha$
- Accepteer  $H_0$  als  $p > \alpha$

(LET OP! Bij een tweezijdige test wordt  $\alpha$  door 2 gedeeld. Bij het vergelijken van  $\alpha$  met  $p$  wordt er gekeken naar  $\alpha$  voordat er door 2 gedeeld wordt)

#### 11.2.4 Betrouwbaarheidsinterval methode

Bij deze methode wordt gekeken in welk interval het verschil tussen twee gemiddelden van twee populaties ligt bij een bepaald betrouwbaarheidsniveau. Hiervoor bestaat een eenvoudige formule:

$$(\bar{x}_1 - \bar{x}_2) - \delta_0 \pm \text{kritieke waarde} \times \text{standaarddeviatie}$$

Het rechterdeel van de formule  $(\bar{x}_1 - \bar{x}_2) - \delta_0$  is afkomstig uit de formule van de toetsinggrootheid  $t$  bij gelijke varianties. De kritieke waarde is hier niets anders dan het opzoeken van  $t_{\alpha/2, \nu}$  en de standaarddeviatie is tevens afkomstig uit de formule van de toetsinggrootheid  $t$  bij gelijke varianties en is

$$\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n} \right)}$$

Een voorbeeld van een conclusie bij deze methode zou kunnen zijn: Als ik deze procedure heel vaak zou doen, dan zou 95% van de tijd het echte verschil tussen de gemiddelden van de populaties in het interval liggen. (bij een betrouwbaarheidsniveau van 95%)

(LET OP! Een betrouwbaarheidsinterval)

### 11.3 ANOVA

Als een kwalitatieve variabele en een kwantitatieve variabele met elkaar vergeleken worden en de kwalitatieve variabele heeft meer dan 2 opties ( $k > 2$ ) dan gebruikt men een ANOVA (Analysis of variance). Een nulhypothese zou dan kunnen zijn  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ .

#### 11.3.1 eenweg ANOVA

Bij ANOVA is de eenzijdige F-test van toepassing. Van belang zijn ook de sum of squares (kwadratensommen). Men onderscheidt hier 2 typen sum of squares: De sum of squares for treatments (SST) een schatter van de variabiliteit tussen de populaties en de sum of squares of error

(SSE) is een schatter van de variabiliteit binnen de populaties. De formules van SST en SSE zijn vrij eenvoudig:

$$SST = \sum (\bar{x} - \bar{\bar{x}})^2$$

$$SSE = \sum (x - \bar{x})^2$$

In het onderstaande schema is een overzicht te vinden van de benodigdheden van het 7 stappen schema. Hierin staat MST voor de mean square for treatments en MSE voor mean square for error.

	$\nu$	SS	MS
Treatment	$\nu_1 = k - 1$	$SST = \sum (\bar{x} - \bar{\bar{x}})^2$	$MST = SST / \nu_T$
Error	$\nu_2 = n - k$	$SSE = \sum (x - \bar{x})^2$	$MSE = SSE / \nu_E$

De toetsingsgrootte voor deze ANOVA is

$$F = \frac{MST}{MSE}$$

7 stappen schema:

1.  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$   $H_1$ : tenminste 1  $\mu_k$  verschilt
2. toetsingsgrootte:  $F = \frac{MST}{MSE}$
3. eenzijdige test: men verworpt  $H_0$  voor grote waarden
4.  $\alpha = A$  is bijvoorbeeld 5% (=0,05)
5. Kritieke waarden worden in de tabel in Appendix A4 gevonden.
6. Bereken  $F = \frac{MST}{MSE}$
7. Conclusie: Op een significantieniveau van 5% is er voldoende bewijs om er van uit te gaan dat tenminste een van de gemiddelden verschilt. Er is sprake van een zogenaamd "treatment effect"

### 11.3.2 tweeweg ANOVA

In de vorige subparagraaf is de eenweg ANOVA behandeld. Hier werden een kwalitatieve en een kwantitatieve variabele met elkaar vergeleken. Met een ANOVA kan men ook twee kwalitatieve variabelen en een kwantitatieve variabele met elkaar vergelijken. Dit noemt men tweeweg of tweefactor ANOVA. Het schema met de benodigheden voor het 7 stappen plan ziet er dan zo uit.

	$\nu$	$SS$	$MS$
Factor 1	a-1	SSA	$MSA = SSA/a-1$
Factor 2	b-1	SSB	$MSB = SSB/b-1$
Factor 1 x Factor 2	(a-1) x (b-1)	SSAB	$MSAB = SSAB/(a-1) \times (b-1)$
Error	n-(a x b)	SSE	$MSE = SSE/n-(a \times b)$

Er worden nu niet 1, maar 3 7 stappen schema's gemaakt die exact op dezelfde wijze worden gemaakt zoals bij eenweg ANOVA, echter zijn er drie verschillende toetsingsgrootheden:

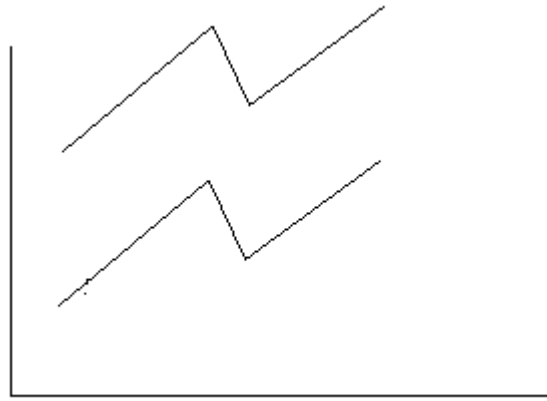
$$F = \frac{MSA}{MSE}$$

$$F = \frac{MSB}{MSE}$$

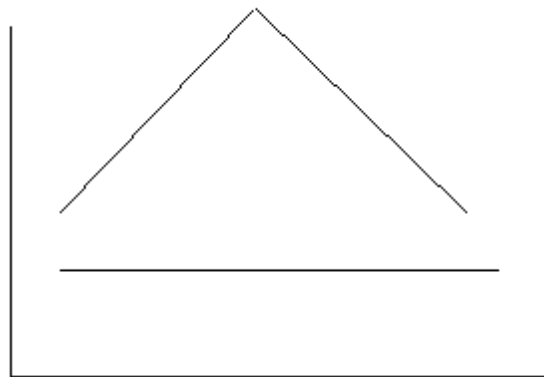
$$F = \frac{MSAB}{MSE}$$

### 11.3.3 Interactie

Interactie houdt in dat het effect van twee variabelen samen anders is dan dat de twee variabelen los van elkaar werken. Dit is te zien in de curve van beide variabelen. De onderstaande figuren geven dit weer.



Figuur 11<sup>a</sup>. Geen interactie



Figuur 11<sup>b</sup>. Interactie

Bij interactie geldt een belangrijke regel. Als de F-test uitwijst dat er interactie is en de curve toont geen interactie, dan is er wel sprake van interactie, daar de F-test als een krachtiger middel wordt beschouwd om dit aan te tonen.

#### 11.4 $X^2$ -test

Voordat aan een  $X^2$ -test begint, waarbij twee kwalitatieve variabelen met elkaar worden vergeleken, dient men eerst een zogenaamde "Goodness of fit"-test te doen.

### 11.4.1 "Goodness of fit"-test

Met een "Goodness of fit"-test kan aangetoond worden hoe goed het model past bij de desbetreffende observaties. Van belang zijn hier de geobserveerde (observed) resultaten en de verwachte (expected) resultaten. De formule voor de toetsingsgrootte is

$$X^2 = \sum \frac{(obs - exp)^2}{exp}$$

waar *obs* staat voor observed resultaten en *exp* staat voor de verwachte resultaten. Een belangrijke regel is dat *exp* minimaal 5 moet zijn wil deze test überhaupt in gang gezet worden.

Hieronder worden het 7 stappen schema gegeven in geval van een multinomiaal experiment. Bij een multinomiaal experiment zijn er meer dan 2 opties/kansen. Die zijn bijvoorbeeld als volgt verdeeld:  $p_1 = 0,3, p_2 = 0,4, p_3 = 0,1, p_4 = 0,2$ . Wat belangrijk is om te weten is dat *exp* te berekenen is door een  $p$  met  $n$  te vermenigvuldigen. De formule voor de vrijheidsgraden is niet ingewikkeld:  $\nu = k - 1$  ( $k$  is gelijk aan het aantal  $p$ 's)

7 stappen schema:

1.  $H_0$ : alle  $p$ 's zijn zoals verwacht  $H_1$ : tenminste 1  $p$  wijkt af van de verwachte waarde
2. toetsingsgrootte:  $X^2 = \sum \frac{(obs - exp)^2}{exp}$
3. eenzijdige test: men verworpt  $H_0$  voor grote waarden
4.  $\alpha = A$  is bijvoorbeeld 5% (=0,05)
5. Kritieke waarden worden in de tabel in Appendix A4 gevonden.
6. Bereken  $X^2 = \sum \frac{(obs - exp)^2}{exp}$
7. Voorbeeldconclusie: Op een significantieniveau van 5% is er voldoende bewijs om er van uit te gaan dat het model past bij de desbetreffende observaties.

### 11.4.2 Test voor onafhankelijkheid

Met een test voor onafhankelijkheid kan worden nagegaan of 2 variabelen onafhankelijk van elkaar zijn. In dit geval is er altijd sprake van een tabel waar beide variabelen in vermeld staan. De toetsingsgrootte is hetzelfde als bij de "Goodness of fit"-test, te weten

$$X^2 = \sum \frac{(obs - exp)^2}{exp}$$

Men berekent *exp* met behulp van de volgende formule:

$$exp = \frac{totaalrijen \times totaalkolommen}{totaalsteekproefgrootte}$$

De formule voor de vrijheidsgraden is ook anders dan bij de "Goodness of fit"-test, te weten

$$v = (aantalrijen - 1)(aantalkolommen - 1)$$

7 stappen schema:

1.  $H_0$ : de variabelen zijn onafhankelijk  $H_1$ : de variabelen zijn afhankelijk
2. toetsingsgrootte:  $X^2 = \sum \frac{(obs - exp)^2}{exp}$
3. eenzijdige test: men verwierpt  $H_0$  voor grote waarden
4.  $\alpha = A$  is bijvoorbeeld 5% (=0,05)
5. Kritieke waarden worden in de tabel in Appendix A4 gevonden.
6. Bereken  $X^2 = \sum \frac{(obs - exp)^2}{exp}$
7. Voorbeeldconclusie: Op een significantieniveau van 5% is er voldoende bewijs om er van uit te gaan dat de 2 variabelen onafhankelijk zijn.

### 11.5 Lineaire regressie

Met lineaire regressie kan men een verband aantonen tussen een stochastische variabele Y en een variabele x. Beiden zijn kwantitatieve variabelen. Zo'n lineaire relatie heeft de volgende vorm:

$$y = \beta_0 + \beta_1 x_i + E$$

Hierbij geldt

$$\beta_0 = \bar{y} - b_1 \cdot \bar{x}$$

$$\beta_1 = \frac{s_{xy}}{s_x^2}$$

Zoals gewoonlijk betreffen de Griekse notaties zoals  $\beta$  populaties en is  $b$  van toepassing op steekproeven.

### 11.5.1 t-test voor de richtingscoëfficiënt

De richtingscoëfficiënt van de regressielijn wordt berekend door middel van een t-test met als toetsingsgrootheid

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

In deze formule is  $\beta_1$  meestal gelijk aan 0 en  $b_1$  en  $s_{b_1}$  zijn meestal af te lezen in bijvoorbeeld output van SPSS. De vrijheidsgraden worden berekend met de onderstaande formule.

$$v = n - k - 1 \text{ waar } k \text{ staat voor het aantal variabelen}$$

7 stappen schema:

1.  $H_0: \beta_1 = 0$   $H_1: \beta_1 \neq 0$  (lineaire relatie) of  $H_1: \beta_1 < 0$  (negatieve lineaire relatie) of  $H_1: \beta_1 > 0$  (positieve lineaire relatie)

2. toetsingsgrootheid:  $t = \frac{b_1 - \beta_1}{s_{b_1}}$

3. Afhankelijk van  $H_1$  is de test eenzijdig of tweezijdig en verwerpt men  $H_0$  voor grote waarden, kleine waarden of voor zowel grote als kleine waarden

4.  $\alpha = A$  is bijvoorbeeld 5% (=0,05)
5. Kritieke waarden worden in de tabel in Appendix A2 gevonden.
6. Bereken  $t = \frac{b_1 - \beta_1}{s_{b_1}}$
7. Conclusie: Op een significantieniveau van 5% is...enz

### 11.5.2 Betrouwbaarheidsinterval

Net als bij de eerder besproken t-test is er bij de regressie sprake van een betrouwbaarheidsinterval.

Het betrouwbaarheidsinterval berekent men met de onderstaande eenvoudige formule:

$$b_1 \pm t_{\alpha/2} \cdot s_{b_1}$$

### 11.5.3 De correlatiecoëfficiënt

Ook om de correlatiecoëfficiënt  $\rho$  te analyseren wordt er een t-test gebruikt. De toetsingsgrootheid hiervoor is

$$t = \sqrt{n-2} \cdot \frac{r}{\sqrt{(1-r)^2}}$$

waar

$$r = \frac{s_{xy}}{s_x s_y}$$

De formule voor de vrijheidsgraden is, net als bij de t-test voor de richtingscoëfficiënt,

$$v = n - k - 1 \text{ waar } k \text{ staat voor het aantal variabelen}$$

7 stappen schema

1.  $H_0: \rho = 0$   $H_1: \rho \neq 0$  of  $H_1: \rho < 0$  of  $H_1: \rho > 0$

2. Toetsingsgrootheden:  $t = \sqrt{n-2} \cdot \frac{r}{\sqrt{(1-r)^2}}$  en  $r = \frac{S_{xy}}{S_x S_y}$
3. Afhankelijk van  $H_1$  is de test eenzijdig of tweezijdig en verwerpt men  $H_0$  voor grote waarden, kleine waarden of voor zowel grote als kleine waarden
4.  $\alpha = A$  is bijvoorbeeld 5% (=0,05)
5. Kritieke waarden worden in de tabel in Appendix A2 gevonden.
6. Bereken  $t = \sqrt{n-2} \cdot \frac{r}{\sqrt{(1-r)^2}}$  en  $r = \frac{S_{xy}}{S_x S_y}$
7. Conclusie: Op een significantieniveau van 5% is...enz