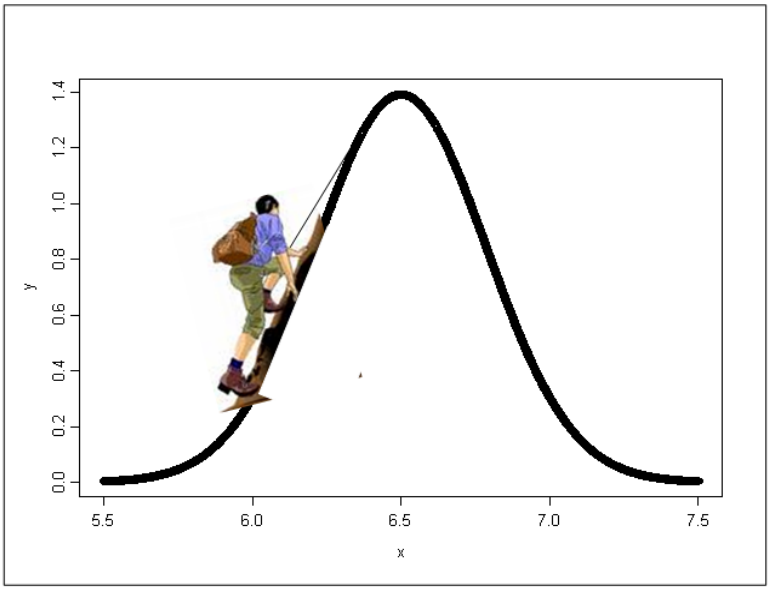
De 3 Mooiste Modellen

Wat elke arts moet weten over statistiek.

Syllabus Medische Statistiek

Hans Burgerhof



BRAINFEED
FOOD FOR THOUGHT

Inhoud

Voorwoord

Toen ik medio 2016 werd gevraagd om een tweedaagse nascholingscursus statistiek te geven voor medici, hoefde ik daar niet lang over na te denken. Na ruim 30 jaar onderwijservaring, waarvan meer dan 25 jaar in de gezondheidszorg, had ik genoeg kennis van het toepassingsgebied en ervaring met de beroepsgroep om hier enthousiast aan te kunnen beginnen.
De door mij geschreven syllabus voor de Junior Scientific Masterclass aan derdejaars studenten Geneeskunde aan de RuG heb ik als basis gebruikt. Door de statistische analyses van twee wetenschappelijke artikelen als uitgangspunt te nemen, heb ik geprobeerd de nadruk op de praktische kant, de interpretatie van de resultaten te leggen. Niettemin duiken we af en toe de theorie in om duidelijk te maken waarom bepaalde statistische methoden onder zekere omstandigheden de voorkeur verdienen.
Ik hoop dat deze cursus bijdraagt in het begrip van statistiek en de cursisten na afloop het wetenschappelijke bos door de statistische bomen zien.

Hans Burgerhof
Groningen, oktober 2016

Inleiding Statistiek en kansrekening

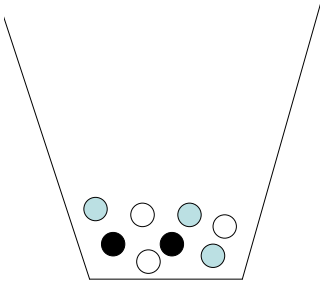
Als je een willekeurige persoon vraagt waar hij of zij aan denkt bij het woord statistiek hoor je regelmatig de antwoorden “grafieken”, “gemiddelden”, “toetsen” en “P-waarde” (niet zelden in combinatie met kwalificaties als “moeilijk” en “saai”). Grafieken en gemiddelden worden gebruikt om de gevonden data te presenteren (beschrijvende statistiek), toetsen en P-waarde zijn termen uit de verklarende statistiek. Bij wetenschappelijk onderzoek komt statistiek echter niet alleen aan de orde op het moment dat alle gegevens verzameld zijn; al voordat men met de dataverzameling begint moet er nagedacht worden over de statistische methode die later gebruikt zal worden en vervolgens over het aantal waarnemingen dat verricht moet worden. Steekproefgroottebepaling en powerberekening zijn begrippen die in deze context gebruikt worden. Als je bijvoorbeeld wilt aantonen dat een nieuw dieet voor personen met obesitas een groter gewichtsverlies geeft dan een standaarddieet, zul je moeten aangeven wat “groter” is en hoeveel mensen met het onderzoek mee moeten doen om met een redelijke kans het verschil tussen de diëten aan te tonen, als dat verschil ook echt bestaat

Laten we voorlopig statistiek omschrijven als de wetenschap die zich bezig houdt met

- het opzetten van onderzoek
- het verzamelen van gegevens
- het overzichtelijk weergeven van de verzamelde gegevens
- het analyseren van de gegevens
- het trekken van conclusies.

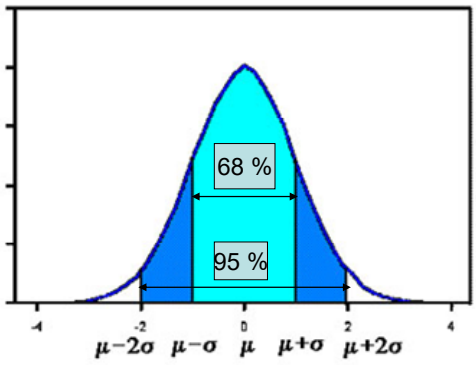
Bij wetenschappelijk onderzoek willen we in de regel een uitspraak doen over een bepaald aspect van een populatie. In de meeste gevallen is de populatie te groot om alle elementen van deze populatie te onderzoeken en zullen we ons moeten baseren op resultaten van een steekproef uit deze populatie. Als we willen onderzoeken hoeveel procent van de zwangere vrouwen in Nederland rookt gedurende de zwangerschap, kunnen we moeilijk alle zwangere vrouwen (nu en in de toekomst) benaderen. We zullen genoegen moeten nemen met een steekproef uit de totale populatie van zwangere vrouwen en op grond van onze bevindingen in deze steekproef een conclusie trekken met betrekking tot de situatie in de populatie. Omdat de waarnemingen in een steekproef van keer tot keer kunnen verschillen, zullen we bij de conclusies ten aanzien van de populatie altijd voorzichtig moeten zijn. Een belangrijke taak van de statistiek is de mate van onzekerheid te kwantificeren.

Om de resultaten uit de steekproef te vertalen naar de populatie is enige kennis van kansrekening noodzakelijk.
Op de middelbare school hebben jullie tijdens de wiskundeles kennis gemaakt met vraagstukken in de vorm van “in een vaas zitten drie rode, drie witte en twee blauwe knikkers. Er worden aselect twee knikkers met terugleggen getrokken. Hoe groot is de kans dat er één rode en één blauwe knikker worden getrokken?” Zie Figuur I.1.



Figuur I.1 *Vaas met knikkers*

Dit is een typisch voorbeeld van een kansrekeningsvraagstuk. De inhoud van de vaas is bekend, er wordt een experiment beschreven en gevraagd wordt om vooraf de kans te berekenen op een bepaalde uitkomst van dat experiment.
In de dagelijkse praktijk van de statistiek gaat het echter anders: de inhoud van de vaas is onbekend (we weten niet hoeveel vrouwen gedurende de zwangerschap roken), maar we weten wel de uitkomst van het experiment (in de steekproef bleken 34 van de 127 vrouwen gedurende de zwangerschap te roken). Wat vertelt ons dit over de inhoud van de vaas? Met andere woorden: wat kunnen we nu zeggen over de populatie? We zullen zelden volstaan met een puntschatting (in de steekproef rookte ongeveer 27 %, dus zal in de populatie ook wel ongeveer 27 % roken) maar liever geven we grenzen waarbinnen met een bepaalde betrouwbaarheid het onbekende percentage zal liggen. Hoe dit in zijn werk gaat zullen we in de volgende hoofdstukken bespreken.
We zullen regelmatig gebruik maken van de normale verdeling en kansen in deze verdeling. Ter herinnering: in iedere normale verdeling ligt tussen de grenzen bepaald door
- één standaarddeviatie (sd) onder het gemiddelde en één sd boven het gemiddelde ongeveer 68 % van de verdeling
- twee sd's onder het gemiddelde en twee sd's boven het gemiddelde ongeveer 95 % van de verdeling
- drie sd's onder het gemiddelde en drie sd's boven het gemiddelde meer dan 99 % van de verdeling, zie Figuur I.2.



Figuur I.2 *De normale verdeling*

Uit bovenstaande blijkt onder andere dat als je een willekeurige waarneming uit een normale verdeling trekt dat de kans ongeveer 0,16 is dat deze waarneming meer dan één sd boven het gemiddelde zit. Voor een gedetailleerder beeld van de kansen in een normale verdeling zijn tabellen ter beschikking. Dit zijn tabellen van de zogenaamde standaard normale verdeling, de normale verdeling met gemiddelde 0 en sd = 1, ook wel de z-verdeling genoemd. Kansvraagstukken met betrekking tot een willekeurige normale verdeling (met gemiddelde μ en sd σ) kunnen vertaald worden naar een kansvraagstuk uit de standaardnormale verdeling. Dit gebeurt door eerst te standaardiseren: $z = \frac{x - \mu}{\sigma}$ en vervolgens gebruik te maken van kansen uit **de tabel van de (standaard) normale verdeling**, zie Tabel I.1.

Tabel I.1. *Één- en tweezijdige P-waarden van de standaard normale verdeling*

z-waarde	Tweezijdige P-waarde (het gebied in beide staarten samen)	Éénzijdige P-waarde (het gebied in de rechterstaart)
0,0	1,00	0,50
0,1	0,92	0,46
0,2	0,84	0,42
0,3	0,76	0,38
0,4	0,69	0,34
0,5	0,62	0,31
0,6	0,55	0,27
0,674	0,50	0,25
0,7	0,48	0,24
0,8	0,42	0,21
0,9	0,37	0,18
1,0	0,32	0,16
1,5	0,13	0,07
1,645	0,10	0,05
1,96	0,05	0,025
2,0	0,045	0,023
2,5	0,012	0,006
2,576	0,010	0,005
3,0	0,003	0,001
3,291	0,001	0,0005

Voorbeeld

In de populatie van vrouwelijke eerstejaars studenten Geneeskunde in Groningen is de lengte normaal verdeeld met gemiddelde 175 cm en sd = 10 cm. Hoe groot is de kans dat een willekeurig gekozen studente uit deze populatie langer is dan 190 cm?
Noem de variabele lengte L, dan geldt

$$P(L > 190) = P(Z > \frac{190 - 175}{10}) = P(Z > 1{,}5)$$

We willen dus weten hoe groot de kans is dat deze studente meer dan anderhalve sd groter is dan het gemiddelde. In bovenstaande tabel zien we dat bij een z = 1,5 een tweezijdige P hoort van 0,13. Dit wordt de tweezijdige overschrijdingskans genoemd. Voor onze vraag zijn we echter alleen geïnteresseerd in de rechteroverschrijdingskans (de kans dat ze groter is dan 190 cm) en dus kijken we in de meeste rechtse kolom voor de éénzijdige P-waarde. De kans is (ongeveer) 0,065.

In veel kansvraagstukken wordt gewerkt met het gemiddelde van een steekproef. Standaardisatie gaat op vergelijkbare wijze als boven, maar bedenk dat de standaarddeviatie van een gemiddelde (de zogenaamde **standard error**) gelijk is aan $\frac{\sigma}{\sqrt{n}}$, waarbij σ de sd van de populatie is en n de grootte van de steekproef.

OPDRACHTEN

Gegeven is dat de lengtes van eerstejaars studentes Geneeskunde normaal verdeeld is met gemiddelde 175 cm en sd = 10 cm.

1.
Hoeveel procent van deze studentes is kleiner dan 170 cm?

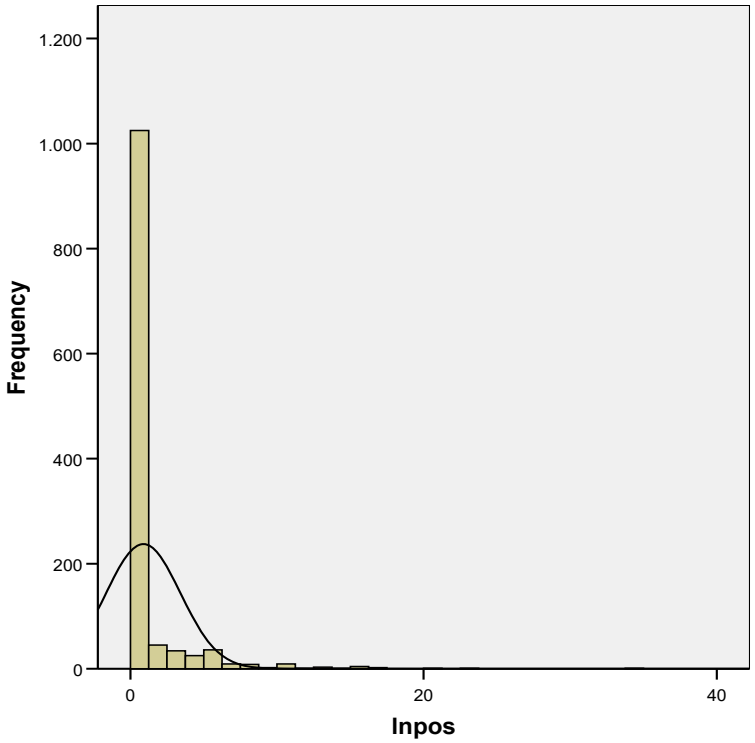
2.
Hoe groot is de kans dat een willekeurige eerstejaars studente Geneeskunde langer is dan 187 cm?

3.
Er wordt een steekproef genomen van 16 studentes. Hoe groot is de kans dat zij gemiddeld langer zijn dan 187 cm?

Deel 1 Een continue responsievariabele

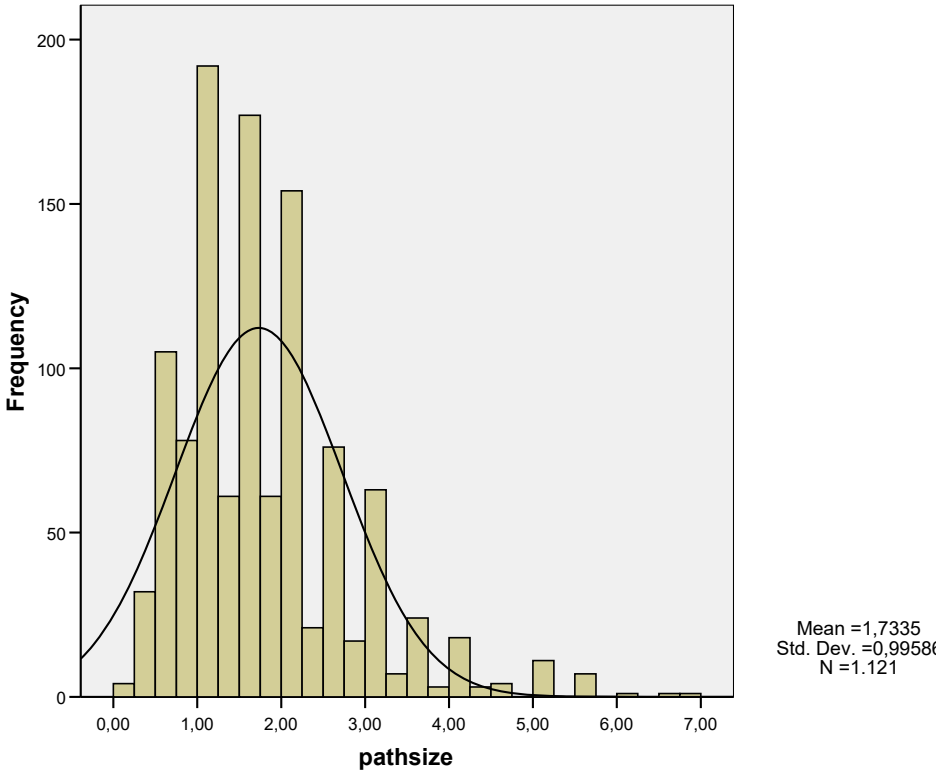
Hoofdstuk 1 Normaal verdeeld?

Bij veel statistische toetsen waarbij de responsievariabele continu is, wordt aangenomen dat deze responsievariabele normaal verdeeld is. De uitkomst van een dergelijke toets is alleen valide als aan die voorwaarde voldaan is. Het is daarom belangrijk om, als je een statistische toets gaat gebruiken, die als voorwaarde heeft dat de betreffende variabele normaal verdeeld is, de aanname van normaliteit te controleren. Dit kan op verschillende manieren, visueel of middels een toets. Omdat het voor veel doeleinden volstaat als de verdeling bij benadering normaal is en volgens de centrale limietstelling het gemiddelde dan redelijk normaal verdeeld is gaan we alleen op de visuele check in.



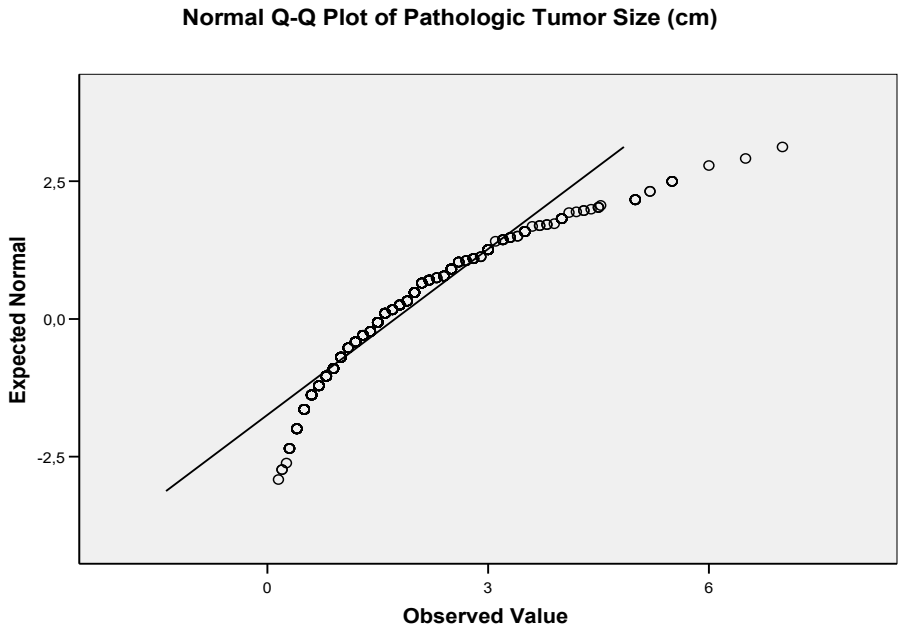
Figuur 1.1 *. Histogram van het aantal positieve lymfeklieren bij 1207 patiënten. In SPSS:* ***Graphs – Legacy Dialogs – histogram***

In Figuur 1.1 zie je het histogram van het aantal positieve lymfeklieren uit de dataset Breastcancer. De grafiek is erg scheef, er was één vrouw met 35 positieve lymfeklieren. Over deze verdeling zal weinig discussie zijn wat betreft de normaliteit.



Figuur 1.2 *Grootte van de tumor (cm³) bij 1121 vrouwen met borstkanker. In SPSS:* ***Graphs - Legacy Dialogs – histogram***

De grafiek in Figuur 1.2 geeft misschien meer aanleiding tot discussie. De verdeling is minder scheef dan die in Figuur 1.1, maar mogen we deze als normaal verdeeld beschouwen? Het is lastig om op het oog staven met vloeiende curven te vergelijken. Een **Q-Q plot** kan uitkomst bieden, zie Figuur 1.3.



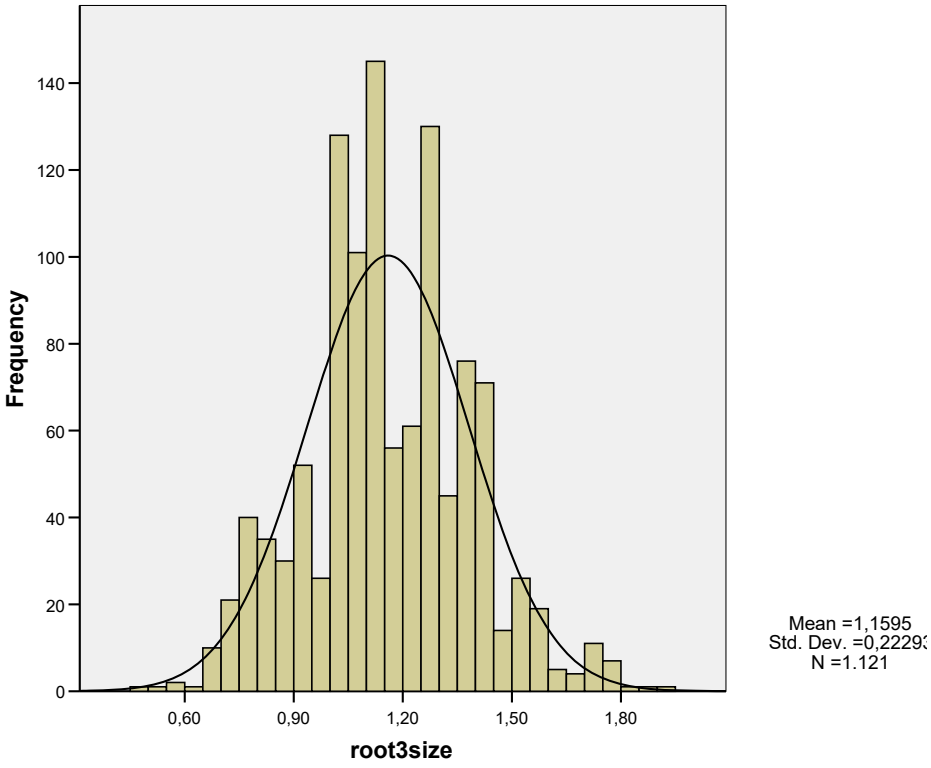
Figuur 1.3 *Q-Q Plot van de grootte van de tumor. In SPSS:* ***Analyze – Descriptive Statistics – Q-Q Plots***

In de Q-Q Plot worden op de Y-as de verwachte kwantielen (als de verdeling normaal zou zijn) afgezet tegen de geobserveerde kwantielen op de X-as. Als de betreffende variabele uit normale verdeling afkomstig is, zouden de punten op de getrokken lijn moeten liggen. Figuur 1.3 laat ons zien dat van een normale verdeling geen sprake is.
Als een variabele niet normaal verdeeld is, is het natuurlijk niet toegestaan om een statistische toets die uitgaat van normaliteit op deze variabele los te laten. We kunnen dan kiezen tussen de volgende opties:

1. een toets toepassen die niet de voorwaarde van normaliteit vereist
2. onze variabele transformeren in de hoop dat het resultaat wel normaal verdeeld is

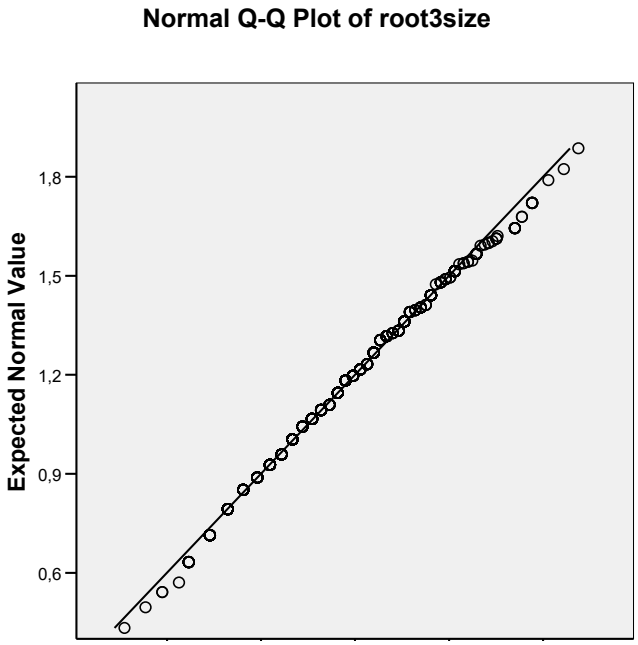
Ad 1. Later in deze syllabus zien we voorbeelden van zogenaamde parametervrije (of non-parametrische) toetsen, die niet de aanname van normaliteit hebben.
Ad 2. Veel gebruikte transformaties op niet-normaal verdeelde variabelen zijn: de logaritmische functie (ln(x)), de reciproce functie (1/x) en de wortel functie ($\sqrt{x}$).

In het voorbeeld van de tumorgrootte blijkt de derde-machts wortel ($\sqrt[3]{x}$) een geschikte transformatie, zie Figuur 1.4. Blijkbaar is de straal van de tumor redelijk normaal verdeeld.

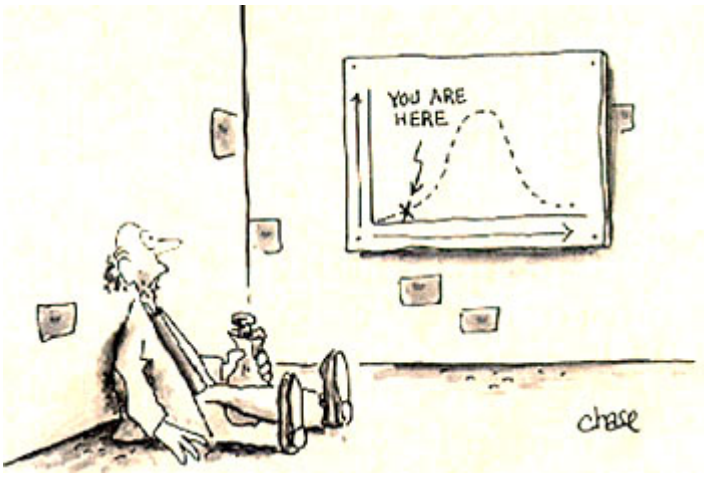


<u>Figuur 1.4</u>** Histogram van de derdemachts wortel van tumorgrootte. In SPSS: **Graphs - Legacy Dialogs – histogram

Dit blijkt ook uit de Q-Q Plot van deze variabele, zie Figuur 1.5.



Figuur 1.5 *Q-Q Plot van de derdemachts wortel van tumorgrootte. In SPSS:* ***Analyze – Descriptive Statistics – Q-Q Plots***

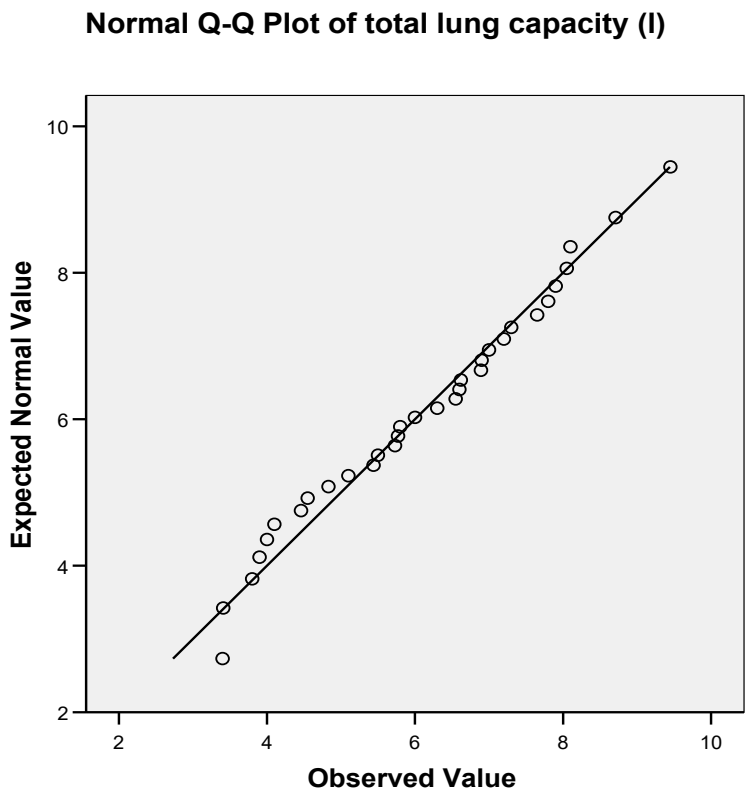


Figuur 1.6 *Niet alles is normaal (noch eerlijk) verdeeld …*

Hoofdstuk 2 Één groep

2.1 Een normaal verdeelde responsievariabele

Schatten
We nemen als voorbeeld de variabele totale longcapaciteit (tlc) van de file totlc.sav. We beschouwen deze gegevens als een aselecte steekproef uit een (niet nader gespecificeerde) populatie en willen de gemiddelde tlc van deze populatie schatten op grond van deze steekproef van 32 individuen. Eerst beoordelen we op grond van een QQ-plot of we de verdeling van de tlc scores als normaal verdeeld mogen beschouwen. Zie Figuur 2.1.



Figuur 2.1 *Q-Q plot van de variabele tlc (totale long capaciteit). In SPSS:* ***Analyze – Descriptive Statistics – Q-Q Plots***

Op grond van de grafiek lijkt de aanname van normaliteit alleszins redelijk. Als maat voor ligging (het "centrum van de verdeling") kiezen we dan voor het gemiddelde.
Het steekproefgemiddelde kunnen we gebruiken als puntschatting van het onbekende populatiegemiddelde μ.
We bestellen in SPSS enkele beschrijvende maten, zie Tabel 2.1.

Tabel 2.1 *. Beschrijvende statistiek van de variabele tlc. In SPSS:* ***Analyze – Descripive Statistics - Frequencies – Statistics*** *(mean, sd, se)*

Statistics

total lung capacity (l)

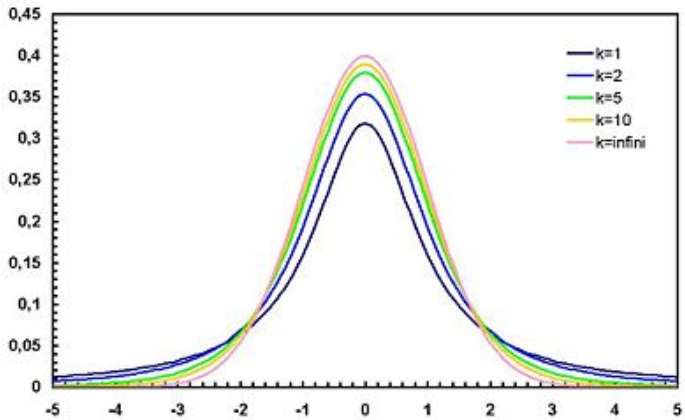
N	Valid	32
	Missing	0
Mean		6,0878
Std. Error of Mean		,28710
Std. Deviation		1,62406

Het steekproefgemiddelde is ongeveer 6,09 en we schatten derhalve dat het populatiegemiddelde ook ongeveer 6,09 liter zal zijn. Maar we willen meer: we willen een gebied aangeven waar met een bepaalde betrouwbaarheid het onbekende gemiddelde in zal liggen. Uitgaande van het steekproefgemiddelde 6,09 en een normale verdeling *met een bekende standaarddeviatie* zal met 95 % betrouwbaarheid het populatiegemiddelde liggen tussen 6,09 – 1,96*SE en 6,09 + 1,96*SE. Hierin is SE (Standard Error of the Mean) de spreidingsmaat voor het gemiddelde en is berekend door de standaarddeviatie (s) te delen door de wortel van het aantal waarnemingen (√n), dus
1,62 / √32 = 0,29. In formule:

Het 95 % betrouwbaarheidsinterval voor het gemiddelde, bij bekende standaarddeviatie:

$$[\overline{X} - 1{,}96 * \frac{\sigma}{\sqrt{n}}; \overline{X} + 1{,}96 * \frac{\sigma}{\sqrt{n}}]$$

In ons voorbeeld zou dat leiden tot [6,09 – 1,96*0,29 ; 6,09 + 1,96*0,29] = [5,53 ; 6,65]. We gaan er hier echter ten onrechte van uit dat de standaarddeviatie van de populatie bekend is. De standaarddeviatie uit bovenstaande tabel (1,62) is een schatting op grond van de steekproef en waarschijnlijk niet exact hetzelfde als de populatie-standaaddeviatie. Deze onzekerheid moet ook tot uitdrukking komen in het betrouwbaarheidsinterval en daarom wordt niet de z-waarde 1,96 gebruikt, maar een waarde uit de **t-verdeling**. Dit is een verzameling van symmetrische verdelingen die qua vorm op de normale verdeling lijken, maar dikkere staarten hebben. De t-verdeling is afhankelijk van het aantal vrijheidsgraden (degrees of freedom (df)), dat gelijk is aan het aantal waarnemingen – 1; df = n – 1.



Figuur 2.2 *. t-verdelingen met k = 1, 2, 5, 10 en oneindig veel vrijheidsgraden. In het laatste geval valt de grafiek samen met de grafiek van de z-verdeling.*

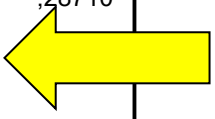
Voor de juiste kritieke waarden van de verschillende t-verdelingen zijn tabellen beschikbaar, zie bijvoorbeeld Altman of Swinscow (table B) of google op "table t-verdeling". Indien de standaarddeviatie uit de steekproef geschat wordt, wordt de formule voor het 95 % betrouwbaarheidsinterval:

$[\overline{X} - t * \frac{s}{\sqrt{n}}; \overline{X} + t * \frac{s}{\sqrt{n}}]$, waarin s de steekproefstandaarddeviatie is. In ons geval, met 31 vrijheidsgraden: [6,09 – 2,04*0,29 ; 6,09 + 2,04*0,29] = [5,50 ; 6,68]
De interpretatie van dit interval: we zijn er voor 95 % zeker van dat het onbekende populatiegemiddelde tussen deze grenzen ligt.
Overigens hadden we ook SPSS om deze grenzen kunnen vragen, zie Tabel 2.2.

Tabel 2.2 *Uitgebreide beschrijvende statistiek van de variabele tlc. In SPSS:* ***Analyze – Descriptive Statistics – Explore***

Descriptives

			Statistic	Std. Error
total lung capacity (l)	Mean		6,0878	,28710
	95% Confidence Interval for Mean	Lower Bound	5,5023	
		Upper Bound	6,6733	
	5% Trimmed Mean		6,0656	
	Median		6,1500	
	Variance		2,638	
	Std. Deviation		1,62406	
	Minimum		3,40	
	Maximum		9,45	
	Range		6,05	
	Interquartile Range		2,66	
	Skewness		,019	,414
	Kurtosis		-,826	,809



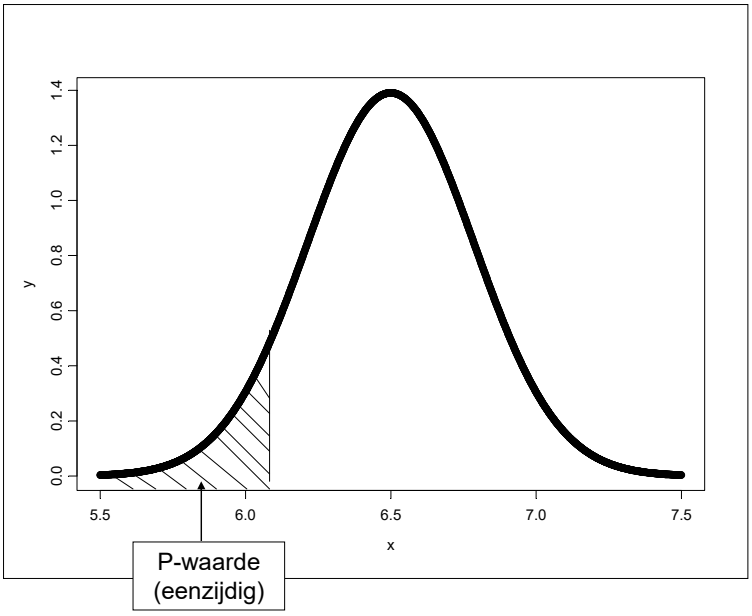
Toetsen
We hebben gezien hoe we een onbekend populatiegemiddelde kunnen schatten. Als de onderzoek(st)er geïnteresseerd is in de vraag of de gemiddelde longcapaciteit gelijk is aan een bepaalde waarde, kan hij/zij besluiten tot een toets. Stel dat deze data afkomstig zijn van een groep proefpersonen met een bepaalde aandoening, bijvoorbeeld chronische bronchitis, en er is bekend dat de gemiddelde longcapaciteit van gezonden 6,5 liter is. Als men vermoedt dat het gemiddelde in de groep met chronische bronchitis lager ligt dan deze 6,5 zullen nulhypothese en alternatief als volgt worden gedefinieerd:

H0: $\mu = 6,5$ ("advocaat van de duivel")

en een alternatieve hypothese die de onderzoeksvraag weerspiegelt, in dit geval

H1 (of Ha): $\mu < 6,5$

Dit is een zogenaamd éénzijdig alternatief. Vaak wordt er tweezijdig getoetst, hier komen we op het eind van deze paragraaf op terug.
Nota bene: deze hypothesen dienen al voorafgaande aan de dataverzameling opgesteld te zijn!
Voorafgaande aan de toetsingprocedure moet nog een significantieniveau (alfa: α) gekozen worden. Dit is de kans die de onderzoeker hooguit wil lopen om ten onrechte de nulhypothese te verwerpen. Omdat we niet graag het gelijk aan onze kant willen claimen terwijl dat niet het geval is, willen we α klein houden; in de regel wordt het significantieniveau α = 0,05 genomen.
Vervolgens worden de data verzameld en gemiddelde en standaarddeviatie van de steekproef berekend. Op grond van deze gegevens wordt dan de zogenaamde **P-waarde** bepaald, de kans op de gevonden steekproefuitkomst of nog extremer, aannemende dat de nulhypothese juist is. Een grote P-waarde wijst er op dat de nulhypothese wel eens juist zou kunnen zijn, terwijl een kleine P-waarde de nulhypothese onwaarschijnlijk maakt. Maar wat is klein? Als de P-waarde kleiner is dan of gelijk aan de gekozen α spreken we van een significant resultaat en verwerpen we de nulhypothese.



Figuur 2.3 *De kans op een steekproefgemiddelde gelijk aan 6,09 of kleiner, als het populatiegemiddelde gelijk is aan 6,5.*

De kans op een steekproefgemiddelde van 6,09 of kleiner als in werkelijkheid het populatiegemiddelde gelijk is aan 6,5 berekenen we met behulp van standaardisatie (naar de z-verdeling als de populatiestandaarddeviatie bekend is, naar de t-verdeling als de standaarddeviatie uit de steekproef geschat wordt):

$$P(\overline{X} < 6{,}09) = P(t < \frac{6{,}09 - 6{,}5}{0{,}287}) = P(t < -1{,}43) \approx 0{,}08 \text{ (uit een tabel)}$$

Bij een significantieniveau van 0,05 kunnen we de nulhypothese niet verwerpen (P > α), we zien dus geen significant verschil en moeten op grond van deze steekproef concluderen dat het gemiddelde in de populatie 6,5 kan zijn.
We kunnen ook aan SPSS de opdracht geven om de P-waarde te berekenen, de uitdraai van de **t-toets** staat in onderstaande tabel. Nota Bene: SPSS geeft standaard de tweezijdige P-waarde.

Tabel 2.3 *De t-toets voor één groep. In SPSS:* ***Analyze – Compare Means – One sample T-test***

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
total lung capacity (l)	32	6,0878	1,62406	,28710

One-Sample Test

	Test Value = 6.5					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
total lung capacity (l)	-1,436	31	,161	-,41219	-,9977	,1733

Éénzijdig versus tweezijdig toetsen

Over het algemeen heeft tweezijdig toetsen de voorkeur boven éénzijdig. Men besluit tot een statistische toets omdat er onduidelijkheid is over de werkelijke situatie. Als je al weet dat therapie A beter werkt dan therapie B dan hoeven we er geen onderzoek aan te wijden. We verkeren in onzekerheid en de uitkomst van het onderzoek kan beide kanten op. Het is helemaal fout om het alternatief te definiëren *nadat* de steekproefgegevens bekend zijn. Je verdubbelt dan het significantieniveau.
Er zijn echter situaties denkbaar waarin éénzijdig toetsen te verdedigen is.

Mag je bij éénzijdig toetsen de tweezijdige P-waarde altijd halveren?

Hoe vanzelfsprekend dat ook lijkt, het antwoord is nee! In het bovenstaand voorbeeld, waarin we de nulhypothese toetsten dat de gemiddelde tlc in deze populatie gelijk was aan 6,5 tegen het alternatief dat het gemiddelde kleiner was dan 6,5 was het geen probleem.
Maar wat zou er gebeurd zijn als ons alternatief was geweest: $\mu > 6{,}5$? SPSS zou exact dezelfde uitdraai hebben geleverd. Het mag duidelijk zijn dat we nu de nulhypothese echter niet kunnen verwerpen. Een steekproefgemiddelde dat *lager* uitvalt dan 6,5 kan natuurlijk nooit reden zijn om de nulhypothese te verwerpen en nu te verkondigen dat het populatiegemiddelde *groter* is dan 6,5!
Wat is hier precies aan de hand?

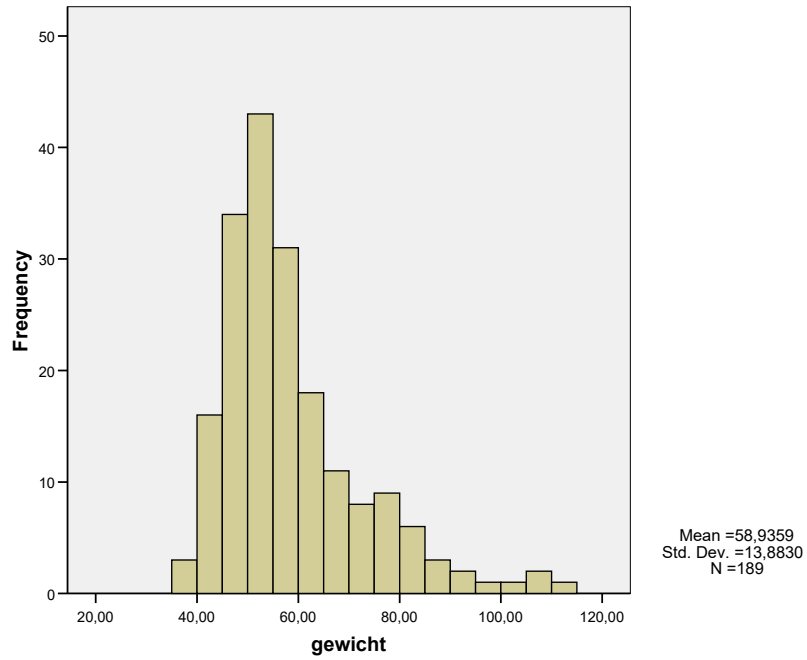
- In het eerste geval (alternatief $\mu < 6{,}5$) zijn we geïnteresseerd in de linkeroverschrijdingskans. Het steekproefgemiddelde valt links van de waarde onder de nulhypothese; we zitten aan de "goede" kant. SPSS heeft de linkeroverschrijdingskans bepaald en verdubbeld. Als wij deze weer halveren vinden we de correcte éénzijdige P-waarde.
- In het tweede geval (alternatief $\mu > 6{,}5$) zijn we geïnteresseerd in de rechteroverschrijdingskans. Het steekproefgemiddelde valt links van de waarde onder de nulhypothese; we zitten aan de "verkeerde" kant. SPSS heeft de linkeroverschrijdingskans bepaald en verdubbeld. Als wij deze weer halveren vinden we een verkeerde éénzijdige P-waarde. Om de correcte (rechteroverschrijdings)kans te vinden moeten we de tweezijdige P-waarde van SPSS halveren (dit is de linkeroverschrijdingskans) en vervolgens van 1 aftrekken.

2.2 Een niet-normaal verdeelde responsievariabele

Schatten

Als de responsievariabele niet normaal verdeeld is, is het gemiddelde in de regel niet de meest interessante maat van ligging (het gemiddelde wordt namelijk sterk beïnvloed door extreme waarnemingen). De mediaan zal dan informatiever zijn. Er zijn verschillende manieren om een betrouwbaarheidsinterval rond de mediaan op te stellen, maar dat valt buiten het bestek van deze cursus.
Een andere optie is om te proberen met een geschikte transformatie (logaritme, wortel, reciproce, …) een nieuwe variabele te creëren die wel (bij benadering) normaal verdeeld is. Vervolgens kan van deze getransformeerde variabele het gemiddelde met bijbehorend betrouwbaarheidsinterval berekend worden. Daarna kunnen de resultaten weer vertaald worden naar de oorspronkelijke meetschaal. Bedenk wel dat dit geen betrouwbaarheidsinterval oplevert van het (ongewogen rekenkundig) gemiddelde maar van het meetkundig gemiddelde in het geval van een logaritmische transformatie (zie appendix A).

Als voorbeeld kijken we naar de variabele gewicht uit de file lowbwt.sav.



Figuur 2.4 *Gewicht van de moeder tijdens de laatste menstruatie. In SPSS:* ***Graphs - Legacy Dialogs – histogram***

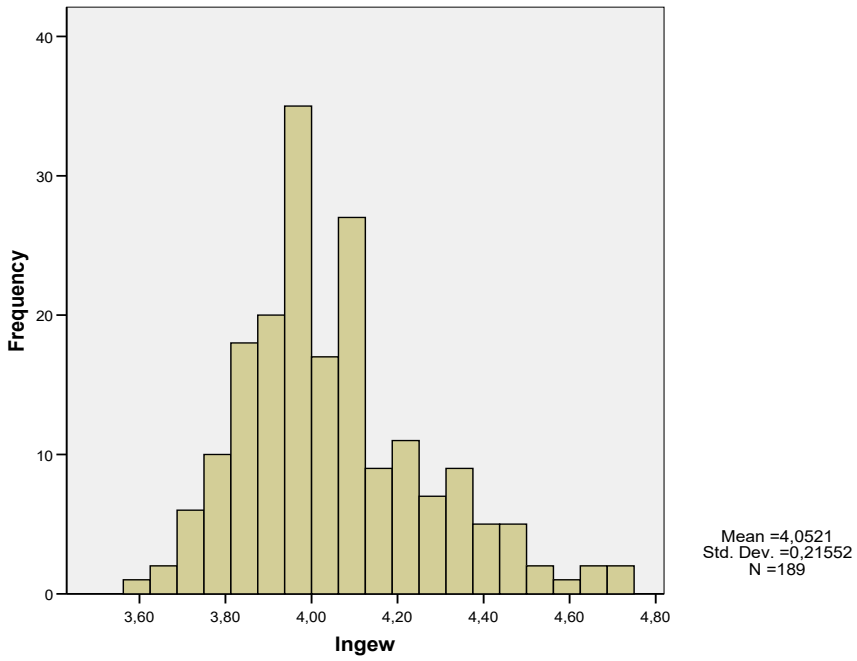
Het histogram in Figuur 2.4 geeft aan dat de verdeling van de gewichten van de moeders scheef naar rechts is. Het 95 % betrouwbaarheidsinterval voor het gemiddelde in Tabel 2.4 is gebaseerd op de aanname van normaliteit en zal dus niet correct zijn.

Tabel 2.4 *Beschrijvende informatie van het gewicht van de moeder. In SPSS:**Analyze – Descriptive statistics - Explore***

Descriptives

			Statistic	Std. Error
gewicht	Mean		58,9359	1,00984
	95% Confidence Interval for Mean	Lower Bound	56,9438	
		Upper Bound	60,9280	
	5% Trimmed Mean		57,8207	
	Median		54,9340	
	Variance		192,739	
	Std. Deviation		13,88304	
	Minimum		36,32	
	Maximum		113,50	
	Range		77,18	
	Interquartile Range		13,85	
	Skewness		1,402	,177
	Kurtosis		2,404	,352

Als we de variabele gewicht logaritmisch transformeren, lijkt de verdeling meer op een normale verdeling, zie Figuur 2.5. (Dit histogram is ter vergelijking van Figuur 2.4, een Q-Q plot was hier beter op zijn plaats geweest om de normaliteit te beoordelen).



Figuur 2.5 *Histogram van de logaritme van gewicht (van de moeder). In SPSS: **Graphs - Legacy Dialogs – histogram***

Tabel 2.5 *Beschrijvende informatie van de logaritme van gewicht. In SPSS:**Analyze – Descriptive statistics - Explore***

Descriptives

			Statistic	Std. Error
lngew	Mean		4,0521	,01568
	95% Confidence Interval for Mean	Lower Bound	4,0212	
		Upper Bound	4,0830	
	5% Trimmed Mean		4,0430	
	Median		4,0061	
	Variance		,046	
	Std. Deviation		,21552	
	Minimum		3,59	
	Maximum		4,73	
	Range		1,14	
	Interquartile Range		,24	
	Skewness		,722	,177
	Kurtosis		,541	,352

Uit Tabel 2.5 kunnen we het 95 % betrouwbaarheidsinterval voor het gemiddelde van de logaritmisch getransformeerde variabele aflezen: [4,02 ; 4,08]. Als we het 95 % betrouwbaarheidsinterval vertalen naar de oorspronkelijke schaal krijgen we $[e^{4,0212}; e^{4,0830}] = [55,77\,; 59,32]$. Dit is dus een 95 % betrouwbaarheidsinterval voor het **meetkundig gemiddelde** (zie Appendix A) van het gewicht en wijkt af van het 95 % BI dat gegeven was in Tabel 2.4. De puntschatting van het meetkundig gemiddelde ($e^{4,05} = 57,52$) ligt tussen de mediaan (54,93) en het gemiddelde (58,94) in.

Toetsen

Als de responsievariable niet normaal verdeeld is, kunnen we de t-toets niet gebruiken. We kunnen weer op zoek naar een geschikte transformatie en, als de getransformeerde variabele redelijk normaal verdeeld is, de t-toets op de getransformeerde variabele toepassen.
Een andere optie is om een **non-parametrische test** op de oorspronkelijke variabele toe te passen. Non-parametrische of parametervrije toetsen maken weinig of geen veronderstellingen over de populatieverdeling. In het geval van één groep komen de **tekentoets** (in het Engels **sign-test**) en de **rangtekentoets** (**Wilcoxon's signed rank test**) daarvoor in aanmerking.

Als voorbeeld kijken we naar de variabele "pathsize", de grootte van de tumor in de file Breastcancer.sav. Zoals te zien in Figuur 1.3 is deze variabele niet normaal verdeeld. Als we willen toetsen of de mediaan van deze verdeling gelijk is aan 1 cm³, kunnen we als volgt redeneren: als H_0 waar is (de mediaan is 1) dan heeft iedere waarneming een kans gelijk aan 0,5 om kleiner te zijn dan 1 en een kans van 0,5 om groter te zijn dan 1 (er van uitgaande dat de kans op exact 1 gelijk is aan 0. Mochten er toch waarnemingen gelijk aan 1 zijn, dan laten we deze buiten beschouwing). Dus het aantal waarnemingen groter dan 1 is een binomiaal verdeelde variabele met steekproefgrootte n en p = 0,5. De nulhypothese wordt verworpen als we significant meer waarnemingen groter of kleiner dan 1 zien, dan we op grond van toeval mogen verwachten.
In SPSS moet je, om deze test te kunnen uitvoeren eerst een kolom met enen maken (deze variabele is "path1" genoemd in Tabel 2.6) om de tumorgroottes mee te vergelijken. In de

datafile Breastcancer.sav zijn er 1121 gegevens over de grootte van de tumor, waarvan er 107 waarnemingen gelijk aan 1, zie Tabel 2.6. Van de overige 1014 waarnemingen verwachten we, als de nulhypothese juist is, dat er 507 kleiner dan 1 zijn en 507 groter dan 1. Op grond van toeval kunnen deze aantallen natuurlijk verschillen, maar als de gevonden aantallen erg afwijken, is dat reden om de nulhypothese te verwerpen. 785 waarnemingen blijken groter dan 1 te zijn, 219 kleiner.

Tabel 2.6 *Beschrijvende informatie bij de Sign test*

Frequencies

		N
path1 - Pathologic Tumor Size (cm)	Negative Differences[a]	795
	Positive Differences[b]	219
	Ties[c]	107
	Total	1121

a. path1 < Pathologic Tumor Size (cm)

b. path1 > Pathologic Tumor Size (cm)

c. path1 = Pathologic Tumor Size (cm)

Het is mogelijk om de exacte P-waarde van de Sign test te laten berekenen, maar SPSS geeft default de normale benadering, zie Tabel 2.7.

Tabel 2.7 *Resultaat van de Sign test. In SPSS:* ***Analyze – Nonparametric Tests – 2 Related Samples –*** *vink aan:* ***Sign test*** *. Voor de exacte P-waarde klik op* ***Exact****. (In dit voorbeeld krijg je dezelfde P-waarde als bij de "asymptotische").*

Test Statistics[a]

	path1 - Pathologic Tumor Size (cm)
Z	-18,057
Asymp. Sig. (2-tailed)	,000

a. Sign Test

De Sign test (tekentoets) kijkt alleen naar het aantal waarnemingen met een positief respectievelijk negatief teken. De P-waarde is kleiner dan 0,05, dus de nulhypothese dat 1 de mediaan is, wordt verworpen. Omdat we meer negatieve verschillen vinden (1 – tumorgrootte < 0) dan positieve zijn er dus relatief meer tumorgroottes die groter zijn dan 1. De mediaan ligt significant meer naar rechts.

Bij de Sign test wordt de grootte van de afwijking ten opzichte van 1 niet in ogenschouw genomen. Wilcoxon's signed rank test (de rangtekentoets) doet dit wel. Van iedere waarneming wordt gekeken naar de absolute afwijking ten opzichte van 1. Deze worden vervolgens gesorteerd van klein naar groot en van rangnummers voorzien. Daarna wordt berekend wat het gemiddelde rangnummer van de positieve afwijkingen is en evenzo van de negatieve afwijkingen (zie Tabel 2.8).

Tabel 2.8 *Resultaat van Wilcoxon's Signed Rank test. In SPSS:* ***Analyze – Nonparametric Tests – 2 Related Samples***

Ranks

		N	Mean Rank	Sum of Ranks
path1 - Pathologic Tumor Size (cm)	Negative Ranks	795[a]	579,46	460672,00
	Positive Ranks	219[b]	246,27	53933,00
	Ties	107[c]		
	Total	1121		

a. path1 < Pathologic Tumor Size (cm)

b. path1 > Pathologic Tumor Size (cm)

c. path1 = Pathologic Tumor Size (cm)

Test Statistics[b,c]

	path1 - Pathologic Tumor Size (cm)
Z	-21,825[a]
Asymp. Sig. (2-tailed)	,000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

c. Some or all exact significances cannot be computed because there is insufficient memory.

We zien dat er niet alleen meer waarnemingen boven de 1 zijn dan eronder, maar bovendien is hun gemiddelde afwijking ten opzichte van 1 groter dan die van de waarnemingen onder 1. Een voorwaarde voor Wilcoxon's toets is dat de gegevens uit een symmetrische verdeling komen. Aangenomen dat dat juist is, zien we dat ook hier de H_0 wordt verworpen. De aanname van symmetrie lijkt hier echter niet terecht (zie Figuur 1.2).

OPDRACHTEN

(voor de opdrachten in dit en de volgende hoofdstukken geldt bij toetsing steeds een significantieniveau $\alpha = 0{,}05$)

1.
Toets of de moeders uit de dataset lowbwt.sav gemiddeld 25 jaar zijn. Gebruik zowel een parametrische als een niet-parametrische toets. Welke toets mag je hier gebruiken?
Wat is de conclusie?

2.
Bij het onderzoek "Groningen beweegt" werden twee verschillende bewegingsprogramma's aangeboden. 20 personen werden random over deze groepen verdeeld en hun gewichten werden twee maal gemeten (in kg): voorafgaande aan het onderzoek en na afloop.

Groep	Gewicht1	Gewicht2
1	80	77
1	74	72
1	77	73
1	64	60
1	69	65
1	88	83
1	82	79
1	75	75
1	68	67
1	72	70
2	74	73
2	65	62
2	89	87
2	74	73
2	69	68
2	75	74
2	81	80
2	76	77
2	68	70
2	80	78

Maak een SPSS file van deze data en bewaar de file onder de naam "Beweeg.sav".
Laat in SPSS een nieuwe variabele berekenen die het verschil in gewicht aangeeft tussen gewicht1 en gewicht2.
Toets of de nulhypothese dat de mensen *van groep 1* niet in gewicht veranderd zijn. (TIP: je kunt groep 1 selecteren met behulp van Data - Select cases).
Welke toets gebruik je en waarom? Wat is de conclusie?

Hoofdstuk 3 Twee groepen

3.1 Twee onafhankelijke groepen met een normaal verdeelde responsievariabele

Schatten

In de klinische praktijk komt het vaak voor dat we twee groepen (patiënten versus gezonden, experimentele groep versus controle groep) met elkaar willen vergelijken op grond van een continue responsievariabele, zoals bij voorbeeld longcapaciteit, bloeddruk of gewichtsverlies. Als de betreffende variabele normaal verdeeld is, richten we ons vaak op het verschil in groepsgemiddelden. Het verschil in populatiegemiddelden $\delta = \mu_1 - \mu_2$ wordt geschat met behulp van het verschil in steekproefgemiddelden: $d = \overline{Y}_1 - \overline{Y}_2$. Het berekenen van deze puntschatting is geen probleem, maar we willen de onzekerheid in deze schatting ook kwantificeren. De wijze waarop het daarvoor benodigde betrouwbaarheidsinterval berekend wordt vergt meer uitleg.

Een stukje theorie: een kansvariabele (of stochastische variabele) is een variabele die op grond van toeval verschillende waarden kan aannemen. Iedere kansvariabele heeft een eigen kansverdeling. Als je de verdeling van een kansvariabele S bekijkt, die de som is van twee kansvariabelen X en Y (dus S = X + Y), dan zal S over het algemeen een grotere spreiding vertonen dan X en Y afzonderlijk. Immers als X en Y beide waarden aan kunnen nemen tussen 0 en 10, kan S waarden aannemen tussen 0 en 20. Je kunt bewijzen dat, als X en Y onafhankelijke kansvariabelen zijn, de variantie van S gelijk is aan de variantie van X plus de variantie van Y. In formule: $\sigma_S^2 = \sigma_X^2 + \sigma_Y^2$.

Aangezien de standaarddeviatie gelijk is aan de wortel van de variantie geldt voor de sd van S: $\sigma_S = \sqrt{\sigma_X^2 + \sigma_Y^2}$.

Ook voor de kansvariabele V die het verschil is tussen X en Y (dus V = X – Y) geldt dezelfde formule:

f1 $\sigma_V = \sqrt{\sigma_X^2 + \sigma_Y^2}$.

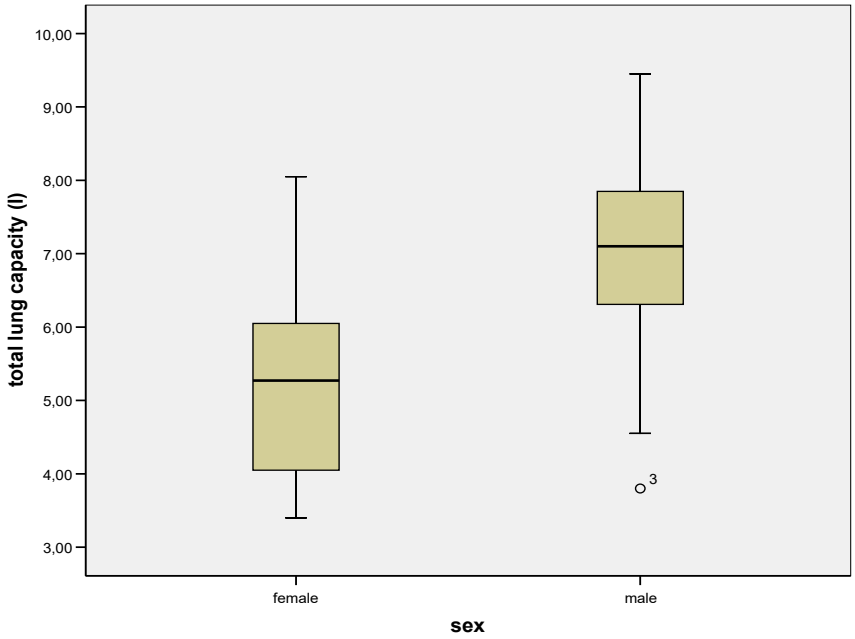
(Misschien had je in eerste instantie verwacht dat de spreiding zou afnemen als je twee kansvariabelen van elkaar aftrekt. Bedenk dan dat, als X en Y beide waarden aan kunnen nemen tussen 0 en 10, V waarden kan aannemen tussen -10 en 10).

Terug naar een praktisch voorbeeld: stel dat onderzoekers het verschil willen schatten in gemiddelde totale longcapaciteit van mannen en vrouwen in afwachting van hun hart-long transplantatie (zie de file totlc.sav).

In Figuur 3.1 zien we een **Boxplot** van de totale longcapaciteit (tlc) voor mannen en vrouwen afzonderlijk. De medianen (de vette strepen in de box) van mannen en vrouwen lijken nogal te verschillen. De verdelingen lijken redelijk symmetrisch, en Q-Q Plots per geslacht (die hier niet getoond worden) bevestigen de normaliteit van tlc. Voor het vervolg nemen we aan dat de verdelingen normaal zijn. Via **Analyse – Descriptive Statistics – Explore** (met geslacht in de **Factor List)** krijgen we onder andere de volgende informatie over tlc:

	n	Gemiddelde	sd
vrouwen	16	5,20	1,30
mannen	16	6,98	1,43

De puntschatting voor het verschil tussen de gemiddelden van mannen en vrouwen is 6,98 – 5,20 = 1,78.
Hoe stellen we een 95 % betrouwbaarheidsinterval op voor het verschil in gemiddelden?



Figuur 3.1 *Boxplot van de Totale longcapaciteit voor mannen en vrouwen. In SPSS:* ***Graphs – Legacy Dialogs – Boxplot – Simple***

De gemiddelde longcapaciteit van ieder geslacht is een kansvariabele en van beide gemiddelden kan de standaarddeviatie ("standard error") geschat worden. De standaarddeviatie van het verschil V tussen deze gemiddelden kan daar uit afgeleid worden met formule **f1** van de vorige bladzijde:

f2 $$s_v = \sqrt{\frac{s_{man}^2}{n_{man}} + \frac{s_{vr}^2}{n_{vr}}},$$

waarin *s* staat voor de standaarddeviatie uit de steekproef. Het 95 % BI kan volgens de volgende formule berekend worden: $[v - t * s_v ; v + t * s_v]$, waarbij de t-waarde bepaald wordt door het percentage betrouwbaarheid en het aantal vrijheidsgraden. (Eigenlijk is dit niet geheel correct. Als de spreidingen in de twee groepen niet gelijk zijn, hebben we hier te maken met een mixture van twee verschillende t-verdelingen. Deze kan worden benaderd door een t-verdeling met een aangepast aantal vrijheidsgraden).
Indien de spreidingen van mannen en vrouwen gelijk verondersteld mogen worden, is er een alternatieve formule voor **f2**. Deze wordt hieronder besproken bij de uitleg van de t-toets.

Voor ons voorbeeld van de tlc is de $s_v = \sqrt{\frac{1{,}43^2}{16} + \frac{1{,}30^2}{16}} = 0{,}48$ en het 95 % BI:

[1,78 – 2,042*0,48 ; 1,78 + 2,042*0,48] = [0,79 ; 2,77]

Het getal 2,042 is de kritieke waarde uit de t-verdeling met (16 – 1) + (16 – 1) = 30 vrijheidsgraden die hoort bij een tweezijdige overschrijdingskans van 0,05.
We weten dus met 95 % waarschijnlijkheid dat het onbekende verschil tussen de twee populatiegemiddelden tussen deze grenzen ligt. Het interval bevat de waarde 0 niet, hetgeen duidt op een echt verschil tussen de gemiddelden van mannen en vrouwen.

Toetsen

Voor het toetsen van de nulhypothese dat de gemiddelden van de normaal verdeelde variabele tlc van mannen en vrouwen gelijk zijn (H0: $\mu_m = \mu_v$) gebruiken we de **t-toets voor onafhankelijke groepen.**

Tabel 3.1 *De t-toets voor onafhankelijke groepen. In SPSS:* ***Analyze – Compare Means – Independent-Samples T Test***

Group Statistics

	sex	N	Mean	Std. Deviation	Std. Error Mean
total lung capacity (l)	female	16	5,1981	1,30082	,32521
	male	16	6,9775	1,43882	,35970

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
total lung capacity (l)	Equal variances assumed	,001	,976	-3,669	30	,001	-1,77938	,48492	-2,76971	-,78904
	Equal variances not assumed			-3,669	29,700	,001	-1,77938	,48492	-2,77013	-,78862

In het bovenste deel van Tabel 3.1 zien we beschrijvende informatie per groep, in de onderste tabel staan drie verschillende toetsen. We zien in de vierde kolom twee t-waarden. Deze zijn hier beide even groot (-3,669), maar dat is meestal niet het geval. Er worden namelijk twee verschillende t-toetsen uitgevoerd. De onderste t-toets ("Equal variances not assumed") maakt gebruik van afzonderlijke schattingen van de standaarddeviaties per groep, zoals in formule **f2**. De bovenste t-toets ("Equal variances assumed") veronderstelt dat de spreiding in beide groepen even groot is en maakt één gemeenschappelijke schatting van de standaarddeviatie (de "gepoolde" standaarddeviatie s_p). Deze wordt als volgt berekend:

f3 $$s_p = \sqrt{\frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}}$$

De geschatte gemeenschappelijke variantie is een gewogen gemiddelde van de varianties van de groepen.

In dit geval leiden beide methoden tot dezelfde afgeronde schatting van de standaarddeviatie van het verschil (achtste kolom) en derhalve tot dezelfde t-waarde. Hier maakt het dus niet uit naar welke t-toets je kijkt, maar in de meeste gevallen zul je verschillende t-waarden (en dus verschillende P-waarden) vinden. Hoe weet je in dat geval welke t-toets je moet gebruiken? In het linker deel van de onderste tabel zie je **Levene's toets**. Deze toetst de nulhypothese dat de varianties in de groepen gelijk zijn. Als deze nulhypothese niet wordt verworpen (Sig. = P-waarde $> 0{,}05$) mogen we de spreidingen aan elkaar gelijk beschouwen en nemen we de bovenste t-toets; als Levene's nulhypothese wel wordt verworpen (Sig. = P-waarde $\leq 0{,}05$) moeten we de onderste t-toets gebruiken.

Merk op dat in het rechter deel van de tabel het 95 % BI voor het verschil in gemiddelde tlc wordt gegeven. Dit is, op de mintekens en afronding na, gelijk aan het BI dat we op de vorige bladzijde hebben berekend.

3.2 Twee onafhankelijke groepen en een niet normaal verdeelde responsievariabele

Schatten

Er is een methode ontwikkeld voor het opstellen van een betrouwbaarheidsinterval voor het verschil in medianen (onder een bepaalde voorwaarde) maar deze wordt weinig gebruikt en is tamelijk complex. Deze methode wordt hier niet verder besproken.

Toetsen

Als de responsievariabele niet normaal verdeeld is, kunnen we de t-toets niet gebruiken. Er is een nonparametrische toets om twee onafhankelijke groepen met elkaar te vergelijken: de **Mann-Whitney U-toets.**

> (Ongeveer in dezelfde tijd dat de heren Mann en Whitney deze toets bedachten heeft de heer Wilcoxon een toets afgeleid die equivalent is aan deze toets. De naam Mann-Whitney wordt echter vaker gebruikt omdat de toets van **Wilcoxon, de Ranksum test**, mogelijk naamsverwarring kan opleveren met een andere toets van Wilcoxon, de Signed-Rank test voor gepaarde data).

De nulhypothese stelt dat de waarnemingen van beide groepen uit één en dezelfde verdeling komen, tegen het alternatief dat de verdelingen niet op dezelfde plek liggen (maar wel dezelfde vorm hebben). De verdeling wordt niet nader gespecificeerd. In het algemeen worden bij parametervrije toetsen minder vooronderstellingen gemaakt dan bij parametrische toetsen (zoals de t-toets). Dat maakt dat de parametervrije toetsen breder inzetbaar zijn, maar vaak minder onderscheidingsvermogen hebben.

Alle waarnemingen (van beide groepen tezamen) worden gerangschikt op grootte en worden van een rangnummer voorzien. Als de nulhypothese juist is, verwacht je dat de waarnemingen van de ene groep gelijkmatig tussen de waarnemingen van de andere groep verspreid liggen, en niet dat alle waarnemingen van de ene groep links liggen (lage rangnummers) en van de andere groep rechts (hoge rangnummers). Vervolgens wordt de som genomen van de rangnummers per groep, zie Tabel 3.2. Hier wordt getoetst of de tumorgrootte van vrouwen boven de 60 jaar uit dezelfde verdeling komt als de tumorgrootte van vrouwen tot en met 60 jaar.

Tabel 3.2 *De Mann-Whitney toets op de variabele tumorgrootte. Vergeleken worden de vrouwen die jonger zijn dan 61 jaar (agegroup = 0) met de vrouwen die ouder zijn dan 60 jaar (agegroup = 1). In SPSS:* ***Analyze – Nonparametric Tests – 2 Independent Samples***

Ranks

	agegroup	N	Mean Rank	Sum of Ranks
Pathologic Tumor Size (cm)	,00	671	606,54	406991,00
	1,00	450	493,09	221890,00
	Total	1121		

Test Statistics[a]

	Pathologic Tumor Size (cm)
Mann-Whitney U	120415,000
Wilcoxon W	221890,000
Z	-5,763
Asymp. Sig. (2-tailed)	,000

a. Grouping Variable: agegroup

Omdat beide groepen niet even groot hoeven zijn, is het niet eerlijk om de sommen van de rangnummers te vergelijken. De kolom met Mean Rank geeft het gemiddelde rangnummer voor de betreffende groep. Als de nulhypothese juist is, zou je verwachten dat deze gemiddelde rangnummers ongeveer gelijk zouden zijn. Dit wordt getoetst in de onderste tabel met een benaderingstoets ("Asymptotic significance"). In dit geval dienen we, bij $\alpha = 0{,}05$, de nulhypothese te verwerpen ($P < 0{,}05$). Aan de gemiddelde rangnummers kunnen we zien dat de vrouwen uit de groep tot en met 60 meer naar rechts liggen en dus een grotere mediane tumor hadden.

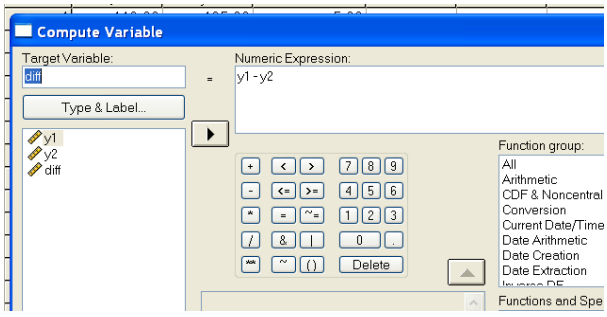
Het is overigens ook mogelijk in SPSS de exacte P-waarde te laten berekenen. Dat is vooral geschikt bij kleine aantallen. Voor grote datasets kost de exacte berekeningswijze veel tijd, terwijl de benadering in dat geval juist goed is.

3.3 Twee afhankelijke (gepaarde) groepen

Indien men een onderzoek heeft met gepaarde (afhankelijke) gegevens, bijvoorbeeld een meting vóór een interventie en een nameting bij dezelfde personen of herhaalde metingen van de bloeddruk bij patiënten aan de nierdialyse, kun je niet doen alsof de gegevens onafhankelijk zijn. De toetsen die besproken zijn in de paragrafen 3.1 en 3.2 zijn derhalve niet geschikt om direct op de data toe te passen. Er is echter een eenvoudige oplossing: door over te gaan op de individuele verschillen tussen de gepaarde waarnemingen krijgen we één groep data en kunnen we de toetsen uit Hoofdstuk 2 gebruiken.
Als voorbeeld gebruiken we een kleine dataset met zes respondenten waarvan tweemaal de systolische bloeddruk is gemeten:

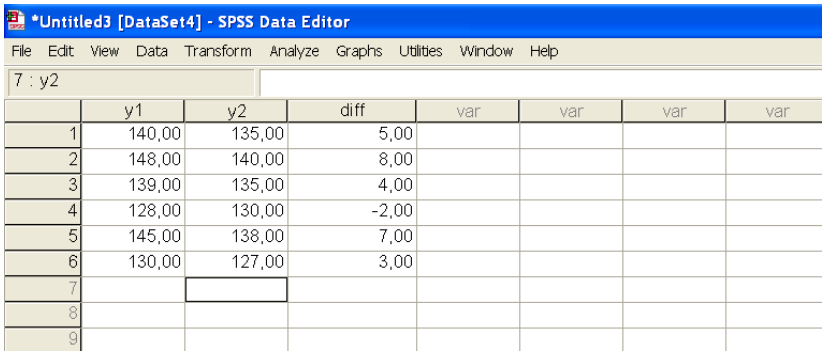
Respondent	Y1	Y2
1	140	135
2	148	140
3	139	135
4	128	130
5	145	138
6	130	127

We vragen ons af of de bloeddrukken op de twee tijdstippen aan elkaar gelijk zijn, tegen het alternatief dat de bloeddruk omlaag is gegaan. Als we de data in SPSS invoeren kunnen we de verschilvariabele laten berekenen met **Transform – Compute –** "diff" bij Target variable, en Y1 – Y2 bij Numeric Expression. Zie Figuur 3.2.



Figuur 3.2 *Het berekenen van een nieuwe variabele in SPSS.*

Dit commando zal tot gevolg hebben dat er een nieuwe variabele wordt berekend die de naam "diff" draagt en het verschil aangeeft tussen de waarden op de variabelen Y1 en Y2. De dataset zal er nu uitzien zoals in Figuur 3.3.



	y1	y2	diff	var	var	var	var
1	140,00	135,00	5,00				
2	148,00	140,00	8,00				
3	139,00	135,00	4,00				
4	128,00	130,00	-2,00				
5	145,00	138,00	7,00				
6	130,00	127,00	3,00				
7							
8							
9							

Figuur 3.3 *De dataset in SPSS (versie 14)*

Als we mogen aannemen dat de variabele diff normaal verdeeld is, wat niet te controleren is met een dataset van slechts zes waarnemingen, kunnen we de t-toets voor één groep toepassen op deze zes getallen. Het toetsen van de nulhypothese dat het gemiddelde van de variabele diff gelijk is aan nul is equivalent aan het toetsen dat het gemiddelde van Y1 gelijk is aan het gemiddelde van Y2. $H_0 : \mu_1 = \mu_2 \Leftrightarrow H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_{diff} = 0$

De resultaten van de t-toets voor één groep staan in Tabel 3.3.

Tabel 3.3 *De één steekproef t-toets op de verschillen van Y1 en Y2. In SPSS:* ***Analyze – Compare Means – One Sample T-Test –*** *diff bij Test Variable(s)*

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
diff	6	4,1667	3,54495	1,44722

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
diff	2,879	5	,035	4,16667	,4465	7,8869

Het gemiddelde verschil wordt geschat op 4,17 mm Hg, wat significant verschilt van 0 (bij een significantieniveau = 0,05). Het 95 % BI voor het verschil loopt van 0,45 tot 7,89 mm Hg. Deze toets is alleen valide als de variabele diff normaal verdeeld is!

Het is echter niet noodzakelijk om eerst de verschilvariabel "diff" te laten berekenen. Je kunt ook een zogenaamde **gepaarde t-toets** op de variabelen Y1 en Y2 laten uitvoeren. Zie Tabel 3.4.

Tabel 3.4 *T-Toets voor gepaarde data op Y1 en Y2. In SPSS:* ***Analyze – Compare Means – Paired Samples T-Test –*** *Y1 bij Variable 1, Y2 bij Variable 2 en klik op ▶*

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	y1	138,3333	6	7,96660	3,25235
	y2	134,1667	6	4,87511	1,99025

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	y1 - y2	4,16667	3,54495	1,44722	,44647	7,88686	2,879	5	,035

De resultaten van de gepaarde t-toets in Tabel 3.4 zijn exact dezelfde als van de t-toets voor één groep verschillen in Tabel 3.3. Beide toetsen zijn alleen valide indien *de individuele verschillen* (de variabele "diff") normaal verdeeld zijn.

Indien de verschillen niet normaal verdeeld zijn, dan kunnen we op de variabele "diff" een non-parametrische toets voor één groep uitvoeren (Sign test of Wilcoxon's signed rank test) of, equivalent daaraan, een non-parametrische toets voor gepaarde groepen. Deze geven onderling weer exact dezelfde resultaten.

OPDRACHTEN

1.
Open de file Beweeg.sav, die je in H2 gemaakt hebt. Toets of de groepen verschillen met betrekking tot hun aanvangsgewicht, aannemende dat deze normaal verdeeld zijn. Hoe luidt de nulhypothese?
Wat is de conclusie?

2.
Toets of de groepen van de file Beweeg.sav evenveel gewicht verliezen. Welke toets gebruik je en waarom? Hoe luidt de nulhypothese? Wat is je conclusie?

3.
Toets, voor alle 20 personen samen, of de gewichten op tijdstip 1 en tijdstip 2 gemiddeld gelijk zijn. Gebruik een gepaarde toets. Welke toets gebruik je en waarom? Hoe luidt de nulhypothese? Wat is je conclusie?

Hoofdstuk 4 Meer dan twee onafhankelijke groepen

Indien er meer dan twee onafhankelijke groepen zijn, zou je paarsgewijze vergelijkingen kunnen doen (groep 1 met groep 2, groep 1 met groep 3, etc) en herhaalde malen de theorie van hoofdstuk 3 toepassen. Deze aanpak heeft echter verschillende bezwaren. Ten eerste is het theoretisch beter – en efficiënter – om alle informatie in één analyse mee te nemen. Een tweede probleem is dat van de **kanskapitalisatie**. Als je verscheidene toetsen uitvoert met een significantieniveau $\alpha = 0{,}05$, dan is de overall-alfa, de kans dat je in minstens één van je toetsen ten onrechte een nulhypothese verwerpt, groter dan 0,05.

> Stel dat je 5 groepen met elkaar wilt vergelijken. Dan moet je $\binom{5}{2} = 10$ toetsen uitvoeren. Je overall-alfa wordt dan 1 – de kans dat je geen enkele fout van de eerste soort maakt = $1-(1-\alpha)^{10} = 1-0{,}95^{10} = 1-0{,}599 = 0{,}401$

In plaats van diverse toetsen willen we een overall-toets voor de nulhypothese dat er geen verschil is tussen meer dan twee onafhankelijke groepen.

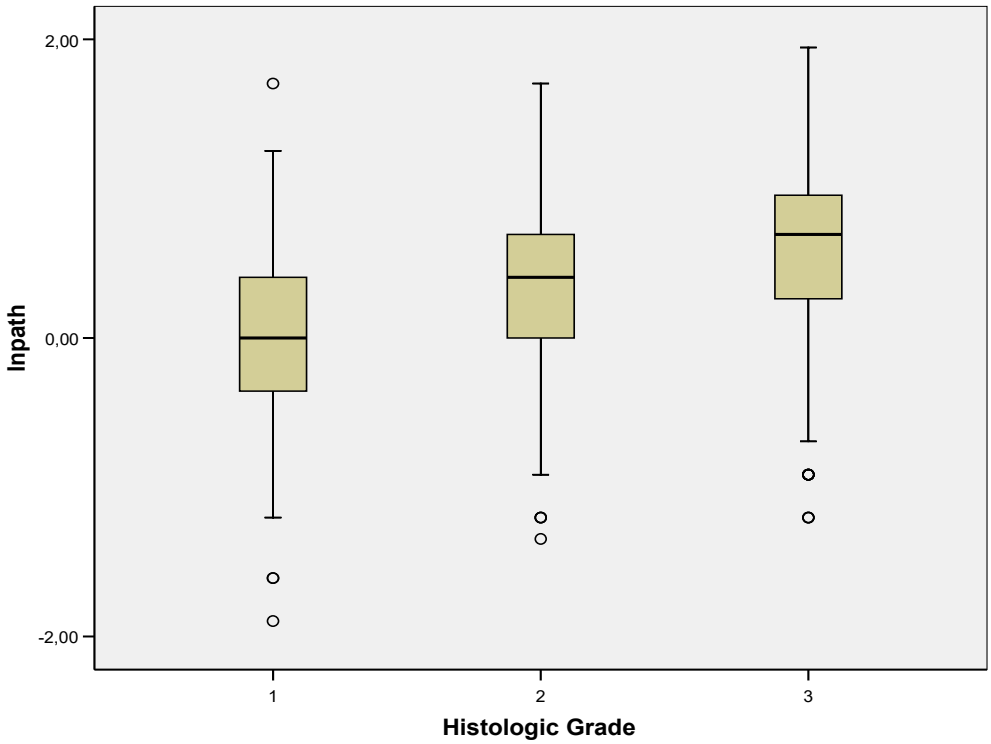
4.1 Een normaal verdeelde responsievariabele

Toetsen
Bij een normaal verdeelde variabele gebruiken we als maat voor ligging het gemiddelde. Als we willen toetsen of er verschillen zijn tussen de *k* groepen, luidt de nulhypothese dat er geen verschillen zijn in de groepsgemiddelden:

H0: $\mu_1 = \mu_2 = ... = \mu_k$

tegen het alternatief dat er minstens twee groepen van elkaar verschillen.
De toets die we hiervoor gebruiken staat bekend onder de naam F-toets, vernoemd naar de statisticus Sir Ronald Fisher. Deze wordt uitgevoerd met een zogenaamde "**One-Way ANOVA**". ANOVA is een afkorting van het Engelse ANalysis Of VAriance, in het Nederlands: **variantieanalyse**. Op het eerste oog lijkt dat misschien vreemd; we zijn geïnteresseerd in de gemiddelden maar we analyseren de varianties. De achterliggende gedachte is de volgende: als de nulhypothese juist is, en alle waarnemingen komen uit (normale) verdelingen met hetzelfde gemiddelde, dan zal de spreiding *tussen* de steekproefgemiddelden van de groepen niet veel anders zijn dan je op grond van de spreiding *binnen* de groepen mag verwachten. In de ANOVA wordt de variantie tussen de groepen vergeleken met de variantie binnen de groepen.
Als voorbeeld kijken we naar de logaritmisch getransformeerde tumorgrootte in de dataset Breastcancer.sav. De ANOVA wordt uitgevoerd op drie groepen gebaseerd op histologische gradering.
In Figuur 4.1 zien we dat de verdelingen van de responsievariabele in de drie groepen redelijk symmetrisch zijn en er enig verschil lijkt in de mediane log-tumorgrootte. De spreiding in de groepen lijkt tamelijk homogeen.
Voor het uitvoeren van een One-Way ANOVA zijn er drie voorwaarden

1. de responsievariabele is normaal verdeeld
2. de waarnemingen zijn onderling onafhankelijk
3. de spreiding in de groepen is gelijk



Figuur 4.1 *Boxplot van de logaritme van tumorgrootte ingedeeld naar histologische gradering. In SPSS:* ***Graphs – Legacy Dialogs – Boxplot – Simple***

Aan de eerste twee voorwaarden is hier voldaan, de laatste lijkt ook geen probleem. Maar we zullen voor de zekerheid met **Levene's toets** de nulhypothese toetsen dat de drie varianties aan elkaar gelijk zijn, H0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$. Vergelijk Levene's toets bij de gepoolde t-toets (hoofdstuk 3.1).
Eerst een stukje beschrijvende statistiek:

Tabel 4.1 *Beschrijvende statistiek van de logaritme van tumorgrootte, als onderdeel van de One-way ANOVA. In SPSS:* ***Analyze – Compare Means – One-Way ANOVA – Options - Descriptives***

Descriptives

lnpath

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	77	-,0034	,64241	,07321	-,1492	,1424	-1,90	1,70
2	484	,3527	,55995	,02545	,3027	,4027	-1,35	1,70
3	313	,5720	,58858	,03327	,5065	,6375	-1,20	1,95
Total	874	,3998	,59951	,02028	,3600	,4396	-1,90	1,95

Als we de gemiddelden van tabel 4.1 terug transformeren naar de oorspronkelijke schaal vinden we respectievelijk 1,00 , 1,42 en 1,77 kubieke centimeter. Hoe hoger de gradering, hoe groter de gemiddelde tumor, maar we weten nog niet of deze verschillen significant zijn.

In tabel 4.2 zien we dat de nulhypothese van gelijke varianties niet wordt verworpen (P = 0,86 > α). Dus aan de voorwaarde van gelijke spreidingen in de groepen is voldaan.

Tabel 4.2 *Levene's toets voor homogeniteit van varianties. In SPSS:* ***Analyze – Compare Means – One-Way ANOVA – Options – Homogeneity of variance test***

Test of Homogeneity of Variances

lnpath

Levene Statistic	df1	df2	Sig.
,147	2	871	,863

Nu aan alle voorwaarden is voldaan, mag de ANOVA uitgevoerd worden. Zie tabel 4.3.

Tabel 4.3 *(One-Way) ANOVA van de logaritme van tumorgrootte voor drie groepen op basis van de histologische gradering. In SPSS:* ***Analyze – Compare Means – One-Way ANOVA***

ANOVA

lnpath

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	22,874	2	11,437	34,245	,000
Within Groups	290,889	871	,334		
Total	313,762	873			

In de tweede kolom van Tabel 4.3 wordt de totale kwadraten som van de responsievariabele Y ($\sum (Y_i - \overline{Y})^2$), een maat voor de spreiding, in twee delen gesplitst. In de derde kolom staan de vrijheidsgraden (df = degrees of freedom) en de getallen in de vierde kolom (de "mean squares") zijn berekend door de kwadratensommen uit kolom 2 te delen door het bijbehorende aantal vrijheidsgraden. Als de nulhypothese waar is, zijn beide mean squares schattingen voor één en dezelfde variantie. De F-waarde in de vijfde kolom is verkregen door de variantieschatting tussen de groepen (11,44) te delen door de variantieschatting binnen de groepen (0,33). Als de nulhypothese waar is, verwachten we dat de F-waarde ongeveer 1 is. Hoe groter F is, hoe meer dat een aanwijzing is dat de gemiddelden in de populaties verschillen. De P-waarde ("Sig.", van significance) is de kans om een F-waarde van 34,245 of groter te vinden als de nulhypothese juist is.
Indien de P-waarde groter is dan het gekozen significantieniveau α (meestal 0,05), dan wordt de nulhypothese niet verworpen en concluderen we dat er, op grond van deze steekproef, geen significante verschillen zijn tussen de gemiddelden. Is P echter kleiner dan of gelijk aan α (zoals in Tabel 4.3), dan verwerpen we de nulhypothese en trekken we de conclusie dat er minstens twee gemiddelden significant verschillen.
We weten echter nog niet welke groepen significant van elkaar verschillen en welke niet. Er zijn vele procedures bedacht om meerdere groepen met elkaar te vergelijken terwijl het risico op kanskapitalisatie in de hand wordt gehouden. Één van deze zogenaamde **Post-hoc** of "**multiple comparisons**" methodes is de **Bonferroni** correctie. De gedachte achter deze procedure is de volgende: alle groepen worden paarsgewijs met elkaar vergeleken met behulp van een t-toets. Er zijn echter twee aanpassingen: ten eerste wordt een schatting van de standaarddeviatie gebruikt op grond van de informatie van *alle* groepen en ten tweede wordt per toets een kleiner significantieniveau gebruikt. Bij de Bonferroni correctie wordt de

overall-alfa (bijvoorbeeld 0,05) gedeeld door het aantal toetsen dat men wenst uit te voeren. Indien er vier groepen zijn en alle groepen moeten met elkaar vergeleken worden, zijn er 6 toetsen. Per toets wordt in dat geval een significantieniveau 0,05 / 6 ≈ 0,0083 gehanteerd. Als er veel groepen, en dus veel vergelijkingen, zijn, is de Bonferroni methode erg conservatief. Er zijn vele andere Post-hoc methoden maar het zou te ver leiden om die hier uitvoerig te bespreken.
In Tabel 4.4 kunnen we zien wat de resultaten van de Bonferroni test zijn voor ons voorbeeld.

Tabel 4.4. *Bonferroni Post-hoc procedure na een significante ANOVA. In SPSS:* ***Analyze – Compare Means – One-Way ANOVA – Post-Hoc - Bonferroni***

Multiple Comparisons

Dependent Variable: lnpath
Bonferroni

(I) Histologic Grade	(J) Histologic Grade	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-,35607*	,07090	,000	-,5261	-,1860
	3	-,57539*	,07351	,000	-,7517	-,3991
2	1	,35607*	,07090	,000	,1860	,5261
	3	-,21932*	,04192	,000	-,3199	-,1188
3	1	,57539*	,07351	,000	,3991	,7517
	2	,21932*	,04192	,000	,1188	,3199

*. The mean difference is significant at the .05 level.

De α per toets is hier 0,05 / 3 = 0,0167. Een lastig getal om je P-waarde mee te vergelijken. SPSS lost het anders op: de P-waarden in Tabel 4.4 zijn reeds met 3 vermenigvuldigd, zodat we deze "Bonferroni gecorrigeerde" P-waarden gewoon met 0,05 kunnen vergelijken.
Uit Tabel 4.4 blijkt dat alle groepen onderling significant verschillen. Conclusie: hoe hoger de gradering, hoe groter de gemiddelde tumor.

Relatie ANOVA en t-toets
De ANOVA (voor meer dan twee onafhankelijke groepen) is te beschouwen als een natuurlijke uitbreiding van de t-toets voor twee onafhankelijke groepen. Als je een ANOVA toepast op twee groepen vind je exact dezelfde P-waarde als bij de t-toets. Je kunt ook aantonen dat de F-waarde van de ANOVA gelijk is aan het kwadraat van de t-waarde uit de t-toets.

4.2 Een niet normaal verdeelde responsievariabele

Als de responsievariabele niet normaal verdeeld is of als de varianties in de groepen niet gelijk zijn, mogen we de ANOVA niet gebruiken. Bij het probleem met twee onafhankelijke groepen bleek er een parametervrije "vervanger" voor de t-toets: de Mann-Whitney toets (zie 3.2). Zo is er bij het probleem met meer dan twee onafhankelijke groepen waar we de ANOVA niet kunnen toepassen ook een uitbreiding van deze Mann-Whitney toets: de **Kruskal-Wallis** toets. Ook hier wordt de nulhypothese getoetst dat alle waarnemingen uit één en dezelfde verdeling komen en is de berekeningswijze gebaseerd op rangnummers. Zie Tabel 4.5 voor een voorbeeld.
De P-waarde is kleiner dan α (0,05), dus de nulhypothese dat alle gegevens uit dezelfde verdeling komen moet worden verworpen. Maar welke groepen zijn significant verschillend?

Helaas kent SPSS geen Post-hoc procedure die toegepast kan worden na een significante Kruskal-Wallis toets. We zullen zelf paarsgewijze Mann-Whitney toetsen moeten toepassen en onze overall-alfa zelf in de gaten moeten houden.

Tabel 4.5 *De Kruskal-Wallis toets. In SPSS:* ***Analyze – Nonparametric Tests – K Independent Groups***

Ranks

	Histologic Grade	N	Mean Rank
Pathologic Tumor Size (cm)	1	77	273,68
	2	484	413,84
	3	313	514,39
	Total	874	

Test Statistics[a,b]

	Pathologic Tumor Size (cm)
Chi-Square	65,972
df	2
Asymp. Sig.	,000

a. Kruskal Wallis Test

b. Grouping Variable: Histologic Grade

OPDRACHTEN

1.
Toets of de rassen in de file lowbwt.sav hetzelfde gemiddelde geboortegewicht hebben. Zo nee, waar zitten de verschillen?

2.
Toets of er verschillen zijn tussen de rassen met betrekking tot het aantal bezoeken aan de dokter in het eerste trimester van de zwangerschap. Welke toets(en) gebruik je en waarom? Wat zijn je conclusies?

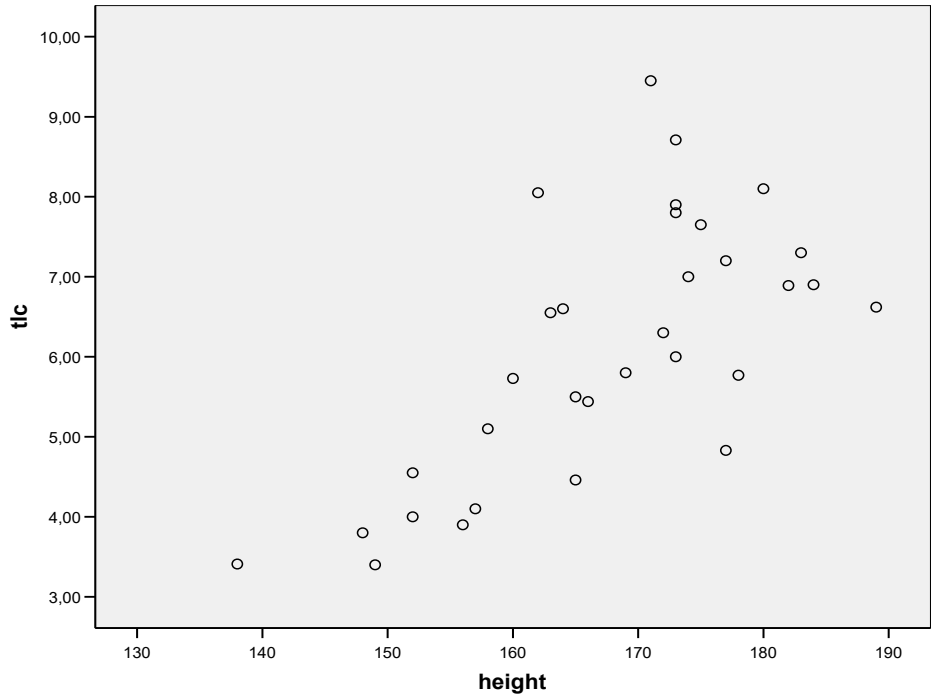
Hoofdstuk 5. Correlatie en lineaire regressie

In dit hoofdstuk kijken we naar de relatie van een **continue responsievariabele** met één of meerdere verklarende variabelen. Ter illustratie gebruiken we de datasets totlc.sav en lowbwt.sav. In 5.1 wordt de samenhang van twee continue variabelen (tlc en lengte) bekeken met behulp van lineaire correlatie waarna in 5.2 de theorie van de enkelvoudige lineaire regressie wordt besproken. In 5.3 zullen we zien hoe met behulp van meervoudige lineaire regressie de relatie tussen de responsievariabelen en meerdere verklarende variabelen onderzocht kan worden.

5.1 Correlatie tussen twee continue variabelen

De sterkte van de samenhang tussen twee continue variabelen wordt vaak uitgedrukt met behulp van een **correlatiecoëfficiënt**. De correlatiecoëfficiënten die in deze tekst besproken worden geven de sterkte van de lineaire samenhang weer in een getal tussen -1 en 1. Deze gestandaardiseerde weergave maakt het mogelijk om diverse samenhangen tussen variabelen die op verschillende schalen gemeten zijn te kunnen vergelijken.
Correlatie is een symmetrische relatie, dat wil zeggen dat de correlatie tussen de variabelen X en Y precies hetzelfde is als de correlatie tussen Y en X.

5.1.1 Correlatie tussen twee normaal verdeelde variabelen

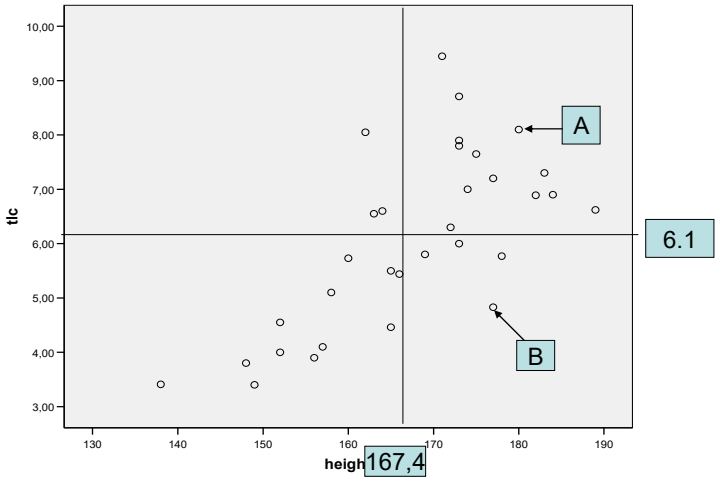


Figuur 5.1 *Strooiingsdiagram ("scatterplot") van tlc versus lengte. In SPSS:* ***Graphs – Legacy Dialogs – Scatter/Dot – Simple Scatter***

In Figuur 5.1 zien we een **strooiings-** of **spreidingsdiagram** van de variabelen tlc en lengte. Op het eerste oog lijkt er een relatie zichtbaar, hoe langer de respondent, hoe groter de totale

long capaciteit. Als deze relatie perfect lineair zou zijn zouden alle punten op een rechte lijn moeten liggen. Dat is hier duidelijk niet het geval, er zitten aanzienlijke verschillen tussen respondenten met dezelfde lengte, zie bijvoorbeeld de mensen met een lengte van ongeveer 173 cm.
Een maat voor lineaire samenhang is de zogenaamde **covariantie**. Deze geeft aan hoe de spreiding van de ene variabele, X, samenhangt ("co-varieert") met de spreiding van de andere variabele, Y. In een formule:

f4 $\text{covariantie(X, Y)} = \dfrac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$

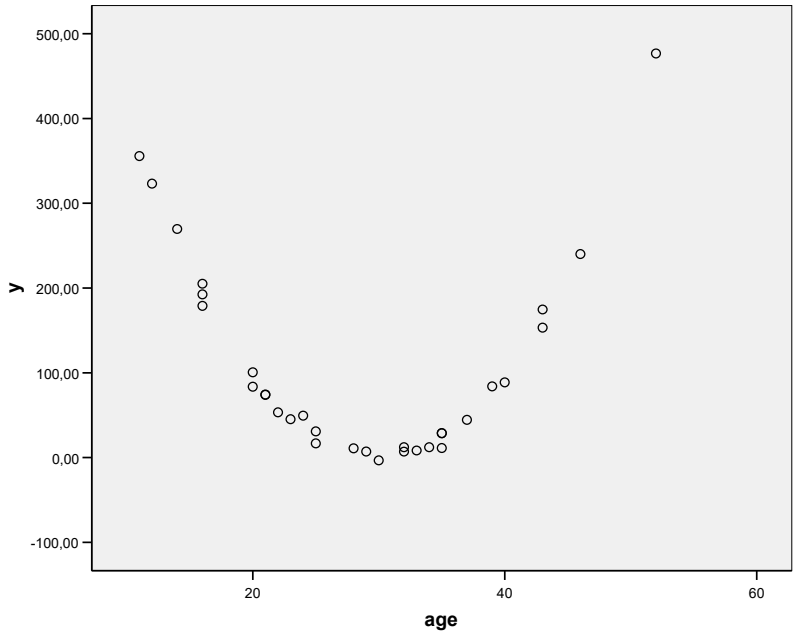


Figuur 5.2 *Toelichting op de covariantie*

In Figuur 5.2 wordt geïllustreerd wat de gedachte achter de covariantie is. In de figuur zijn twee lijnen aangegeven ter hoogte van de gemiddelde lengte (167,4 cm) en de gemiddelde tlc (6,1 liter). Het snijpunt van deze lijnen wordt het zwaartepunt van de dataset genoemd. Een punt dat zich rechtsboven het zwaartepunt bevindt, zoals de gegevens van persoon A, levert een positieve bijdrage aan de covariantie. Immers deze persoon heeft een lengte boven de gemiddelde lengte (en dus een positieve deviatie $X_i - \overline{X}$) en ook een positieve deviatie wat betreft de tlc. Het product van deze deviaties levert dus een positieve bijdrage aan de som in formule f4. Omdat negatief maal negatief positief is, geldt voor de punten linksonder het zwaartepunt ook dat zij een positieve bijdrage leveren aan deze som. Punten linksboven en rechtsonder het zwaartepunt (zoals de gegevens van persoon B) hebben een negatieve bijdrage aan de covariantie; zij doen als het ware afbreuk aan de positieve relatie.
De covariantie is afhankelijk van de eenheid waarin wordt gemeten. Als we de lengte in meters zouden weergeven, is de covariantie 100 maal zo klein als de covariantie die we krijgen als we de lengte in centimeters weergeven.
Pearson's correlatiecoëfficiënt komt aan dit probleem tegemoet door de covariantie te delen door het product van de standaarddeviaties van de betrokken variabelen. Dit levert een getal dat per definitie tussen -1 en 1 ligt.

f5 Pearson's correlatiecoëfficiënt $r = \dfrac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$

Als r = -1 liggen alle punten precies op een dalende lijn, bij r = 1 liggen alle punten precies op een stijgende lijn. Als r = 0 is er geen sprake van een lineair verband. Dit wil niet zeggen dat er geen sprake kan zijn van samenhang tussen X en Y; mogelijk is er sprake van niet-lineaire samenhang.



Figuur 5.3 *Voorbeeld van niet-lineaire samenhang*

In Figuur 5.3 zien we hier een voorbeeld van, Pearson's r is hier bijna 0 (-0,08) maar er is wel degelijk sprake van samenhang tussen beide variabelen, in dit geval een kwadratisch verband.

> Het is mogelijk om een betrouwbaarheidsinterval op te stellen voor de populatie correlatiecoëfficiënt, die aangegeven wordt met de Griekse letter rho: ρ. Voor de berekeningswijze verwijzen we naar Altman.

Indien we willen toetsen of er sprake is van een lineair verband tussen twee continue, normaal verdeelde, variabelen zal de nulhypothese luiden dat er geen lineair verband is in de populatie, tegen het alternatief dat er wel zo'n verband is:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Voor de gegevens uit Figuur 5.1 zien we het resultaat in Tabel 5.1.

Tabel 5.1 *Pearson's correlatiecoëfficiënt van lengte en tlc. In SPSS:* ***Analyze – Correlate - Bivariate***

Correlations

		height (cm)	total lung capacity (l)
height (cm)	Pearson Correlation	1	,696**
	Sig. (2-tailed)		,000
	N	32	32
total lung capacity (l)	Pearson Correlation	,696**	1
	Sig. (2-tailed)	,000	
	N	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

De correlatiecoëfficiënt wordt geschat op 0,696 en de nulhypothese wordt bij een significantieniveau van 0,05 verworpen. We kunnen stellen dat er *op dit domein* sprake lijkt van een lineair verband tussen lengte en tlc. Of dit de beste manier is om de samenhang weer te geven blijkt echter niet uit deze analyse.

5.1.2 Correlatie tussen twee variabelen, niet beide normaal verdeeld

Indien niet beide variabelen normaal verdeeld zijn, zijn de toets en de berekeningswijze voor het betrouwbaarheidsinterval die gebruikt worden bij Pearson's r niet valide. We kunnen wel gebruik maken van **Spearman's correlatiecoëfficiënt**. Deze gebruikt dezelfde formule als Pearson, alleen wordt deze niet toegepast op de waargenomen data, maar op de rangnummers van deze gegevens (zie Swinscow of Altman).

5.2 Enkelvoudige lineaire regressie

Bij regressietechnieken maken we onderscheid tussen een **responsievariabele** (ook wel de afhankelijke variabele genoemd) en één of meerdere verklarende variabelen (ook wel onafhankelijke variabelen genoemd). In grafische presentaties zorgen we er dan voor dat de responsievariabele op de verticale as staat afgebeeld. We willen onderzoeken in hoeverre de responsievariabele verklaard en / of voorspeld kan worden met behulp van de verklarende variabele(n). In deze paragraaf bestuderen we de **enkelvoudige lineaire regressie**, dat wil zeggen dat er slechts één verklarende variabele is.

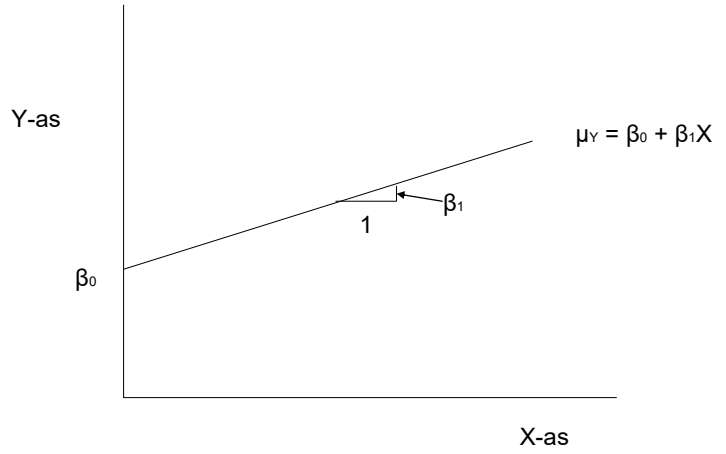
5.2.1 Een continue verklarende variabele

Bij lineaire regressie met één continue verklarende variabele gaan we er van uit dat het verband tussen de variabelen X en Y in de populatie voor een individu met index *i* beschreven kan worden met het volgende model:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

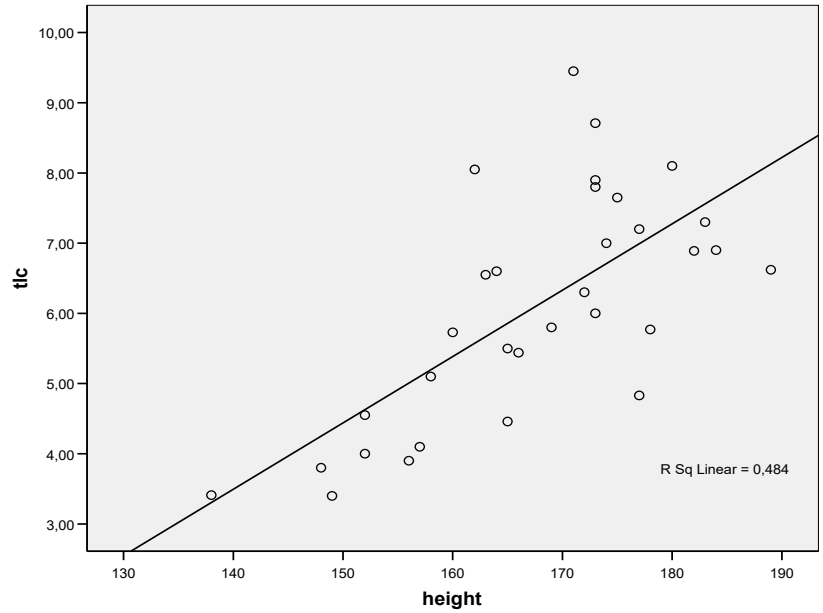
Hierin zijn de β's de coëfficiënten van de lijn, β_0 wordt het **intercept** (of de constante) genoemd en β_1 de **regressiecoëfficiënt** (of de **richtingscoëfficiënt** van de lijn). e_i wordt het **residu** genoemd en het model veronderstelt dat deze afkomstig is uit een normale verdeling met gemiddelde 0 en een constante spreiding die onafhankelijk is van de waarde van X.

We zouden – equivalent hieraan – ook kunnen zeggen dat de gemiddelde Y-waarden voor vaste X-waarden op een rechte lijn liggen: $\mu_{Y|X} = \beta_0 + \beta_1 X$ en de individuele waarden normaal verdeeld liggen met gelijke spreiding rond deze gemiddelden. Zie Figuur 5.4.
In de meeste gevallen zullen we geïnteresseerd zijn in de regressiecoëfficiënt die aangeeft hoeveel de responsievariabele Y gemiddeld verandert als X één eenheid groter wordt.



Figuur 5.4 *De vergelijking van de lijn in de populatie*

Op grond van onze steekproefgegevens proberen we de populatieparameters zo goed mogelijk te schatten. We kunnen in SPSS vragen om de best passende lijn in het spreidingsdiagram te tekenen, zie Figuur 5.5.



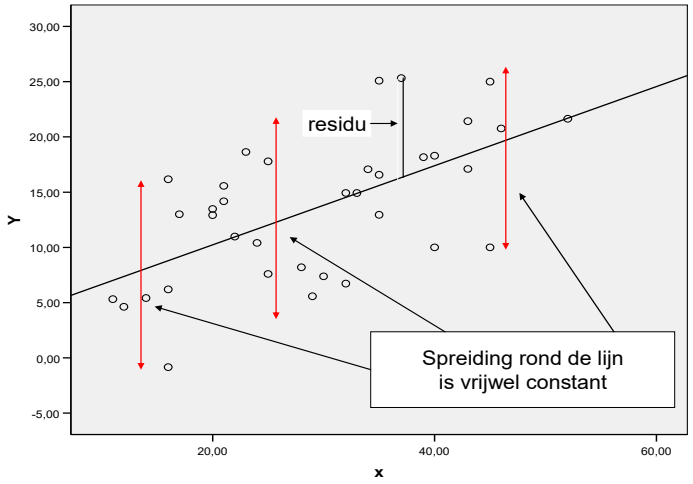
Figuur 5.5 *Strooiingsdiagram ("scatterplot") van tlc versus lengte met regressielijn. In SPSS:* ***Graphs – Legacy Dialogs – Scatter/Dot – Simple Scatter*** *en vervolgens in de output dubbelklikken op de grafiek. In het nieuwe menu, de Chart Editor,* ***Elements – Fit Line at Total***

Wat wordt precies verstaan onder "de best passende lijn"? SPSS berekent de coëfficiënten van de lijn zodanig dat de som van de kwadraten van de afstanden van alle punten tot die lijn zo

klein mogelijk is. We kunnen SPSS vragen om de coëfficiënten van de lijn voor ons te schatten en te toetsen of er een significant verband is tussen tlc en lengte, maar eerst kijken we naar de **voorwaarden** voor een lineaire regressie:

1. de waarnemingen zijn onderling onafhankelijk
2. het verband tussen Y en X is lineair
3. de residuen zijn normaal verdeeld
4. de spreiding van de residuen is homogeen

De eerste voorwaarde valt alleen te controleren als er informatie is over de wijze waarop de gegevens zijn verzameld; een persoon mag niet meerdere keren in de dataset voorkomen. Met de tweede voorwaarde wordt bedoeld dat als op grond van een spreidingsdiagram al duidelijk is dat het verband tussen Y en X niet lineair is (zie Figuur 5.3) het geen zin heeft een lineaire regressie uit te voeren. De laatst twee voorwaarden betreffen de residuen, de afstanden in verticale richting van de punten tot de lijn. Deze dienen normaal verdeeld te zijn en de spreiding moet even groot zijn, ongeacht de waarde van X, zie Figuur 5.6.



Figuur 5.6 *Voorbeeld van homogene spreiding van de residuen*

De voorwaarden 3 en 4 zijn pas te checken als de regressielijn al berekend is. We zullen later zien hoe we dat in SPSS kunnen bewerkstelligen.
Aannemende dat aan de eerste twee voorwaarden voldaan is, gaan we in SPSS een lineaire regressie laten uitvoeren van tlc op lengte. In de uitvoer zien we drie relevante tabellen, zie Tabel 5.2.

Tabel 5.2 *Uitvoer van (enkelvoudige) lineaire regressie van ltc op lengte. In SPSS:* ***Analyze – Regression – Linear –*** *vul in tlc bij "Dependent" en height bij "Independent(s)"*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,696[a]	,484	,467	1,18610

a. Predictors: (Constant), height (cm)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	39,559	1	39,559	28,119	,000[a]
	Residual	42,205	30	1,407		
	Total	81,764	31			

a. Predictors: (Constant), height (cm)

b. Dependent Variable: total lung capacity (l)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-9,742	2,993		-3,255	,003
	height (cm)	,095	,018	,696	5,303	,000

a. Dependent Variable: total lung capacity (l)

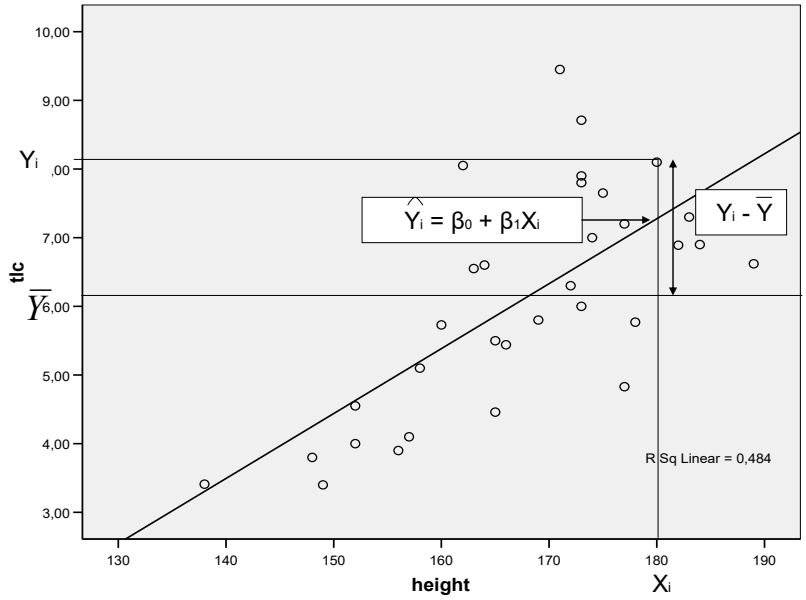
In de bovenste tabel kunnen we lezen hoe goed dit lineaire model de spreiding van de tlc verklaart, in de middelste tabel wordt getoetst of dit model significant is en in de onderste tabel staan de schattingen van de coëfficiënten van de lijn. Voor een meer gedetailleerdere beschouwing beginnen we in de onderste tabel.

Onderste tabel van Tabel 5.2

De vergelijking van de lijn in de populatie wordt geschat als tlc = -9,742 + 0,095*lengte.
De geschatte regressiecoëfficiënt heeft de volgende interpretatie: we schatten dat een groep mensen met een bepaalde lengte, vergeleken met een andere groep die allen 1 cm kleiner zijn, gemiddeld 0,095 l meer totale longcapaciteit hebben.
Hoe interpreteren we de -9,742 van de constante? Dit zou de geschatte tlc zijn van een persoon met lengte 0 cm. Dat is natuurlijk onzin. Je kunt een regressievergelijking niet extrapoleren naar een gebied waar geen waarnemingen gedaan zijn; de kleinste persoon in deze steekproef was 138 cm.
Behalve de geschatte coëfficiënten zien we ook de standaardfouten van deze schattingen, twee t-waarden (verkregen door de schattingen te delen door hun standaardfouten) en de bijbehorende P-waarden. Hier worden twee nulhypothesen getoetst. De bovenste t- toets toetst de nulhypothese dat de constante nul is, met andere woorden dat de lijn door de oorsprong gaat. Deze toets zal ons zelden interesseren; ook in dit voorbeeld is het tamelijk onzinnig om te toetsen of een persoon van lengte 0 een tlc gelijk aan 0 heeft. We zijn benieuwd naar de relatie tussen tlc en lengte en dat wordt getoetst in de tweede t-toets. De nulhypothese luidt

$H_0 : \beta_1 = 0$. (Als de richtingscoëfficiënt van de lijn gelijk is aan 0, loopt de lijn horizontaal. Als de lijn horizontaal loopt, maakt het blijkbaar niet uit of je klein dan wel groot bent; de verwachte tlc blijft gelijk. Als $\beta_1 = 0$ is er dus geen verband tussen lengte en tlc).
In dit voorbeeld zien we een P < 0,001, duidelijk kleiner dan een significantieniveau van 0,05, en we verwerpen de nulhypothese van geen verband. Blijkbaar hebben langere mensen een grotere totale long capaciteit, een conclusie die ons niet onlogisch voorkomt.
Tot slot staat in de onderste tabel een kolom met kop Beta waar we in paragraaf 5.3 op terug komen.

Middelste tabel van Tabel 5.2
In de middelste tabel van Tabel 5.2 wordt de nulhypothese dat het totale model niets verklaart getoetst met behulp van variantieanalyse. Het "totale model" bestaat hier maar uit één verklarende variabele, lengte, en het wekt dan ook geen verbazing dat de P-waarde van deze toets gelijk is aan de P-waarde van de t-toets van de regressiecoëfficiënt. Dit is per definitie zo in elke enkelvoudige lineaire regressie, in paragraaf 5.3 zullen we zien dat bij een meervoudige lineaire regressie deze equivalentie niet langer opgaat.
Wat gebeurt er in deze ANOVA? De totale spreiding van de responsievariabele Y (in dit geval tlc), uitgedrukt in de totale kwadratensom SS = $\sum (Y_i - \overline{Y})^2$ wordt in twee delen gesplitst, een verklaard deel en een onverklaard deel. Waarom heeft in Figuur 5.7 de persoon van 1.80 meter een totale long capaciteit (Y_i) die boven het gemiddelde ($\overline{Y}$) ligt? Voor een deel kunnen we dat verklaren: omdat deze persoon langer is dan gemiddeld en er een positief verband is tussen lengte en tlc zal de voorspelde tlc van deze persoon ($\widehat{Y}_i$) boven het gemiddelde liggen. Maar waarom de tlc van deze persoon *boven* de regressielijn ligt en niet *op* de regressielijn kunnen we niet verklaren. Dus van het verschil $Y_i - \overline{Y}$ kunnen we het deel $\widehat{Y}_i - \overline{Y}$ wel verklaren, maar het deel $Y_i - \widehat{Y}_i$ niet. Zie ter illustratie Figuur 5.7.



Figuur 5.7 *Opsplitsing van de Sum of Squares*

Je kunt aantonen dat geldt: $\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y})^2 = \sum (Y_i - \widehat{Y}_i)^2 + \sum (\widehat{Y}_i - \overline{Y})^2$

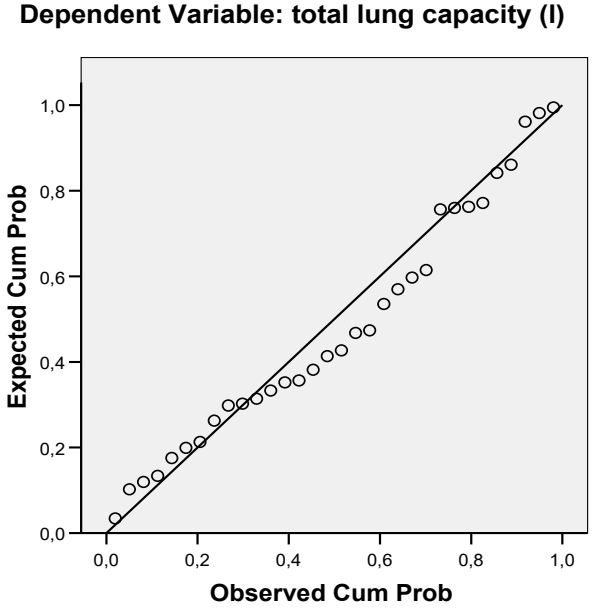
Hiermee laten we zien dat de totale kwadratensom in het linkerlid te splitsen is in een onverklaard en een verklaard deel in het rechterlid. In de ANOVA tabel van de lineaire regressie (zie Tabel 5.2) gebeurt dat in de tweede kolom. De totale kwadratensom (81,8) wordt opgedeeld in een verklaard deel (Sum of Sqaures of Regression, 39,6) en een onverklaard deel (Sum of Squares of Residual, 42,2). Als we deze kwadratensommen delen door het bijbehorende aantal vrijheidsgraden uit kolom 3, dan krijgen we de Mean Squares uit kolom 4. Als de nulhypothese juist is ("het model verklaart niets"), dan zijn beide Mean Squares schattingen van dezelfde variantie (vergelijk ook de theorie in paragraaf 4.1). Dit wordt getoetst met de F-toets. De F-waarde in kolom 5 is verkregen door de Mean Square of Regression te delen door de Mean square of Residual. In de rechtse kolom staat de bijbehorende P-waarde die zoals gebruikelijk wordt vergeleken met een $\alpha = 0{,}05$. Aangezien $P < 0{,}05$ verwerpen we de nulhypothese dat het model niets verklaart, hetgeen we ook al wisten uit de t-toets voor de regressiecoëfficiënt.

Bovenste tabel van Tabel 5.2

De **R Square** uit de derde kolom is verkregen door de Sum of Squares of Regression (39,6) te delen door de totale kwadratensom (81,8). Het verkregen getal is dus de proportie verklaarde spreiding van tlc. In dit geval verklaart lengte dus 0,48*100% = 48 % van de spreiding van tlc. Dit is een ietwat optimistische schatting, de adjusted R Square in de vierde kolom is realistischer. De R in kolom 2 is de wortel van de R Square en is, bij enkelvoudige lineaire regressie, gelijk aan de absolute waarde van Pearson's correlatiecoëfficiënt tussen X en Y. Tot slot de "Standard Error of the Estimate", dit is de wortel van de Mean Square of Residual en is te interpreteren als de gemiddelde absolute afstand van de punten tot de regressielijn, dus de gemiddelde absolute grootte van de residuen.

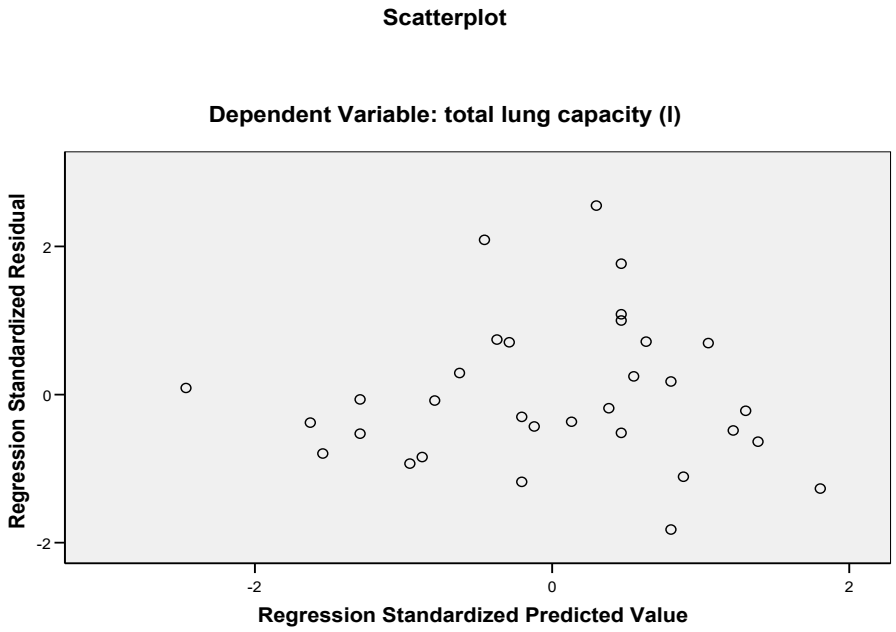
Controle voorwaarden met betrekking tot de residuen

Om te controleren of de uitgevoerde toetsen valide zijn moeten we de twee voorwaarden betreffende de residuen controleren: zijn de residuen normaal verdeeld en is de spreiding van de residuen homogeen? Dit doen we aan de hand van twee grafieken: een **P-P Plot** (vergelijkbaar met een Q-Q Plot, zie Hoofdstuk 1) voor de normaliteit en een spreidingsdiagram voor de homogeniteit.



Figuur 5.8 *P-P Plot van de residuen behorende bij de lineaire regressie van tlc op lengte. In SPSS:* ***Analyze – Regression – Linear – Plots – Normal probability plot***

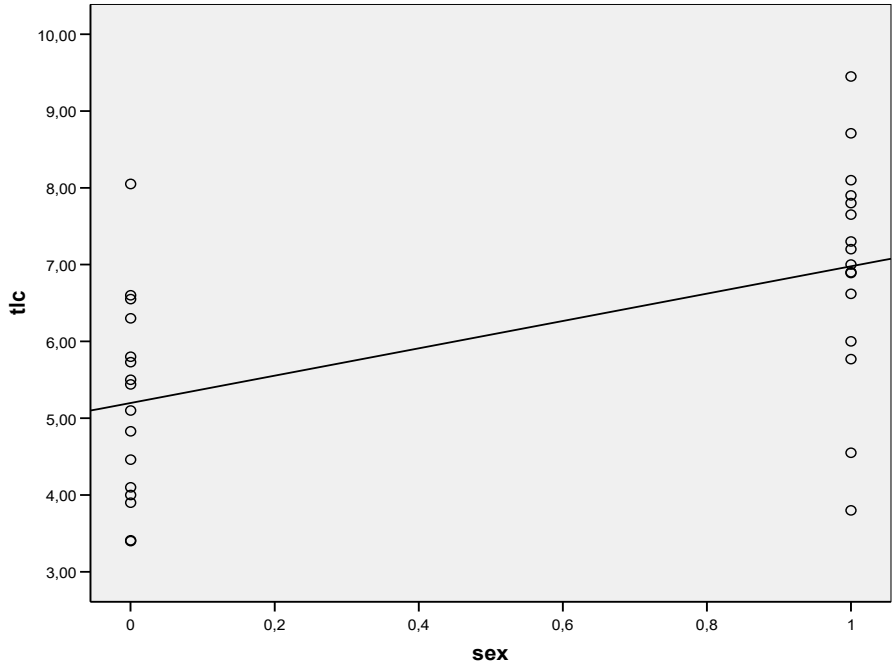
Hoewel de punten in de P-P Plot in Figuur 5.8 niet precies op de lijn vallen, lijkt de verdeling redelijk normaal.
In Figuur 5.9 zijn de gestandaardiseerde residuen afgezet tegen de gestandaardiseerde voorspelde waarde. De spreiding van de residuen lijkt niet al te homogeen, er is meer spreiding bij de hogere voorspelde waarden dan bij de kleinere, wat er op lijkt te duiden dat de tlc met dit model nauwkeuriger te voorspellen is voor kleine mensen dan voor grote. Een dergelijke constatering kan aanleiding zijn om je af te vragen of het onderhavige model wel het juiste is. Misschien is het beter een transformatie te gebruiken of behoeft dit model nog extra verklarende variabelen. Als het laatste het geval is, gaan we over op een meervoudige lineaire regressie, het onderwerp van paragraaf 5.3.



Figuur 5.9 *De gestandaardiseerde residuen afgezet tegen de gestandaardiseerde voorspelde waarde bij de lineaire regressie van tlc op lengte. In SPSS:* ***Analyze – Regression – Linear – Plots –*** *vul in: *ZRESID bij Y en *ZPRED bij X*

5.2.2 Een dichotome verklarende variabele

We willen onderzoeken of er verschil is in gemiddelde tlc tussen mannen en vrouwen. We hebben in hoofdstuk 3 gezien hoe we dat kunnen aanpakken. Als tlc normaal verdeeld is, kunnen we de vraag beantwoorden met behulp van een t-toets voor onafhankelijke groepen. Maar kunnen we vraag ook binnen de context van lineaire regressie beantwoorden? Met andere woorden: is het ook toegestaan een enkelvoudige lineaire regressie uit te voeren als de verklarende variabele dichotoom is?
In Figuur 5.10 hebben we SPSS de best passende lijn laten trekken door het spreidingsdiagram van tlc en geslacht. Als we SPSS vragen om de lineaire regressie uit te voeren krijgen we tabellen zoals we die in de vorige subparagraaf zagen, zie Tabel 5.3. Maar wat gebeurt er eigenlijk?



Figuur 5.10 *Regressielijn voor tlc en geslacht (0 = vrouw, 1 = man). In SPSS:* ***Graphs – Legacy Dialogs – Scatter/Dot – Simple Scatter –*** *tlc op Y-as, Geslacht op X-as, dubbelklikken op de output en in de Chart Editor* ***Elements – Fit line at total***

Tabel 5.3 *Deel van de uitvoer van lineaire regressie van tlc op geslacht. In SPSS:* ***Analyze – Regression - Linear***

Coefficients[a]

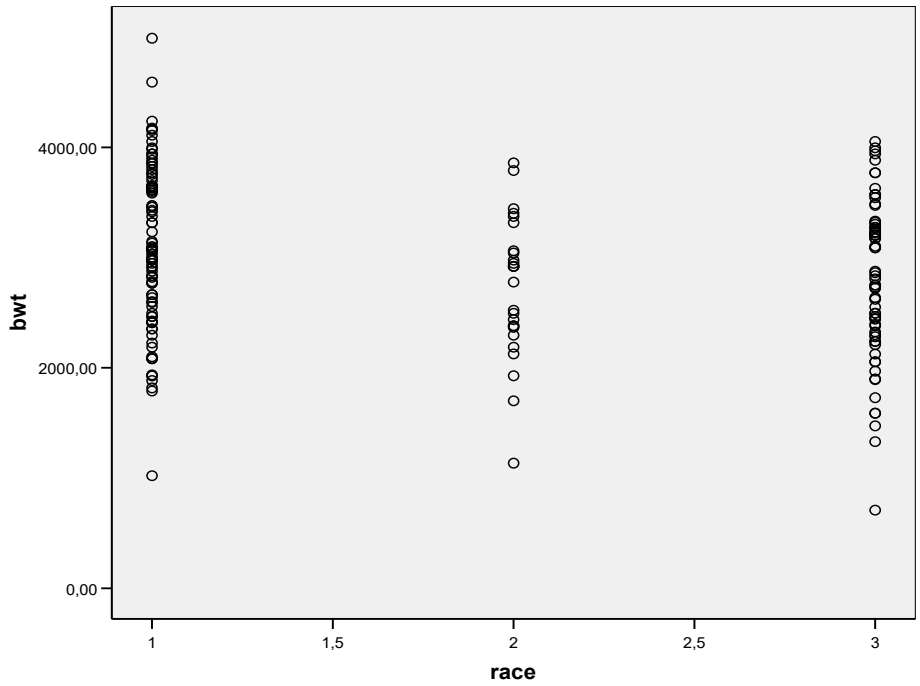
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5,198	,343		15,160	,000
	sex	1,779	,485	,557	3,669	,001

a. Dependent Variable: total lung capacity (l)

De best passende lijn door de punten is die lijn waarvoor de residuen (de afstanden van de punten tot de lijn) absoluut gezien zo klein mogelijk zijn. De lijn zal daarom door de gemiddelde tlc-score van de vrouwen en door de gemiddelde tlc-score van de mannen gaan. De regressiecoëfficiënt is de richtingscoëfficiënt (rc) van de lijn. Maar als we vanuit de gemiddelde tlc van de vrouwen één eenheid naar rechts gaan, komen we bij de mannen uit. Hoeveel moeten we dan omhoog om weer op de lijn te komen? Het verschil tussen de gemiddelde tlc van mannen en de gemiddelde tlc van vrouwen! Vervolgens wordt er getoetst of deze rc gelijk is aan nul, dus of de gemiddelde tlc van vrouwen gelijk is aan de gemiddelde tlc van mannen. Met andere woorden: hier wordt een t-toets voor twee onafhankelijke groepen uitgevoerd. De P-waarde in tabel 5.3 is inderdaad exact dezelfde als de P-waarde in Tabel 3.1. Ook zien we dat het intercept van de lineaire regressie gelijk is aan het gemiddelde van de mannen en de regressiecoëfficiënt gelijk aan het verschil tussen de mannen en de vrouwen.
Het is dus volstrekt legitiem een dichotome verklarende variabele in een lineaire regressie te gebruiken, mits aan alle voorwaarden is voldaan. (Merk op dat de voorwaarden voor het uitvoeren van de gepoolde t-toets dezelfde zijn als voor de lineaire regressie).

5.2.3 Een nominale verklarende variabele (met meer dan twee categorieën)

Een dichotome verklarende variabele is blijkbaar geen probleem in een lineaire regressie, maar hoe zit het met een nominale variabele met meer dat twee categorieën?
Als voorbeeld bekijken we in de file over de geboortegewichten de responsievariabele geboortegewicht en de verklarende variabele ras met drie categorieën, zie Figuur 5.11.



Figuur 5.11 *Scatterplot van geboortegewicht tegen ras (1 = blank, 2 = zwart, 3 = ander)*

Mogen we de variabele ras zondermeer in een lineaire regressievergelijking stoppen? Het antwoord is nee. De gebruikte codering is willekeurig terwijl het gebruikte computerprogramma deze codes als numerieke waarden zal gebruiken; dat wil zeggen dat er van uit gegaan wordt de afstand tussen 1 en 2 (blank en zwart) dezelfde is als tussen 2 en 3 (zwart en "ander"). Dus het verschil tussen blank en "ander" wordt twee maal zo groot als het verschil tussen blank en zwart. Als de onderzoeker een andere (eveneens willekeurige) code zou gebruiken, zouden de resultaten van de analyse verschillen.
Indien we een nominale variabele met meer dat twee categorieën in een lineaire regressie willen gebruiken, moeten we onze toevlucht zoeken tot hulpvariabelen, zogenaamde **dummy's**. Er zijn meerdere manieren van dummycodering, maar de meest gebruikte is de zogenaamde **referentiegroep codering**. Één groep wordt als referentie gekozen en de andere groepen worden met behulp van een dummy vergeleken met de referentiegroep. We kiezen in het bovenstaand voorbeeld groep 1 (blank) als referentiegroep en maken twee hulpvariabelen, d1 en d2, aan met de volgende codes:

Ras	d1	d2
1 (blank)	0	0
2 (zwart)	1	0
3 (ander)	0	1

Vervolgens voeren we een lineaire regressie uit met beide dummy's als verklarende variabelen, zie Tabel 5.4. (We voeren hier eigenlijk een meervoudige lineaire regressie uit; meer theorie daarover volgt in paragraaf 5.3)

Tabel 5.4 *Deel van de uitvoer van een lineaire regressie met twee dummy variabelen*

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5070608	2	2535303,816	4,972	,008[a]
	Residual	9E+007	186	509927,124		
	Total	1E+008	188			

a. Predictors: (Constant), d2, d1

b. Dependent Variable: birth weight in grams

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3103,740	72,882		42,586	,000
	d1	-384,047	157,874	-,182	-2,433	,016
	d2	-299,725	113,678	-,197	-2,637	,009

a. Dependent Variable: birth weight in grams

In deze lineaire regressieanalyse wordt het volgende lineaire model geschat:

$$geboortegewicht = \beta_0 + \beta_1 * d1 + \beta_2 * d2$$

Als we de gebruikte codes invullen :

Blank: geboortegewicht = 3104
Zwart: geboortegewicht = 3104 + -384 = 2720
Ander: geboortegewicht = 3104 + -300 = 2804

In de bovenste tabel van Tabel 5.4 zien we een significante ANOVA (P = 0,008), wat inhoudt dat de nulhypothese "het model verklaart niets" verworpen moet worden.
Uit de onderste tabel van Tabel 5.4 is af te leiden dat beide dummy's significant zijn, dus dat de rassen zwart en ander significant van blank verschillen. Over een eventueel verschil tussen zwart en ander kunnen we op grond van deze analyse niets zeggen.
De uitvoer van de lineaire regressie met de twee dummy's die samen staan voor de variabele ras in de bovenste tabel van Tabel 5.4 komt overeen met de uitvoer die wordt verkregen door een One-Way ANOVA toe te passen met responsievariabele geboortegewicht en verklarende variabele ras, zoals we hebben geleerd in paragraaf 4.1, zie Tabel 5.5.
Na afloop van de regressieanalyse dienen de residuen nog gecontroleerd te worden op normaliteit en homogeniteit van variantie. Dit zijn dezelfde voorwaarden die ook in een ANOVA gecheckt worden.
(Merk op dat je in de lineaire regressie met behulp van de dummies twee groepen met de referentiegroep, in dit geval blank, vergelijkt. Met de Bonferroni procedure van paragraaf 4.1 vergeleken we alle groepen met elkaar).

Tabel 5.5 *One-Way ANOVA met responsievariabele geboortewicht en verklarende variabele ras. In SPSS:* ***Analyze – Compare Means – One-Way ANOVA – Options - Descriptive***

Descriptives

birth weight in grams

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
white	96	3103,740	727,72424	74,27304	2956,2889	3251,1902	1021	4990
black	26	2719,692	638,68388	125,2562	2461,7223	2977,6623	1135	3860
other	67	2804,015	721,30115	88,12096	2628,0758	2979,9541	709	4054
Total	189	2944,656	729,02242	53,02858	2840,0486	3049,2636	709	4990

ANOVA

birth weight in grams

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5070608	2	2535303,816	4,972	,008
Within Groups	9E+007	186	509927,124		
Total	1E+008	188			

We zien dus dat het toegestaan is om een nominale variabele in een regressievergelijking op te nemen als we gebruik maken van hulpvariabelen. Het aantal hulpvariabelen (dummy's) dat je nodig hebt is gelijk aan het aantal categorieën – 1.

Er blijken duidelijke relaties tussen verschillende statistische technieken. Een lineaire regressie met een categorische verklarende variabele blijkt gelijk aan een t-toets (bij twee groepen) of een ANOVA (bij meer dan twee groepen). Je kunt je afvragen waarom we een regressie zouden gebruiken als er ook alternatieven zijn. Het antwoord is: bij een regressie heb je de mogelijkheid om rekening te houden met de mogelijke invloed van andere verklarende variabelen op de responsievariabele. Als we willen weten of er verschillen zijn tussen mannen en vrouwen wat betreft hun totale longcapaciteit terwijl we rekening willen houden met (of corrigeren voor) hun lengte dan kan dat niet met een t-toets. Dat kan wel als we gebruik maken van een (lineaire) regressie, zoals we zullen zien in paragraaf 5.3.

5.3 Meervoudige lineaire regressie

In de vorige paragraaf hebben we gezien dat de totale long capaciteit op positieve wijze afhangt van de lengte en dat mannen gemiddeld een grotere tlc hebben dan vrouwen. Dit verschil was afgerond 1,8 liter. Maar we weten ook dat mannen gemiddeld langer zijn dan vrouwen. Wat is het effect van het geslacht als we rekening houden met het lengte verschil? In een regressieanalyse kun je meerdere verklarende variabele in je model gebruiken, bij lineaire regressie veronderstellen we dat de relatie tussen de responsievariabele Y en de verklarende variabelen X_1, X_2, … , X_k te schrijven is als

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + e$$

Waarin e het normaal verdeelde residu is.
In Tabel 5.6 wordt de uitvoer van de **meervoudige lineaire regressie** van tlc op lengte en geslacht gegeven. Deze zal worden vergeleken met de enkelvoudige regressie van tlc op lengte in Tabel 5.2.

Tabel 5.6 *Meervoudige lineaire regressie van tlc op lengte en geslacht. In SPSS:* ***Analyze – Regression – Linear –*** *tlc invullen bij* ***Dependent****, geslacht en lengte bij* ***Independent***

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,724[a]	,524	,491	1,15899

a. Predictors: (Constant), sex, height (cm)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	42,810	2	21,405	15,935	,000[a]
	Residual	38,954	29	1,343		
	Total	81,764	31			

a. Predictors: (Constant), sex, height (cm)

b. Dependent Variable: total lung capacity (l)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-7,034	3,403		-2,067	,048
	height (cm)	,076	,021	,560	3,607	,001
	sex	,772	,496	,241	1,556	,131

a. Dependent Variable: total lung capacity (l)

De uitvoer van de meervoudige lineaire regressie in Tabel 5.6 vertoont grote overeenkomst met die van de enkelvoudige lineaire regressie in Tabel 5.2. Een logisch verschil is dat de tabel met geschatte coëfficiënten bij de meervoudige lineaire regressie uitgebreider is.
De relatie tussen tlc en de verklarende variabelen wordt geschat als

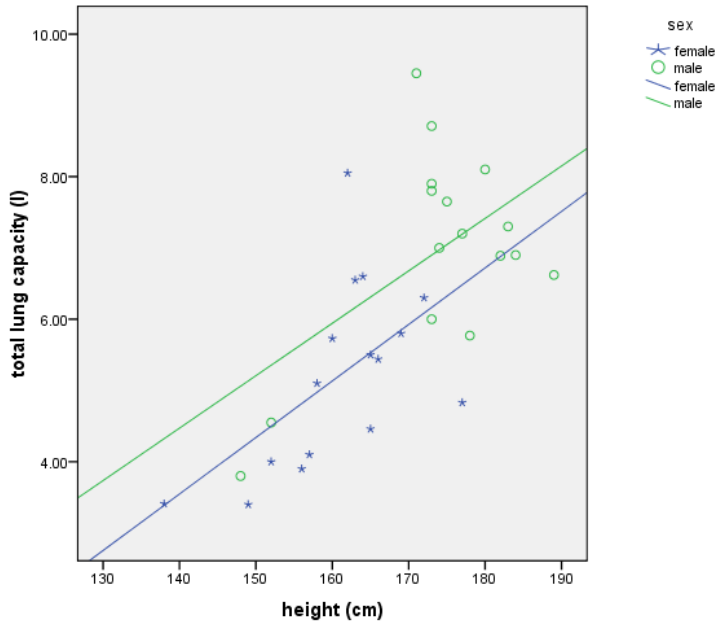
Tlc = -7,03 + 0,08*lengte + 0,77*geslacht

De vrouwen zijn gecodeerd als 0, de mannen als 1, dus per geslacht wordt de vergelijking:

Vrouwen: tlc = -7,03 + 0,08*lengte
Mannen: tlc = -7,03 + 0,08*lengte + 0,77 = -6,26 + 0,08*lengte

De regressiecoëfficiënt voor lengte wordt in deze analyse geschat als 0,08 terwijl deze 0,095 was in de enkelvoudige lineaire regressie. Het geschatte effect van geslacht (bij gelijke lengte) is nu 0,77 liter, in de enkelvoudige regressie 1,8 liter. We zien dat het voor de schattingen van de coëfficiënten duidelijk verschil uitmaakt of er rekening wordt gehouden met andere variabelen of niet. Zo ook voor de t-toetsen. De t-toetsen die hier uitgevoerd worden (in de onderste tabel van Tabel 5.6) zijn zogenaamde **partiële t-toetsen**. Hierbij wordt de nulhypothese getoetst dat de coëfficiënt van de betreffende variabele gelijk is aan 0, gegeven de invloed van de andere verklarende variabelen op de responsievariabele. We zien dat, bij een $\alpha = 0,05$, lengte wel significant is als we rekening houden met het geslacht, maar geslacht niet langer significant is, gegeven de lengte. De mannen hebben in deze steekproef een

gemiddelde tlc die 0,77 liter hoger is dan die van de vrouwen (bij gelijke lengte) maar dat is geen significant verschil. In de populatie zou dat verschil ook nul kunnen zijn.
In Figuur 5.12 zijn de geschatte regressielijnen in het spreidingsdiagram getrokken.



Figuur 5.12 *Spreidingsdiagram van tlc tegen lengte met afdruksymbool geslacht. Getrokken zijn regressielijnen voor vrouwen (onderste lijn) en mannen apart op grond van de geschatte coëfficiënten in Tabel 5.6*

In de middelste tabel van Tabel 5.6 wordt de nulhypothese getoetst "dat het hele model niets verklaart". Netter geformuleerd $H_0 : \beta_1 = \beta_2 = 0$. Deze nulhypothese wordt verworpen (P < 0,001). We zien dat er geen correspondentie meer is tussen de P-waarde van de ANOVA en de P-waarde(n) van de t-toets(en), zoals die wel bestond bij de enkelvoudige lineaire regressie. Dat is logisch, want er wordt nu ook een andere nulhypothese getoetst bij de ANOVA dan bij de individuele t-toetsen.

De R Square in de bovenste tabel geeft aan welk deel van de spreiding van de responsievariabele verklaard wordt door de verklarende variabelen samen. R wordt de meervoudige (multiple) correlatiecoëfficiënt genoemd en is de correlatiecoëfficiënt tussen de waargenomen en de voorspelde Y-waarde. Lengte en geslacht samen verklaren ongeveer 49 % van de spreiding van tlc, voor lengte alleen was dat bijna 47 %. Geslacht voegt dus niet veel informatie toe.

In de onderste tabel van Tabel 5.6 staat een kolom genaamd beta. Deze bevat schattingen van gestandaardiseerde coëfficiënten. Omdat de gewone regressiecoëfficiënt afhankelijk is van de gebruikte eenheid, kunnen deze coëfficiënten niet gebruikt worden om de variabelen onderling te vergelijken in de mate waarin ze invloed hebben op de responsievariabele. De gestandaardiseerde coëfficiënt beta wordt berekend door de regressiecoëfficiënt te vermenigvuldigen met de geschatte standaarddeviatie van de betreffende verklarende variabele en te delen door de geschatte standaarddeviatie van de responsievariabele. De beta van lengte is 0,56 en dat vertelt ons dat een verschil van één standaarddeviatie voor de lengte een verschil van 0,56 standaarddeviatie van tlc tot gevolg heeft.

Op grond van bovenstaande analyse kunnen we de conclusie trekken dat het effect van geslacht op tlc in de univariate analyse niet correct geschat werd omdat we geen rekening hielden met de lengte We noemen lengte in zo'n geval een **confounder**.

Opbouw van een regressiemodel

Bij veel medisch wetenschappelijk onderzoek zijn er meerdere verklarende variabelen. Hoe bouw je een regressiemodel op? Een aan te raden procedure is:

1. Voer univariate analyses uit met al je potentiële verklarende variabelen
2. Selecteer de belangrijkste variabelen op grond van theoretische overwegingen en de gevonden P-waarden (neem een ruime α, bijvoorbeeld 0,25)
3. Bouw je model stap voor stap op, begin bij je belangrijkste (meest significante) variabele
4. Voeg variabelen toe zolang het model significant beter wordt
   Je kunt ook "backward" te werk gaan; begin met alle variabelen in je model en verwijder één voor één de minst significante variabele tot je een model overhoudt met alleen maar significante of theoretisch relevante variabelen.
5. Als alle relevante variabelen in het model zitten, kijk dan eventueel naar interacties (zie hieronder)
6. Controleer de residuen van het uiteindelijke model op normaliteit en homogeniteit van de spreiding

Interactie in een lineaire regressie

We willen onderzoeken wat het effect van gewicht van de moeder en het rookgedrag tijdens de zwangerschap op het geboortegewicht van een kind is. De meervoudige lineaire regressie leidt tot de volgende schattingen:
Geboortegewicht = 2500 + 9,34*gewicht -270*roken (zie Tabel 5.7)

Tabel 5.7 *Deel van de uitvoer van lineaire regressie van geboortegewicht op gewicht (van de moeder) en rookgedrag tijdens de zwangerschap. In SPSS:* ***Analyze – Regression - Linear***

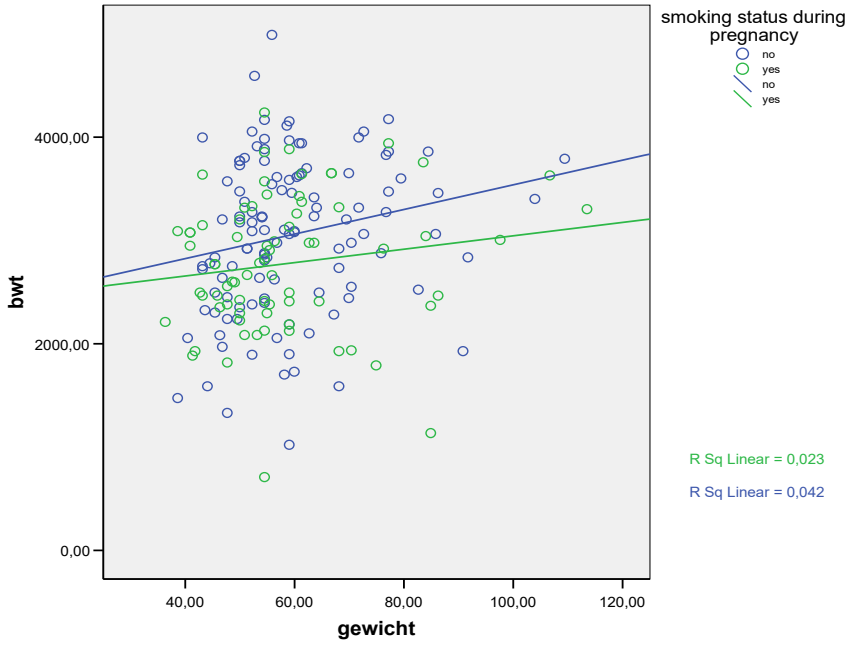
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2500,174	230,833		10,831	,000
	gewicht	9,336	3,722	,178	2,508	,013
	smoking status during pregnancy	-270,013	105,590	-,181	-2,557	,011

a. Dependent Variable: birth weight in grams

Beide variabelen zijn significant bij een α = 0,05. In dit model gaan we er van uit dat de effecten van de verklarende variabelen op het geboortegewicht onafhankelijk van elkaar zijn. Dat wil zeggen dat het effect op het geboortegewicht van het kind van het rookgedrag bij minder zware moeders het zelfde effect heeft als het effect van het rookgedrag bij zwaardere moeders. Met andere woorden, we gaan er – grafisch gezien – van uit dat de regressielijnen parallel lopen.
Als de verklarende variabelen niet onafhankelijk van elkaar zijn, spreken we van **interactie** of **effectmodificatie**. Om een indruk te krijgen of de regressielijnen parallel lopen kunnen we

een spreidingsdiagram maken met het rookgedrag als afdruksymbool en SPSS vragen per subgroep de best passende regressielijn te trekken, zie Figuur 5.13.



Figuur 5.13 *Spreidingsdiagram van het geboortegewicht van het kind tegen het gewicht van de moeder met regressielijnen per rokers groep. In SPSS:**Graphs – Legacy Dialogs – Scatter/Dot – Simple Scatter –** geboortegewicht op de Y-as, gewicht op de X-as, roken bij **Set Markers By***

De onderste regressielijn is die der rooksters. De grafiek suggereert dat het negatief effect van roken op het geboortegewicht bij zwaardere moeders sterker is dan bij minder zware moeders. Om te toetsen of de richtingscoëfficiënten van de twee regressielijnen significant verschillen gaan we een meervoudig regressiemodel opstellen dat naast de variabelen gewicht en rookgedrag een derde verklarende variabele kent. Deze derde variabele is een interactieterm tussen gewicht en rookgedrag. Deze interactieterm kan op verschillende manieren gedefinieerd worden, maar het meest gebruikelijk is om het product te nemen van gewicht en roken. Voor dit doel is in SPSS een nieuwe variabele aangemaakt:

Intgewsmoke = gewicht * smoke

We gaan de coëfficiënten schatten van de vergelijking

$$geboortegewicht = \beta_0 + \beta_1 * gewicht + \beta_2 * smoke + \beta_3 * Intgewsmoke$$

Tabel 5.8 *Lineaire regressie van geboortegewicht op gewicht van de moeder, rookgedrag tijdens de zwangerschap en de interactie van deze variabelen*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2347,507	312,717		7,507	,000
	gewicht	11,905	5,143	,227	2,315	,022
	smoking status during pregnancy	47,867	451,163	,032	,106	,916
	intgewsmoke	-2,456	3,388	-,223	-,725	,470

a. Dependent Variable: birth weight in grams

Als we de schattingen uit Tabel 5.8 invullen in bovenstaande vergelijking en ons realiseren dat de variabele ”smoke” gecodeerd was als 0 voor niet-rooksters en 1 voor rooksters krijgen we:

voor niet-rooksters geboortegewicht = 2348 + 11,9*gewicht + 48*0 – 2,5*gewicht*0 =
= 2348 + 11,9*gewicht

voor rooksters geboortegewicht = 2348 + 11,9*gewicht + 48*1 - 2,5*gewicht*1 =
= 2396 + 9,4*gewicht

We zien dat de richtingscoëfficiënt voor rooksters kleiner is dan die voor niet-rooksters. Dit betekent dat voor zwaardere moeders de lijnen verder uit elkaar lopen dan voor minder zware moeders, zoals ook in Figuur 5.13 te zien is. Maar zijn de verschillen significant? Met andere woorden is er sprake van significante interactie tussen de variabelen gewicht en rookgedrag? Dit kunnen we zien in Tabel 5.8 bij de t-toets van de interactieterm. In dit geval is de P-waarde 0,47 en dus groter dan $\alpha = 0,05$. De nulhypothese dat $\beta_3 = 0$, er is geen interactie-effect, wordt niet verworpen. Blijkbaar maakt het, in deze dataset, niet uit wat het gewicht van de moeder is voor het effect van het rookgedrag tijdens de zwangerschap op het geboortewicht van het kind. De interactieterm kan dus weer verwijderd worden uit het regressiemodel.
Uit Tabel 5.8 zou je – foutief – de conclusie kunnen trekken dat rookgedrag geen invloed heeft op geboortegewicht. Deze variabele heeft immers een P-waarde = 0,916, wat beduidend groter is dan 0,05. We dienen ons echter te realiseren dat het een partiële t-toets betreft, dat wil zeggen dat er rekening wordt gehouden met de andere variabelen in het getoetste model. Één van die andere variabelen is de interactieterm, die natuurlijk gecorreleerd is met de variabele rookgedrag. De interactieterm is gelijk aan 0 indien roken 0 is of als gewicht 0 is. Dus de coëfficiënt van gewicht is de richtingscoëfficiënt van gewicht voor de niet-rokers en de coëfficiënt van roken is het effect van roken *voor mensen met gewicht 0*. Dit laatste is niet echt bruikbaar, maar uit Tabel 5.7 weten we dat rookgedrag wel degelijk invloed heeft. Indien een interactieterm significant is, horen de bijbehorende hoofdeffecten ook in het model. Deze hoofdeffecten zijn dan niet langer individueel te interpreteren en moeten steeds in combinatie met elkaar bekeken worden.

Het vergelijken van twee geneste modellen

We noemen twee modellen **genest** als alle variabelen uit het kleinste model ook onderdeel zijn van het grootste model. Soms komt het voor dat we een model willen uitbreiden met meer dan één variabele tegelijkertijd, denk bijvoorbeeld aan het toevoegen van twee dummies voor het toetsen van het effect van een nominale variabele met drie categorieën.
We beginnen met een lineaire regressie met één verklarende variabele, zie Figuur 5.9.

Tabel 5.9 *Lineaire regressie van het gewicht van de moeder (tijdens de laatste menstruatie) op de leeftijd van de moeder*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.180[a]	.032	.027	13.69256

a. Predictors: (Constant), age of mother (years)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1174.965	1	1174.965	6.267	.013[a]
	Residual	35059.923	187	187.486		
	Total	36234.887	188			

a. Predictors: (Constant), age of mother (years)

b. Dependent Variable: gewicht

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	47.972	4.491		10.681	.000
	age of mother (years)	.472	.188	.180	2.503	.013

a. Dependent Variable: gewicht

Als we willen weten of ras een significante bijdrage levert in de verklaring van het gewicht, gegeven de leeftijd, voegen we twee dummies toe, zie Figuur 5.10. Per dummy wordt met behulp van een t-toets getoetst of de bijbehorende groep significant verschilt van de referentiegroep. In dit geval zien we dat beide groepen (zwart en ander) significant verschillen van blank, maar dat hoeft niet bij iedere nominale variabele het geval te zijn. We hebben nog geen overall-toets om te beoordelen of de variabele ras in zijn totaliteit een significante bijdrage levert aan het model.

Tabel 5.10 *Lineaire regressie van het gewicht van de moeder (tijdens de laatste menstruatie) op de leeftijd van de moeder en het ras*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.340[a]	.116	.101	13.16190

a. Predictors: (Constant), d2, age of mother (years), d1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4186.322	3	1395.441	8.055	.000[a]
	Residual	32048.565	185	173.235		
	Total	36234.887	188			

a. Predictors: (Constant), d2, age of mother (years), d1

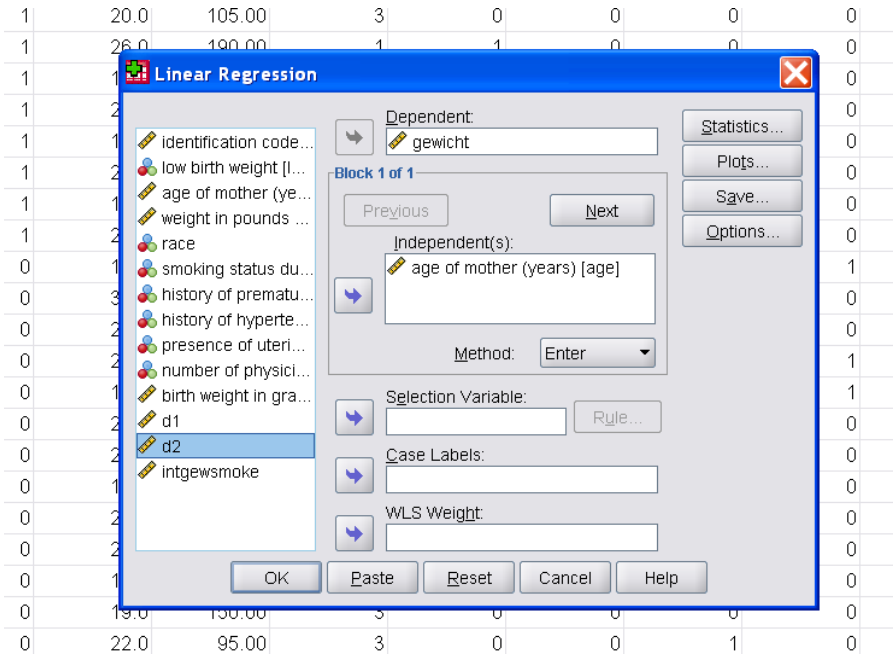
b. Dependent Variable: gewicht

Coefficients[a]

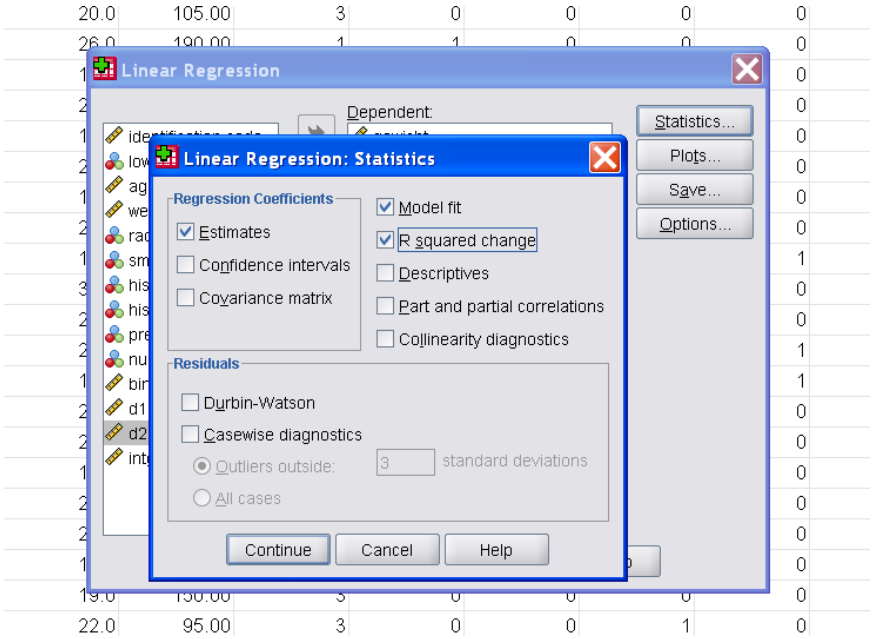
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	48.041	4.696		10.230	.000
	age of mother (years)	.490	.185	.187	2.647	.009
	d1	8.049	2.954	.200	2.725	.007
	d2	-4.532	2.125	-.157	-2.133	.034

a. Dependent Variable: gewicht

We kunnen SPSS vragen om twee geneste modellen met elkaar te vergelijken door in het regressiecommando twee blokken te definiëren: het eerste blok is het kleine model (met alleen leeftijd als verklarende variabele) het tweede blok bevat de variabelen die we gezamenlijk willen toevoegen (in ons geval d1 en d2).
Nadat leeftijd is ingevuld, klik op Next (zie Figuur 5.14) om het volgende blok te definiëren. Vul d1 en d2 in, in het nieuwe blok. Om de twee modellen (het kleinste model bestaat uit de variabelen uit blok 1, het grootste model uit de variabelen van de blokken 1 en 2 samen) met elkaar te kunnen vergelijken bestellen we onder Statistics de “R Square change” die toetst of het grote model significant meer verklaart dan het kleinste model. Zie Figuur 5.15.



Figuur 5.14 *Het definiëren van blokken om geneste modellen met elkaar te kunnen vergelijken. Klik op Next om het volgende blok te definiëren*



Figuur 5.15 *Voor de toets: klik op Statistics en vink R squared change aan*

In Tabel 5.11 zie je een deel van de uitvoer. Gedeeltelijk zien we dezelfde informatie als in de Tabellen 5.9 en 5.10, zoals de significanties en de R-kwadraten van de afzonderlijke modellen. Maar daarnaast wordt er getoetst of de verandering in R-kwadraat (het verklaarde deel van de spreiding van Y) significant is. Het gaat hier om de verandering in R^2 van model 2 ten opzichte van model 1, een verschil van 0,116 – 0,032 = 0,083 (afgerond). Aan de P-waarde (< 0,001) kunnen we zien dat model 2 significant meer verklaart dan model 1.

Tabel 5.11 *Deel van de uitvoer van de lineaire regressie van gewicht van de moeder op leeftijd en ras. Door het definiëren van blokken is het mogelijk verschillende geneste modellen met elkaar te vergelijken.*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.180[a]	.032	.027	13.69256	.032	6.267	1	187	.013
2	.340[b]	.116	.101	13.16190	.083	8.692	2	185	.000

a. Predictors: (Constant), age of mother (years)

b. Predictors: (Constant), age of mother (years), d2, d1

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1174.965	1	1174.965	6.267	.013[a]
	Residual	35059.923	187	187.486		
	Total	36234.887	188			
2	Regression	4186.322	3	1395.441	8.055	.000[b]
	Residual	32048.565	185	173.235		
	Total	36234.887	188			

a. Predictors: (Constant), age of mother (years)

b. Predictors: (Constant), age of mother (years), d2, d1

c. Dependent Variable: gewicht

OPDRACHTEN

1.
Open de file totlc.sav. Maak een spreidingsdiagram van tlc en leeftijd. Welke variabele komt op de Y-as? Lijkt het verband lineair?
Selecteer alle personen jonger dan 26 jaar.
Voer een lineaire regressie uit van tlc op leeftijd. Interpreteer de resultaten, check de voorwaarden en geef je conclusie.

2.
Open lowbwt.sav. Voer een enkelvoudige lineaire regressie uit van geboortegewicht op de variabele ui (geschiedenis van geïrriteerde uterus). Interpreteer de resultaten, check de voorwaarden en geef je conclusie.

3.
Kies op voorhand uit de dataset lowbwt.sav naast de responsievariabele geboortegewicht vijf verklarende variabelen, waaronder ras en gewicht van de moeder. Doel van deze opdracht is om met lineaire regressie een model te vinden dat het geboortegewicht zo goed mogelijk voorspelt.

Voer eerst univariate toetsen uit en maak geschikte grafieken om een indruk van de relaties tussen de verklarende variabelen en geboortegewicht te krijgen.
Bouw je (lineaire) regressiemodel stap voor stap op. Controleer met je uiteindelijke model de voorwaarden en formuleer je conclusies.

Deel 2 Een dichotome responsievariabele

Bij veel medisch wetenschappelijk onderzoek is de belangrijkste uitkomstvariabele niet continu maar **dichotoom** (of ook wel **binair**): de behandeling is wel of geen succes, de patiënt heeft wel of geen bijwerkingen, de respondent is wel of niet overleden. De statistische technieken uit Deel 1 zijn dan niet geschikt om dergelijke gegevens mee te analyseren.
In deel 2 van deze syllabus worden methoden besproken die wel geschikt zijn voor dichotome uitkomstvariabelen.

Hoofdstuk 6 Één groep

Schatten
De overheid overweegt een voorlichtingscampagne gericht op zwangere vrouwen over de gevaren van actief roken voor het ongeboren kind. Maar hoe groot is het probleem eigenlijk? Welk deel van de vrouwen rookt tijdens de zwangerschap?
Er moet hier een onbekende proportie in de populatie geschat worden, die we aan zullen geven met de Griekse letter pi (π). Als we een aselecte steekproef trekken kunnen we de steekproefproportie p gebruiken als puntschatting van de onbekende π. Als we een betrouwbaarheidsinterval willen opstellen voor de proportie rooksters in de populatie, hebben we ook de standaarddeviatie (sd) van de proportie nodig. Het is te bewijzen dat deze sd alleen afhangt van de proportie zelf en de grootte van de steekproef (n). We schatten de sd in de populatie met de volgende formule: $\sqrt{\frac{p(1-p)}{n}}$. Als n groot genoeg is (n > 20) en p niet te klein en niet te groot (np > 5 en n(1 – p) > 5), kan de verdeling van een proportie goed benaderd worden door een normale verdeling en mogen we voor een betrouwbaarheidsinterval waarden uit de z-verdeling gebruiken. Het 95 % BI voor de proportie in de populatie wordt dan:

f6 $$[p - 1{,}96 * \sqrt{\frac{p(1-p)}{n}}; p + 1{,}96 * \sqrt{\frac{p(1-p)}{n}}]$$

In de file lowbwt.sav geeft de variabele “smoke” aan of de moeder tijdens de zwangerschap heeft gerookt. Als we deze dataset als een representatieve steekproef mogen beschouwen van de populatie van alle zwangere vrouwen, kunnen we op grond van deze steekproef een schatting maken van de proportie in de populatie. Van de 189 vrouwen in de steekproef blijken er 74 tijdens de zwangerschap gerookt te hebben. Dat is, afgerond, een proportie van 0,39 ofwel 39 %. Invullen in formule **f6** levert [0,39 – 1,96*0,035 ; 0,39 + 1,96*0,035] = [0,32 ; 0,46].
Als we bedenken dat de variabele smoke gecodeerd is als 1 voor de rooksters en 0 voor de niet-rooksters, kunnen we SPSS het betrouwbaarheidsinterval ook laten berekenen. Het gemiddelde van de enen en nullen levert precies de proportie rooksters op, zie Tabel 6.1.

Tabel 6.1 *Beschrijvende statistiek van "smoke" (deel), proportie rooksters met standaarddeviatie. In SPSS:* ***Analyze – Descriptive Statistics - Explore***

Descriptives

			Statistic	Std. Error
smoking status during pregnancy	Mean		,39	,036
	95% Confidence Interval for Mean	Lower Bound	,32	
		Upper Bound	,46	

Toetsen

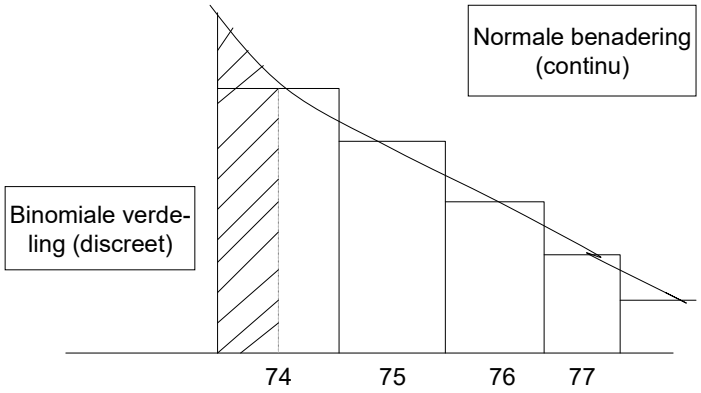
Als het percentage rooksters in de populatie der zwangere vrouwen 25 % of lager is, wil de overheid geen stappen ondernemen. Een onderzoeker die verwacht dat het percentage hoger ligt dan 25 %, of equivalent daaraan dat de proportie hoger ligt dan 0,25, zal de volgende nulhypothese willen toetsen H_0: $\pi = 0{,}25$ tegen het alternatief: H_1: $\pi > 0{,}25$.
Dit is het volgende binomiaal probleem: het aantal rooksters in de steekproef noemen we X. Als de nulhypothese juist is, is X een **binomiaal verdeelde** variabele met n = 189 en $\pi = 0{,}25$. Dit wordt kortweg genoteerd als $X \sim B(189, 0{,}25)$. Als we bovenstaande nulhypothese en alternatief willen toetsen, berekenen we de kans op het aantal rooksters in de steekproef (genoteerd als k) of nog meer, als de nulhypothese juist is. Bij een kleine P-waarde (kleiner dan α) verwerpen we de nulhypothese en besluiten dat significant meer dan 25 % rookt.
In dit geval: $P(X \geq k) = P(X \geq 74)$.

De kans op precies 74 rooksters is $P(X = 74) = \binom{189}{74} 0{,}25^{74} 0{,}75^{115}$

Het is al veel rekenwerk om het aantal combinaties van 74 uit 189 te berekenen, en dan zijn we er nog lang niet. Ook de kans op 75, 76, 77 , … , 189 successen moet nog berekend worden. Al met al 116 kansen.
De binomiale verdeling kan goed worden **benaderd door een normale verdeling** als n groot genoeg is ($n > 20$) en π niet te klein en niet te groot ($n\pi > 5$ en $n(1 - \pi) > 5$). Aan deze voorwaarden is hier voldaan. In plaats van de rechteroverschrijdingskans te bepalen in de Binomiale verdeling door de som te nemen van de oppervlaktes van 116 staven, kijken we naar de oppervlakte van een normale verdeling die "goed past" bij deze binomiale verdeling. We nemen daarvoor de normale verdeling met dezelfde verwachting ("gemiddelde") en dezelfde standaarddeviatie als de te benaderen binomiale verdeling. Van een binomiale verdeling met steekproefaantal n en succeskans π zijn verwachting en standaarddeviatie respectievelijk $n*\pi$ en $\sqrt{n*\pi*(1-\pi)}$. In dit geval is de verwachting 189*0,25 = 47,25 en de sd 5,95.
In plaats van de kans $P(X \geq 74)$ berekenen we van de normaal verdeelde variabele $Y \sim N(47{,}25 , 5{,}95^2)$ een gebied dat overeenkomt met de oppervlakte in de binomiale verdeling, zie Figuur 6.1.
Als we $P(X \geq 74)$ zouden benaderen met $P(Y \geq 74)$ zouden we een stukje missen. Omdat X een discrete variabele is (alleen gehele uitkomsten) en Y een continue variabele sluiten de kansen niet helemaal aan. Bij $P(Y \geq 74)$ missen we het linker deel van $P(X = 74)$, het gearceerde deel in Figuur 6.1. Om de gevraagde oppervlakte beter te benaderen wordt daarom een zogenaamde **continuïteitscorrectie** toegepast: in plaats van $P(Y \geq 74)$ berekenen we $P(Y \geq 74 - 0{,}5) = P(Y \geq 73{,}5)$.

$$P(Y \geq 73{,}5) = P(Z \geq \frac{73{,}5 - 47{,}25}{5{,}95}) = P(Z \geq 4{,}41) < 0{,}001$$

Dit leidt tot de conclusie dat we de nulhypothese verwerpen: significant meer dan 25 % van de vrouwen rookt tijdens de zwangerschap.



Figuur 6.1 *Normale benadering van de binomiale verdeling, met continuïteitscorrectie*

De gebruikte continuïteitscorrectie is vooral van belang bij kleine aantallen; in dit geval zou het voor de P-waarde niets uitgemaakt hebben als we deze correctie achterwege gelaten zouden hebben.
Het is mogelijk om deze test in SPSS te laten uitvoeren, zie Tabel 6.2. SPSS berekent hier de kans met behulp van de normale benadering (éénzijdig), maar het is mogelijk om de exacte P-waarde van de Binomiale verdeling te laten berekenen. Daarvoor dien je het knopje Exact aan te klikken.

Tabel 6.2 *Binomiale test voor de nulhypothese dat de proportie rooksters gelijk is aan 0,25. In SPSS:* ***Analyze – Nonparametric Tests – Binomial –*** *Smoke in "Test Variable List" en vul in 0,25 bij "Test proportion".*

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
smoking status during pregnancy	Group 1	yes	74	,39	,25	,000[a]
	Group 2	no	115	,61		
	Total		189	1,00		

a. Based on Z Approximation.

OPDRACHTEN

1.
In een grote populatie heeft 30 % van de mensen bloedgroep A. Bereken de kans dat van 5 willekeurig gekozen personen uit deze populatie er precies 2 bloedgroep A hebben.

2.
Toets met behulp van de file lowbwt.sav de nulhypothese dat 10 % van de zwangere vrouwen een geïrriteerde uterus heeft gehad tegen het alternatief dat dit percentage hoger ligt.

Hoofdstuk 7 Twee groepen

7.1 Twee onafhankelijke groepen

Schatten
Een laag geboortegewicht is, ook bij voldragen zwangerschappen, een risico voor verschillende complicaties. In het onderzoek waaruit de data van de file lowbwt.sav afkomstig is, is de variabele "low birth weight" gecodeerd als 0 als het geboortegewicht groter dan of gelijk aan 2500 gram was en als 1 als het geboortegewicht kleiner dan 2500 gram was.
Één van de onderzoeksvragen betrof het effect van roken tijdens de zwangerschap op het krijgen van een kind met laag geboortegewicht. Een puntschatting van het verschil in de proporties "laag geboortegewicht" tussen rooksters en niet-rooksters is snel berekend, maar hoe stellen we het betrouwbaarheidsinterval op?

Tabel 7.1 *Kruistabel van rookgedrag en laag geboortegewicht. In SPSS:* ***Analyze – Descriptive Statistics – Crosstabs –*** *smoke in de rows, low in de columns –* ***Cells*** *– kies voor Percentages "Row"*

smoking status during pregnancy * low birth weight Crosstabulation

			low birth weight		
			no (bwt >= 2500 g)	yes (bwt < 2500)	Total
smoking status during pregnancy	no	Count	86	29	115
		% within smoking status during pregnancy	74,8%	25,2%	100,0%
	yes	Count	44	30	74
		% within smoking status during pregnancy	59,5%	40,5%	100,0%
Total		Count	130	59	189
		% within smoking status during pregnancy	68,8%	31,2%	100,0%

In Tabel 7.1 zien we dat in de groep der rooksters de proportie kinderen met laag geboortegewicht afgerond 0,405 is, terwijl deze bij de niet-rooksters 0,252 bedraagt. Een puntschatting van het verschil in proporties is dan afgerond 0,15.
Om een betrouwbaarheidsinterval te maken voor dit verschil in proporties gebruiken we de theorie die in paragraaf 4.1 is besproken. De standaarddeviatie voor een verschil in twee proporties $\pi_1 - \pi_2$ wordt geschat met de formule

f7 $$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Het 95% BI wordt dan:

$$[\, p_1 - p_2 - 1{,}96 * \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \,;\, p_1 - p_2 + 1{,}96 * \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \,].$$

Dat leidt in dit geval tot: [0,15 – 1,96*0,070 ; 0,15 + 1,96* 0,070] = [0,01 ; 0,29].
We kunnen met 95 % waarschijnlijkheid stellen dat het onbekende verschil in proporties tussen 1% en 29 % ligt. Aangezien het BI de waarde 0 niet bevat, denken we dat er "een echt verschil" is.

Toetsen

Als we willen toetsen of de proporties "laag geboortegewicht" in de twee populaties gelijk zijn, kan dat op verschillende manieren. We hebben de keus uit:

1. (normale benadering van) de toets voor twee proporties
2. de chi-kwadraattoets (χ^2)
3. Fisher's exacte toets

Ad 1. ***Toets voor twee proporties***

We toetsen de $H_0 : \pi_1 = \pi_2$, tegen het tweezijdige alternatief.
De kansvariabele waarop wij onze toets gaan baseren is het verschil in steekproefproporties $p_1 - p_2$ en bepalen de P-waarde met behulp van standaardisatie: $z = \frac{p_1 - p_2}{se(p_1 - p_2)}$. Het berekenen van de standaarddeviatie van het verschil in proporties gaat hier een beetje anders dan hierboven bij het samenstellen van het BI. Als de nulhypothese waar is, en beide proporties zijn gelijk, dan kunnen we een betere schatting van de sd van de onbekende proportie krijgen door alle informatie te combineren. We gebruiken dan formule **f7**, maar vullen bij p_1 en bij p_2 dezelfde gemeenschappelijke schatting in. Deze gemeenschappelijke schatting is het gewogen gemiddelde van de twee steekproefgemiddelden:

$\overline{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{k_1 + k_2}{n_1 + n_2}$, waarin de k's het aantal "successen" (aantal kinderen met laag geboortegewicht) in de groepen zijn. In ons voorbeeld is $\overline{p} = 0{,}31$ en wordt de schatting voor de sd van het verschil in proporties: 0,069, niet veel verschillend van de schatting die eerder gebruikt werd.
De toets voor twee proporties levert dan $P(Z > \frac{0{,}15}{0{,}069}) = P(Z > 2{,}18) = 0{,}0146$ als éénzijdige P-waarde en 0,029 voor de tweezijdige toets.
Conclusie bij een α = 0,05: verwerp de nulhypothese; bij rooksters komt "laag geboortegewicht" relatief vaker voor.
Opmerking: hoewel het theoretisch juist is om ook hier een continuïteitscorrectie toe te passen, heeft dat bij deze aantallen nauwelijks effect op de P-waarde.

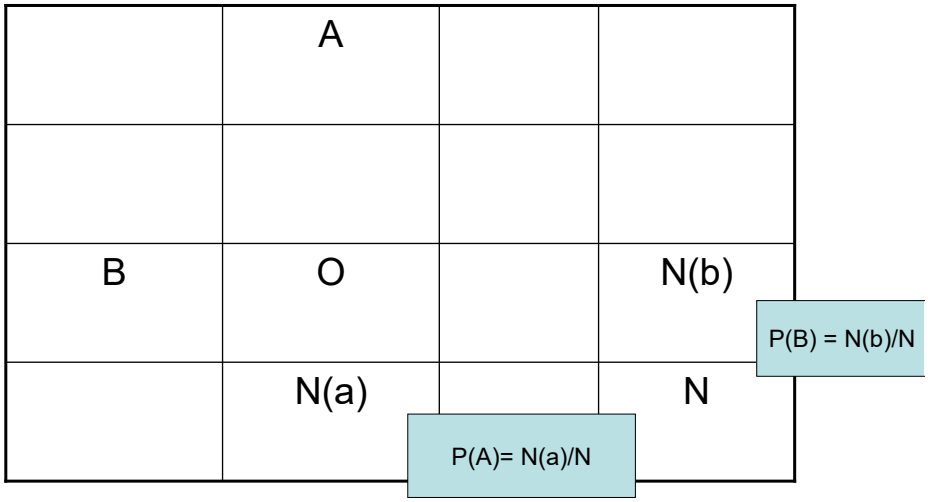
Ad 2: de ***Chi-kwadraat toets***

De nulhypothese $H_0 : \pi_1 = \pi_2$ kan ook anders geformuleerd worden. Als de kans op laag geboortegewicht in beide groepen even groot is, zijn de gebeurtenissen "roken" en "laag geboortegewicht" blijkbaar onafhankelijk van elkaar. Als twee gebeurtenissen A en B onafhankelijk zijn van elkaar dan moet gelden P(A ∩ B) = P(A)*P(B). Deze regel gaan we gebruiken in Tabel 7.1. Als de nulhypothese van onafhankelijkheid waar is, dan verwachten we dat het aantal geobserveerde waarnemingen ("observed") per cel van de kruistabel ongeveer gelijk is aan het aantal verwachte waarnemingen ("expected") per cel. De kansvariabele die als toetsingsgrootheid wordt gebruikt is de chi-kwadraatwaarde die berekend wordt volgens formule **f8**.

f8 $\hat{\chi}^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

De verwachte waarde per cel wordt berekend door het aantal van de betreffende kolom te vermenigvuldigen met het aantal van de betreffende rij en het resultaat te delen door het totaal aantal waarnemingen. De gedachte hierachter is dat als de gebeurtenissen onafhankelijk zijn, de kans om in de cel te komen die zowel in kolom A als in rij B ligt gelijk moet zijn aan de kans om in kolom A te komen maal de kans om in rij B te komen. In formuletaal: $P(A \cap B)$ $P(A)*P(B) = \{N(a)/N\}*\{N(b)/N\}$, zie Tabel 7.2. Deze kans geldt voor elk der N individuen, dus het verwachte aantal mensen in deze cel (E) is deze kans maal het aantal individuen: $E = \{N(a)/N\}*\{N(b)/N\}*N = \{N(a)*N(b)\}/N$

Tabel 7.2 *Uitleg kruistabelanalyse*



Dit geïllustreerd met de data uit Tabel 7.1 geeft de verwachte aantallen in Tabel 7.3:

Tabel 7.3 *De verwachte aantallen in de cellen als de nulhypothese van onafhankelijkheid waar is. In SPSS:* ***Analyze – Descriptive Statistics – Crosstabs –*** *smoke in de rows, low in de columns –* ***Cells*** *– kies voor* ***Expected***

smoking status during pregnancy * low birth weight Crosstabulation

			low birth weight		
			no (bwt >= 2500 g)	yes (bwt < 2500)	Total
smoking status during pregnancy	no	Count	86	29	115
		Expected Count	79,1	35,9	115,0
	yes	Count	44	30	74
		Expected Count	50,9	23,1	74,0
Total		Count	130	59	189
		Expected Count	130,0	59,0	189,0

Merk op dat de relatieve verwachte waarden per rij allemaal gelijk zijn (35,9 / 115 = 23,1 / 74 = 59 / 189 = 0,31), wat je ook verwacht indien de gebeurtenissen onafhankelijk zijn.
De geobserveerde aantallen zullen nooit geheel overeenkomen met de verwachte aantallen, maar als de verschillen tussen Observed (O) en Expected (E) groot zijn is dat een aanwijzing dat de nulhypothese verworpen moet worden. Maar wat is groot? De uitkomst van de

kansvariabele wordt vergeleken met de chi-kwadraat verdeling. Er zijn echter – net als bij de t-verdeling – meerdere chi-kwadraat-verdelingen, afhankelijk van het aantal vrijheidsgraden. Het aantal **vrijheidsgraden** (df) bij een kruistabel analyse wordt bepaald door het aantal rijen en het aantal kolommen. In Tabel 3.5 is sprake van een 2 bij 2 tabel, omdat beide variabelen slecht twee waarden hebben. Als je, bij vaste rij- en kolomtotalen, de verwachte waarde van één cel invult, liggen in een 2 x 2 tabel alle andere verwachte waarden vast. We zeggen dan: deze tabel heeft maar één vrijheidsgraad. Algemeen geldt voor een kruistabel met *r* rijen en *k* kolommen dat het aantal vrijheidsgraden gelijk is aan $(r-1)*(k-1)$.
Formule **f8** levert voor Tabel 6.5:

$$\frac{(86-79{,}1)^2}{79{,}1}+\frac{(29-35{,}9)^2}{35{,}9}+\frac{(44-50{,}9)^2}{50{,}9}+\frac{(30-23{,}1)^2}{23{,}1}=4{,}921$$

Deze waarde kan vergeleken worden met kritieke waarden uit de chi-kwadraat tabel met 1 vrijheidsgraad, zie Swinscow of Altman. We kunnen ook aan SPSS vragen om de P-waarde uit te rekenen, zie Tabel 7.4.

Tabel 7.4 *De chi-kwadraat toets en Fisher's exacte toets. In SPSS:* ***Analyze – Descriptive Statistics – Crosstabs –*** *smoke in the rows, low in the columns,* ***Statistics – Chi-square***

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,924[b]	1	,026		
Continuity Correction[a]	4,236	1	,040		
Likelihood Ratio	4,867	1	,027		
Fisher's Exact Test				,036	,020
Linear-by-Linear Association	4,898	1	,027		
N of Valid Cases	189				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 23,10.

We zien, op afronding na, dezelfde (Pearson) Chi-kwadraat waarde als hierboven berekend is. De P-waarde blijkt 0,026 te zijn, reden voor ons om, bij een significantieniveau van 0,05, de nulhypothese van onafhankelijkheid te verwerpen. Blijkbaar komt er vaker "laag geboortegewicht" voor in de groep rooksters.
De uitgevoerde chi-kwadraattoets heeft twee **voorwaarden**: alle waarnemingen moeten onderling onafhankelijk zijn en de verwachte waarden in alle cellen moet minstens 1 zijn terwijl de verwachte waarden in minstens 80 % van de cellen minstens 5 moet zijn. Dit laatste betekent voor de 2 bij 2 tabel dat de verwachte waarde voor alle vier de cellen minstens 5 moet zijn. Zoals wij kunnen zien in Tabel 7.3 en ook SPSS onder Tabel 7.4 constateert: er zijn hier geen verwachte waarden onder de 5.
Behalve de Pearson chi-kwadraat toets wordt er nog een tweede chi-kwadraat toets uitgevoerd: de chi-kwadraat toets met **continuïteitscorrectie**, ook wel de **Yates-correctie** genoemd. Net als bij de benadering van de binomiale verdeling door de normale verdeling is er bij deze chi-kwadraat toets sprake van een benadering van een discrete verdeling (de geobserveerde aantallen zijn geheel) door een continue. Deze benadering is (iets) beter indien gebruik gemaakt wordt van een correctie door middel van de volgende formule:

$$\chi^2_{cc}=\sum\frac{(|O_i-E_i|-0{,}5)^2}{E_i}$$

Deze chi-kwadraat waarde is altijd kleiner dan die van Pearson en de bijbehorende P-waarde daarom groter. De toets met continuïteitscorrectie zal dus minder snel verwerpen en wordt om die reden aangeduid als conservatiever.

Ad 3. Fisher's exacte toets

De bovenstaande toetsen zijn benaderingstoetsen. Een toets die een exacte P-waarde levert is **Fisher's exacte toets**. SPSS geeft deze automatisch voor de 2 bij 2 tabel als je om de chi-kwadraat toets vraagt, zie Tabel 7.4.
De berekening van deze P-waarde is erg tijdsintensief en met de hand geen pretje. Het komt er op neer dat, gegeven de randtotalen van rijen en kolommen, de kansen berekend worden op alle mogelijke uitkomsten van geobserveerde aantallen in de tabel. Vervolgens wordt de P-waarde bepaald door de som te nemen van de kans op de geobserveerde aantallen plus alle kleinere kansen. Voor een uitgewerkt voorbeeld, zie Swinscow, Altman of Conover. In Tabel 7.4 zien we dat de tweezijdige P-waarde van Fisher's toets tussen de P-waarden van de chi-kwadraat toetsen in ligt. In het gebruikte voorbeeld leiden alle toetsen tot dezelfde conclusie: Bij rooksters komt relatief vaker laag geboortegewicht voor.

De Likelihood Ratio toets toetst dezelfde nulhypothese als de chi-kwadraattoets maar heeft een andere berekeningsgrondslag en zal hier niet worden besproken.

OPDRACHTEN

1.
Geef een 95 % BI voor het verschil in proporties kinderen met "laag geboortegewicht" voor vrouwen met en zonder een historie van geïrriteerde uterus.

2.
Toets of de variabelen laag geboortegewicht en geïrriteerde uterus onafhankelijk zijn.
Welke toets gebruik je? Controleer eventuele voorwaarden voor deze toets en geef je conclusie.

3.
In een onderzoek naar kleurenblindheid bij 14 jongens en 31 meisjes vond men respectievelijk 4 en 2 kleurenblinde kinderen. Toets of er onafhankelijkheid is tussen de variabelen kleurenblindheid en geslacht. Welke toets gebruik je? Wat is de conclusie?
(je kunt in SPSS 45 cases invoeren met elk waarden twee variabelen, maar het is eenvoudiger een datafile te maken met vier cases en drie variabelen:

Geslacht	kleurenblind	aantal
0	0	10
0	1	4
1	0	29
1	1	2

en voor je de analyse laat uitvoeren met **Data – weight cases** aangeeft dat je de variabele aantal als wegingsfactor wilt gebruiken).

7.2 Twee afhankelijke groepen

Bij een onderzoek naar een homeopatisch middel tegen wratten wordt een gepaard onderzoek uitgevoerd bij mensen die op beide armen in gelijke mate last hebben van wratten. Op één arm wordt het homeopatische middel aangebracht, op de andere arm wordt een placebo zalf gebruikt. De keuze van de middelen per arm geschiedt random. Na afloop van het experiment wordt gekeken of de middelen wel of niet geholpen hebben. De resultaten kunnen samengevat worden zoals in tabel 7.5.

Tabel 7.5 *Gepaarde dichotome data*

		Placebo		
		Wel geholpen	Niet geholpen	Totaal
Homeopatisch middel	Wel geholpen	a	b	a+b
	Niet geholpen	c	d	c + d
	Totaal	a + c	b + d	n

Een puntschatting voor het verschil in werkzaamheid wordt verkregen door de proporties "wel geholpen" van elkaar af trekken: $\frac{a+b}{n} - \frac{a+c}{n} = \frac{b-c}{n}$. Voor de constructie van een betrouwbaarheidsinterval rond deze puntschatting, verwijzen we naar Altman.
De toets die gebruikt wordt om data uit een dergelijke gepaarde proefopzet te analyseren is de toets van **McNemar**. De gedachte achter deze toets is de volgende: de personen die geen verschil aangeven tussen beide armen (a en d) geven ons geen informatie over een eventueel verschil tussen de werkzaamheid van beide middelen. Deze mensen worden voor de toets buiten beschouwing gelaten. Als de nulhypothese (geen verschil tussen beide middelen) juist is, verwachten we dat van de mensen die wel een verschil bemerken (b + c) de helft in het voordeel van het ene middel zal aangeven, de andere helft ten faveure van het andere middel.
De toetsgrootheid die bij deze toets gebruikt wordt: $z = \frac{b-c}{\sqrt{b+c}}$. De bijbehorende P-waarde kan met behulp van de tabel van de normale verdeling bepaald worden.
De McNemar test kan in SPSS worden uigevoerd via ***Analyze – Descriptive Statistics – Crosstabs – Statistics - McNemar***

OPDRACHTEN

1.
Van het in paragraaf 7.2 beschreven onderzoek volgen hier de gegevens: er deden 68 mensen mee met het onderzoek. Hiervan gaven 22 personen aan dat op beide armen de middelen geholpen hadden en 26 dat op geen van beide armen het middel had geholpen. Er waren 12 mensen die (na deblindering) aangaven dat het homeopatisch middel had geholpen, maar het placebomiddel niet.
Toets of er een significant verschil is tussen beide middelen. Welke toets gebruik je? Wat is de conclusie?

Hoofdstuk 8 Meer dan twee onafhankelijke groepen

Is de kans op het krijgen van een kind met laag geboortegewicht afhankelijk van het ras? Indien de variabele ras meer dan twee categorieën heeft, zoals in de dataset lowbwt.sav, Gaat het hier om het vergelijken van meerdere proporties. Het is natuurlijk mogelijk om per subgroep (blank, zwart en "ander") een proportie te schatten en paarsgewijs betrouwbaarheidsintervallen op te stellen zoals dat in Hoofdstuk 7 is gebeurd. Maar we willen ook een overall-toets voor alle groepen. Hiervoor wordt dezelfde chi-kwadraat toets gebruikt als in Hoofdstuk 7 is besproken, alleen wordt er over meer dan vier cellen gesommeerd en zal ook het aantal vrijheidgraden verschillen.
In Tabel 8.1 zien we de kruistabel met steekproefpercentages.

Tabel 8.1 *Kruistabel van laag geboortegewicht tegen ras. In SPSS:* ***Analyze – Descriptive Statistics – Crosstabs –*** *race in Columns, low in Rows –* ***Cells – Percentages: Column***

low birth weight * race Crosstabulation

			race			
			white	black	other	Total
low birth weight	no (bwt >= 2500 g)	Count	73	15	42	130
		% within race	76,0%	57,7%	62,7%	68,8%
	yes (bwt < 2500)	Count	23	11	25	59
		% within race	24,0%	42,3%	37,3%	31,2%
Total		Count	96	26	67	189
		% within race	100,0%	100,0%	100,0%	100,0%

We zien dat de geschatte percentages laag geboortegewicht per ras verschillen. Maar zijn de verschillen significant? In Tabel 8.2 wordt de chi-kwadraat toets uitgevoerd volgens formule **f8.** Het aantal vrijheidsgraden wordt bepaald met de formule (r – 1)*(k – 1), waarin r het aantal rijen en k het aantal kolommen is, in dit geval (2 – 1)*(3 – 1) = 2 df. De berekende Chi-kwadraat waarde van Pearson wordt vergeleken met de chi-kwadraat verdeling met 2 vrijheidsgraden wat leidt tot de P-waarde in de laatste kolom. De nulhypothese $H_0 : \pi_1 = \pi_2 = \pi_3$ (ofwel: laag geboortegewicht en ras zijn onafhankelijk) wordt in dit geval, bij een significantieniveau 0,05, niet verworpen.

Tabel 8.2 *Chi-kwadraat toets van laag geboortegewicht en ras. In SPSS:* ***Analyze – Descriptive Statistics – Crosstabs –*** *race in Columns, low in Rows –* ***Statistics – Chi-square***

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5,005[a]	2	,082
Likelihood Ratio	5,010	2	,082
Linear-by-Linear Association	3,570	1	,059
N of Valid Cases	189		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 8,12.

In Tabel 8.2 staan nog twee toetsen. De Likelihood Ratio toets heeft een andere berekeningsgrondslag dan de toets van Pearson en zal hier niet worden besproken. De Linear-

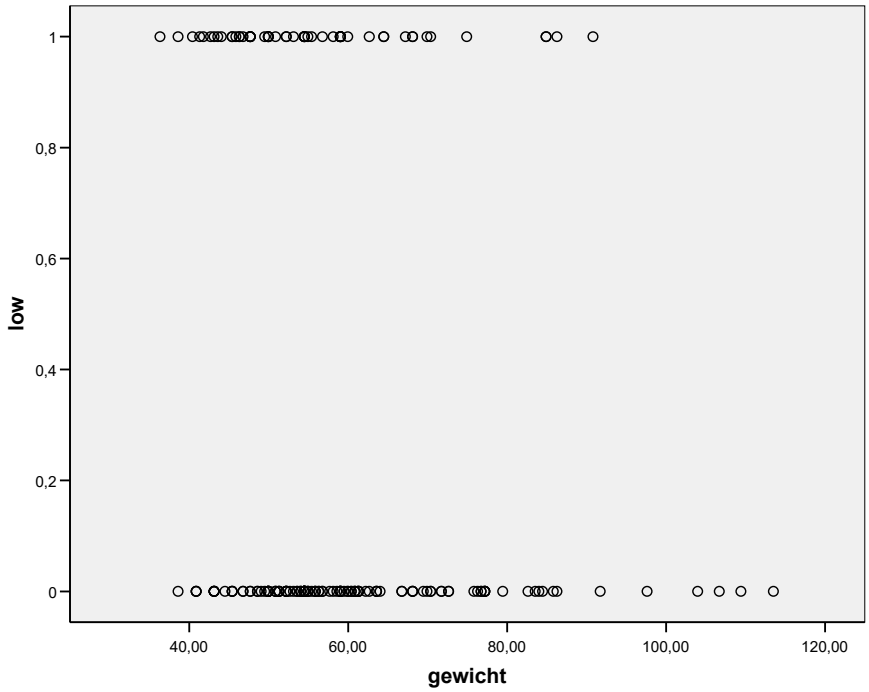
by-Linear Association is een toets voor trend en kan gebruikt worden als de verklarende variabele ordinaal is, bijvoorbeeld oplopende doses van een medicijn, en men verwacht een consistent toenemend (of consistent afnemend) effect.

OPDRACHTEN

1.
Open de dataset Breastcancer.sav.
Toets de nulhypothese van onafhankelijkheid voor de variabelen histologische gradering en status. Welke toets gebruik je? Controleer eventuele voorwaarden bij deze toets. Wat is je conclusie (bij $\alpha = 0{,}05$)

Hoofdstuk 9 Logistische regressie

We willen onderzoeken met welke variabelen de responsievariabele “laag geboortegewicht” samenhangt. Laag geboortegewicht is een dichotome variabele die gecodeerd is als 0 als het geboortegewicht 2500 gram of meer was, en 1 anderszins. Kunnen we met deze responsievariabele met behulp van een lineaire regressie onderzoeken door welke verklarende variabele laag geboortegewicht het beste voorspeld kan worden? We kijken om te beginnen naar een spreidingsdiagram van laag geboortegewicht (“low”) tegen het gewicht van de moeder tijdens de laatste menstruatie, zie Figuur 9.1.

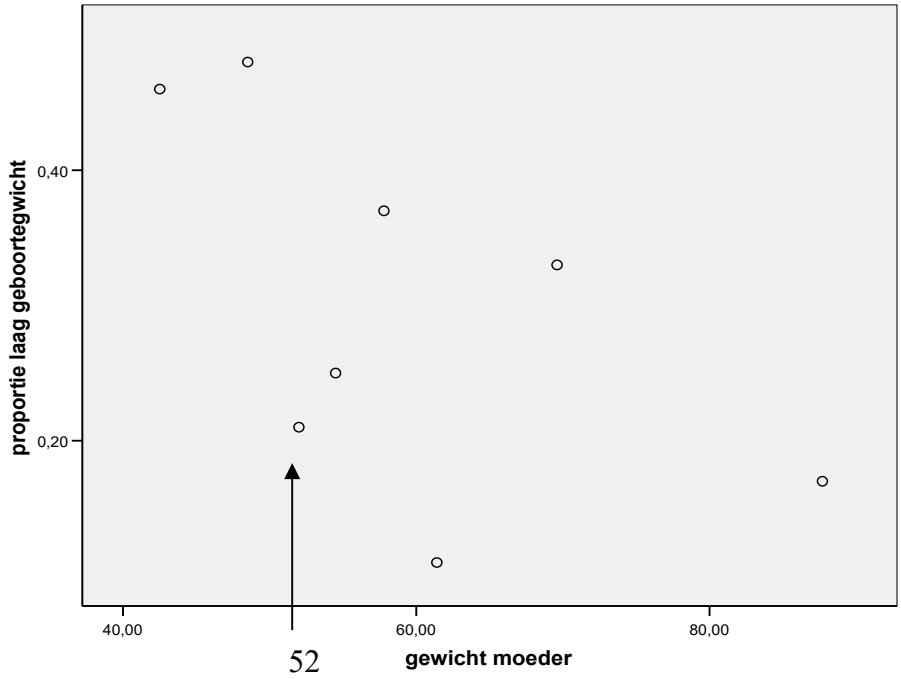


Figuur 9.1 *Strooiingsdiagram van laag geboortegewicht en gewicht van de moeder*

Heeft het zin om door een dergelijke dataset een “zo goed mogelijk passende” rechte lijn te trekken? Behalve dat de relatie tussen beide variabelen niet lineair oogt, is de interpretatie van de lijn moeilijk; wat zou een voorspelde waarde van bijvoorbeeld 0,21 voor iemand van 70 kg betekenen? En een voorspelde waarde van 1,2 voor een moeder van 40 kg? Bovendien zal niet voldaan worden aan de voorwaarde van normaliteit van de residuen. Kortom: lineaire regressie is geen optie als de responsvariabele dichotoom is. Wat dan wel?
Er zijn verschillende modellen mogelijk, maar het meest gebruikte is het **logistische regressie** model. Net als bij een lineaire regressie modelleren we het gemiddelde van Y gegeven X. Bedenk dat bij een dichotome responsvariabele, gecodeerd als 1 voor “Y wel aanwezig” en 0 voor “Y niet aanwezig”, het gemiddelde de proportie respondenten is die de eigenschap Y heeft. Het grote verschil met lineaire regressie is dat niet het gemiddelde van Y zelf, maar een transformatie van het gemiddelde van Y voorspeld wordt door een lineaire combinatie van de onafhankelijke variabelen.

9.1 Enkelvoudige logistische regressie

Om een idee te krijgen wat de gedachte achter logistische regressie is, kijken we eerst naar de **proportie** respondenten met laag geboortegewicht voor acht gewichtsklassen, zie figuur 9.2. De klassen zijn zodanig gekozen dat zij elk ongeveer 12,5 % van de steekproef bevatten. Als de responsvariabele gecodeerd is als 1 voor de kinderen met laag geboortewicht en 0 voor de kinderen zonder laag geboortegewicht, geldt dat de proportie kinderen met laag geboortewicht gelijk is aan het gemiddelde van de responsvariabele. Uit de grafiek blijkt bijvoorbeeld dat de proportie kinderen met laag geboortewicht voor moeders van 52 kg ongeveer 0,21 (21 %) bedraagt. Je kunt dit ook op individueel niveau beschouwen: de kans voor een willekeurige 52 kilo zware moeder om een kind met laag geboortegewicht te krijgen is 0,21. Verder lijkt de proportie laag geboortegewicht af te nemen met het gewicht van de moeder.



Figuur 9.2 *De proportie kinderen met laag geboortegewicht uitgezet tegen gewichtsklasse van de moeder.*

We zoeken naar een statistisch model waarmee we de proportie kinderen met laag geboortegewicht kunnen voorspellen (of verklaren) voor een willekeurige waarde van de verklarende variabele X (in dit geval voor een willekeurig gewicht van de moeder). We zullen deze proportie noteren als $\pi(x)$. Het enkelvoudig logistisch regressiemodel heeft de volgende vorm:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Waarin e het getal van Euler is, het grondtal van de natuurlijke logaritme (e ≈ 2,718).

Een kleine toelichting op de formule: aangezien een macht met een positief grondtal altijd positief is, zijn de teller en noemer van bovenstaande breuk positief en dus de breuk zelf ook. Met andere woorden $\pi(x) > 0$. Omdat de noemer in deze formule altijd groter is dan de teller, is de breuk altijd kleiner dan 1. Dus $0 < \pi(x) < 1$, wat op 0 en 1 na de mogelijke waarden zijn voor een proportie . De parameter β_0 vertelt ons iets over de "ligging" van de grafiek en β_1 geeft informatie over de relatie tussen X (het gewicht van de moeder) en de kans op laag geboortegewicht ("low").
In plaats van naar de kans op "low" wordt vaak naar de **odds** gekeken. De odds is een **kansverhouding:** de kans op laag geboortegewicht gedeeld door de kans op geen laag geboortegewicht:

$$odds(low) = \frac{P(low)}{P(geen\ low)} = \frac{P(low)}{1 - P(low)}$$

Om hier een klein beetje gevoel voor te ontwikkelen de volgende voorbeelden:

- de kans om 2 te gooien bij een worp met een eerlijke dobbelsteen is 1/6. De odds is $\frac{1/6}{5/6} = \frac{1}{5}$, ofwel "1 staat tot 5" (tegenover iedere keer dat je 2 gooit zul je gemiddeld 5 keer geen 2 gooien)
- als de kans ½ is, is de odds 1 (ofwel: de kans dat het wel gebeurt is net zo groot als de kans dat het niet gebeurt)
- als de kans op verkoudheid 0,8 is, is de odds 0,8 / 0,2 = 4
- een kans zit per definitie tussen 0 en 1, een odds kan waarden aannemen tussen 0 en oneindig.

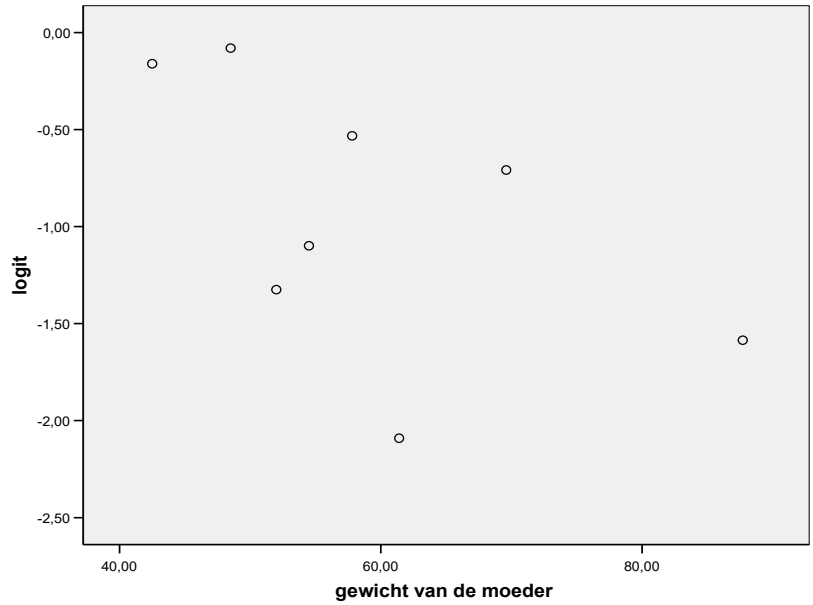
In het geval van een logistisch model wordt de odds:

$$\frac{\pi(x)}{1-\pi(x)} = \frac{e^{\beta_0+\beta_1 x} \big/ 1+e^{\beta_0+\beta_1 x}}{1 \big/ 1+e^{\beta_0+\beta_1 x}} = e^{\beta_0+\beta_1 x}$$

De natuurlijke logaritme van de odds (de ln(odds)) wordt de **logit** genoemd. Hiervoor geldt:

$$\ln(odds) = \ln(e^{\beta_0+\beta_1 x}) = \beta_0 + \beta_1 x$$

In deze formule zien we dat de logit van Y lineair van X afhangt.



Figuur 9.3 *de logit van laag geboortegewicht afgezet tegen de gewichtsklasse van de moeder*

De parameters van het logistisch regressiemodel kunnen we laten schatten door SPSS, maar hoe gebeurt dat en hoe interpreteren we deze getallen?
De onbekende parameters β_0 en β_1 worden geschat met de **maximum likelihood** methode. Een korte uitleg zonder in de wiskundige details te treden: eerst wordt een zogenaamde likelihood-functie **L** opgesteld, een functie die de waarschijnlijkheid van de geobserveerde data aangeeft als functie van de onbekende parameters. Door deze functie te maximaliseren worden de schattingen voor de onbekende parameters verkregen. Maar wat betekenen deze getallen?
Om één en ander te verduidelijken kijken we eerst naar de relatie tussen "low" en een dichotome verklarende variabele, waarna we terugkeren naar de situatie waarin we een continue verklarende variabele hebben.

9.1.1 Een dichotome verklarende variabele

We zijn geïnteresseerd in de vraag of laag geboortegewicht afhankelijk is van het roken tijdens de zwangerschap. Als eerste oriëntatie laten we een kruistabel van deze twee dichotome variabelen maken, zie Tabel 9.1.

Tabel 9.1 *kruistabel van laag geboortewicht en smoke tijdens de zwangerschap. In SPSS:* ***Analyze – Descriptive Statistics – Crosstabs – Cells – Percentages: Row***

smoking status during pregnancy * low birth weight Crosstabulation

			low birth weight		
			no (bwt >= 2500 g)	yes (bwt < 2500)	Total
smoking status during pregnancy	no	Count	86	29	115
		% within smoking status during pregnancy	74,8%	25,2%	100,0%
	yes	Count	44	30	74
		% within smoking status during pregnancy	59,5%	40,5%	100,0%
Total		Count	130	59	189
		% within smoking status during pregnancy	68,8%	31,2%	100,0%

Uit de gegevens van Tabel 9.1 schatten we de kans op laag geboortegewicht voor rooksters als 0,405 en voor niet-rooksters als 0,252. In plaats van naar "kans" kunnen we ook kijken naar de **odds** die in de vorige paragraaf is geïntroduceerd.
Voor rooksters leidt dat tot een odds van 0,405 / 0,595 = 0,681, voor niet-rooksters wordt dat 0,252 / 0,748 = 0,337. Als we geïnteresseerd zijn in de relatie tussen rookgedrag en (de kans op) laag geboortegewicht, kunnen we kijken naar de **Odds Ratio (OR)**, de odds van de rooksters gedeeld door de odds van de niet-rooksters:

$$OR = \frac{odds_{rooksters}}{odds_{nietrooksters}} = \frac{0{,}681}{0{,}337} = 2{,}02$$

Dit geeft aan dat, in deze dataset, de odds van rooksters 2,02 maal zo groot is als de odds van niet-rooksters en roken tijdens de zwangerschap dus een hoger risico geeft op laag geboortegewicht. Maar is het significant hoger? En hoe moeten we te werk gaan als we rekening willen houden met mogelijke confounders zoals leeftijd? Antwoorden op deze vragen kunnen we vinden met behulp van **logistische regressie**.
Als we in SPSS een logistische regressie uitvoeren met "low" als afhankelijke variabele en rookgedrag als verklarende variabele vinden we, onder andere, tabel 9.2.

Tabel 9.2 *deel van de uitvoer van een logistische regressie van laag geboortegewicht op rookgedrag; de coëfficiënten en de Wald test. In SPSS:* ***Analyze – Regression – Binary Logistic –*** *Low bij Dependent, Smoke bij Covariate(s)*

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	smoke	,704	,320	4,852	1	,028	2,022
	Constant	-1,087	,215	25,627	1	,000	,337

a. Variable(s) entered on step 1: smoke.

We zien dat de constante β_0 geschat wordt als -1,087 en de coëfficiënt van geslacht β_1 als 0,704. Dat wil zeggen dat de kans op laag geboortegewicht als volgt van het rookgedrag afhangt:

$$\pi(low) = \frac{e^{-1,087+0,704*smoke}}{1+e^{-1,087+0,704*smoke}}$$

Om dit te kunnen interpreteren, moeten we weten wat de codering van rookgedrag in deze dataset is: niet-rookster = 0, rookster = 1. Op grond van dit model is de kans voor een willekeurige niet-rookster en een willekeurige rookster:

$$\pi(niet-rookster) = \frac{e^{-1,087}}{1+e^{-1,087}} = \frac{0,337}{1+0,337} = 0,252$$

$$\pi(rookster) = \frac{e^{-1,087+0,704}}{1+e^{-1,087+0,704}} = \frac{0,682}{1+0,682} = 0,405$$

Merk op dat dit dezelfde kansen zijn die wij uit de kruistabel (Tabel 9.1) hadden berekend. In de laatste kolom van Tabel 9.2, getiteld Exp(B), staat op de rij van de variabele smoke het getal 2,022. Dit getal is berekend door de coëfficiënt van smoke (0,704) in de exponent van de e-macht te stoppen: $e^{0,704} = 2,022$. Dit getal hebben we hierboven al berekend uit de kruistabel als de odds ratio voor rooksters ten opzichte van niet-rooksters op het krijgen van een kind met laag geboortegewicht. We weten nu dus hoe we de geschatte coëfficiënt uit de logistische regressie moeten interpreteren:

e^{β_1} **is de Odds Ratio**, dat wil zeggen de odds van de groep met code 1 (in dit geval de rooksters) gedeeld door de odds van de groep met code 0 (de niet-rooksters) op de gebeurtenis Y = 1 (het krijgen van een kind met laag geboortegewicht).

Als de schatting voor β_1 gelijk is aan 0, is de geschatte odds ratio 1, dat wil zeggen: er is geen effect van de betreffende verklarende variabele. Als de schatting voor de coëfficiënt positief is, wordt de odds ratio groter dan 1 en is er dus een positief effect. Is de geschatte coëfficiënt negatief dan is de odds ratio kleiner dan 1 en constateren we een negatief effect.

Verder geldt dat e^{β_0} gelijk is aan 0,337, dit is de odds voor de respondenten met smoke = 0, de niet-rooksters.

Een formeel bewijs voor de interpretatie van de coëfficiënten in een logistische regressie met een dichotome verklarende variabele staat in Appendix B van deze syllabus.

Als we willen toetsen of er een significante relatie is tussen rookgedrag en laag geboortegewicht, kan dat met behulp van verschillende toetsen. De nulhypothese luidt: er is geen relatie tussen rookgedrag en laag geboortegewicht. Of in formuletaal: $H_0 : \beta_1 = 0$.

Bedenk dat deze nulhypothese inhoudt dat de odds ratio gelijk is aan $e^0 = 1$, dus dat de odds van rooksters gelijk is aan de odds van niet-rooksters.
Één van de mogelijke toetsen is de **Wald-toets**, zie Tabel 9.2. Naast de coëfficiënt van geslacht wordt ook de standaardfout van deze coëfficiënt geschat (S.E.). Het getal in de kolom "Wald" wordt berekend door het quotiënt van B en S.E. te kwadrateren: $(0,704/0,32)^2 = 4,852$. Deze waarde wordt vergeleken met een chi-kwadraat verdeling met één vrijheidsgraad wat resulteert in een P-waarde = 0,028. Bij een significantieniveau $\alpha = 0,05$ wordt de nulhypothese verworpen en concluderen we dat rookgedrag een significante relatie heeft met

laag geboortegewicht, ofwel dat er een verschil is tussen rooksters en niet-rooksters wat betreft het risico op laag geboortegewicht.

De Wald test is niet altijd betrouwbaar, met name als de schatting voor β of de schatting van de S.E van β een groot getal geeft. Een meer valide test is de **Likelihood Ratio toets**, die we kort zullen bespreken zonder in mathematische details te treden. Deze test stelt ons in staat om verschillende modellen met elkaar te vergelijken. Zo kunnen we ook toetsen of het model met verklarende variabele rookgedrag beter is dan een model zonder verklarende variabelen (het zogenaamde "nulmodel").

Tabel 9.3 *Deel van de uitvoer van een logistische regressie van laag geboortegewicht op rookgedragt; de Likelihood Ratio test. In SPSS:* ***Analyze – Regression – Binary Logistic –*** *Low bij Dependent, Smoke bij Covariate(s)*

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	4,867	1	,027
	Block	4,867	1	,027
	Model	4,867	1	,027

De Likelihood Ratio test vergelijkt de waarde van de likelihoodfunctie **L** van het model met verklarende variabele rookgedrag (**L1)** met die van het model zonder verklarende variabelen (**L0**). Om precies te zijn: er wordt gekeken naar de volgende toetsgrootheid LR = 2ln(L1/L0) = 2ln(L1) – 2 ln(L0). De nulhypothese luidt dat beide modellen even veel verklaren.
SPSS geeft standaard drie toetsen:

- "Step": de toets voor de variabele die bij deze stap van modelopbouw wordt toegevoegd aan het model
- "Block": een toets die gebruikt kan worden om te toetsen of een aantal variabelen dat gezamenlijk aan het model wordt toegevoegd een significante verbetering van het model geeft
- "Model": de toets voor het gehele model.

Bij dit voorbeeld zijn deze drie toetsen identiek omdat er slechts één variabele toegevoegd wordt, en leidt de Likelihood Ratio toets tot een chi-kwadraat waarde van 4,867 met een bijbehorende P-waarde gelijk aan 0,027. De resultaten zijn hier nagenoeg dezelfde als bij de Wald toets en de conclusie luidt ook hier dat de nulhypothese wordt verworpen: rookgedrag heeft een significante relatie met laag geboortegewicht.
De likelihood Ratio test waren we ook al tegen gekomen in tabel 7.4.

9.1.2 Een continue verklarende variabele

Terug naar onze vraag over de invloed van gewicht van de moeder op een kind met laag geboortegewicht. Hoe interpreteren we nu de coëfficiënt van een continue variabele zoals gewicht (van de moeder)? Een enkelvoudige logistische regressie in SPSS met responsvariabele "low" en verklarende variabele gewicht van de moeder geeft onder andere Tabel 9.4.

Tabel 9.4 *deel van de uitvoer van een logistische regressie van laag geboortegewicht op gewicht van de moeder: de coëfficiënten en de Wald test. In SPSS: :* ***Analyze – Regression – Binary Logistic –*** *Low bij Dependent, Gewicht bij Covariate(s)*

Variables in the Equation

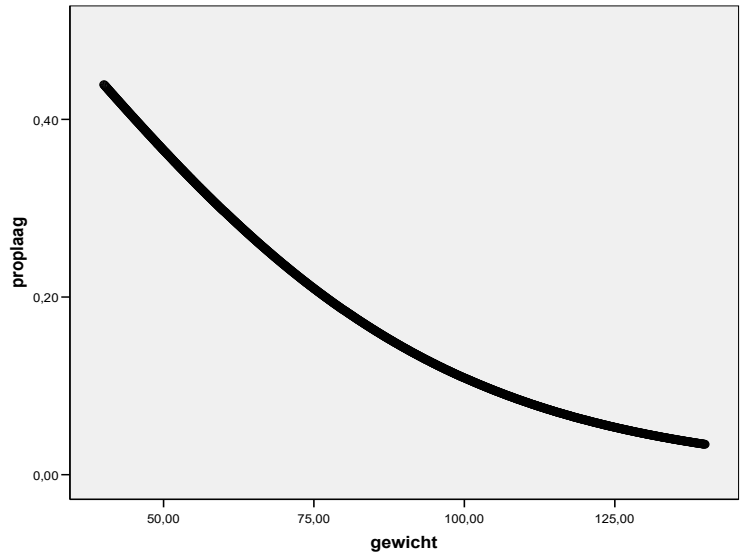
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	gewicht	-,031	,014	5,192	1	,023	,970
	Constant	,998	,785	1,616	1	,204	2,714

a. Variable(s) entered on step 1: gewicht.

We kunnen constateren dat de nulhypothese $H_0 : \beta_1 = 0$ verworpen wordt, want de P-waarde is kleiner dan 0,05. De proportie kinderen met laag geboortegewicht blijkt af te hangen van het gewicht van de moeder (tijdens de laatste menstruatie), maar hoe precies?
Volgens de geschatte coëfficiënten is het best passende logistische model met gewicht als verklarende variabele voor de kans op een kind met laag geboortegewicht:

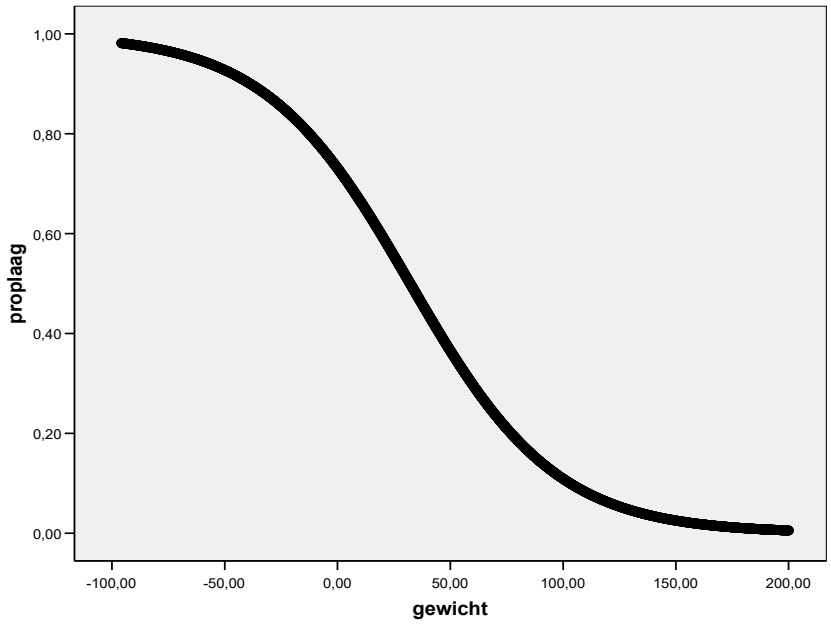
$$\pi(gewicht) = \frac{e^{0,998-0,031*gewicht}}{1+e^{0,998-0,031*gewicht}}$$

Als we van deze functie een grafiek tekenen op het domein van 55 tot en met 85 jaar vinden we Figuur 9.4.



Figuur 9.4 *de geschatte relatie tussen de proportie kinderen met laag geboortegewicht en het gewicht van de moeder op het domein van 40 tot en met 140 kilogram*

Deze grafiek geeft aan dat er een negatief verband is; hoe hoger het gewicht, hoe lager de kans op een laag geboortegewicht. Het verband lijkt op dit domein vrijwel lineair. De typische S-vorm van de grafiek die bij een logistische functie hoort wordt in dit geval echter pas duidelijk als we het domein kunstmatig vergroten, zie Figuur 9.5.



Figuur 9.5 *de geschatte relatie tussen de proportie kinderen met laag geboortegewicht en het gewicht van de moeder op het hypothetische domein van -100 tot en met 200 kilogram*

Analoog aan het geval van een dichotome verklarende variabele kunnen we afleiden dat de odds voor een moeder met gewicht x gelijk is aan

$$e^{\beta_0+\beta_1 x}$$

Voor een persoon die één kilogram zwaarder is (dus gewicht x + 1 heeft) is de odds

$$e^{\beta_0+\beta_1(x+1)}$$

De **Odds Ratio** voor die moeder die één kilo zwaarder is ten opzichte van de één kilo lichtere moeder wordt dan

$$\frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_0+\beta_1(x+1)-(\beta_0+\beta_1 x)} = e^{\beta_1}$$

De odds voor een moeder van bijvoorbeeld 73 kilo om een kind met laag geboortegewicht te krijgen is dus ongeveer $e^{-0,031} = 0,97$ maal zo groot (dus 1,03 maal zo klein) als de odds van een moeder van 72 kilo. Deze moeder van 72 kilo heeft weer een 0,97 maal zo grote odds als een moeder van 71 kilo. Daaruit volgt dat de moeder van 73 kilo een $0,97*0,97 = 0,97^2 \approx 0,94$ maal zo grote odds heeft als de 71 kilo wegende moeder.

In het algemeen geldt bij een continue variabele X dat een persoon met waarde X = x +k, een $\left(e^{\beta_1}\right)^k = e^{k*\beta_1}$ maal zo grote odds heeft als een persoon met waarde

X = x. Zo is de odds op een kind met laag geboortegewicht van een 70 kilo zware moeder ten opzichte van een 60 kilo zware moeder $\left(e^{-0,031}\right)^{10} \approx (0,97)^{10} \approx 0,74$ maal zo groot, ofwel 1/0,74 = 1,36 maal zo klein.
e^{β_0} is de odds voor een respondent met waarde 0 op de verklarende variabele. Dat is een beetje onzinnig wanneer de verklarende variabele gewicht is, maar met behulp van lineaire transformaties kan deze schatting wel degelijk praktisch nut hebben.

9.2 Meervoudige logistische regressie

Hangt de kans op het krijgen van een kind met laag geboortegewicht alleen van het gewicht van de moeder af of zijn er meerdere variabelen gerelateerd aan de kans op deze gebeurtenis? Net als bij het lineaire regressiemodel laat het enkelvoudige logistische regressiemodel zich eenvoudig uitbreiden tot een meervoudig logistisch regressiemodel:

$$\pi(x_1,...,x_k) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...\beta_k x_k}}$$

Of, hieruit afgeleid: de odds is $e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}$, wat equivalent is met

$$\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k .$$

De ln(odds) wordt, zoals al eerder opgemerkt, in de literatuur ook wel aangeduid als de **logit**. We zien dat niet de oorspronkelijke responsievariabele laag geboortegewicht, maar een transformatie van de kans dat laag geboortegewicht = 1 als lineaire combinatie van de verklarende variabelen wordt geschreven. Als we willen onderzoeken door welke combinatie van variabelen laag geboortegewicht het best verklaard wordt, kunnen we op dezelfde wijze te werk gaan als bij een lineaire regressie. Dat wil zeggen dat we de volgende stappen kunnen nemen:

1. Welke variabelen zijn op grond van theorie kandidaten voor het meervoudige logistische regressiemodel?
2. Welke variabelen zijn op grond van een univariate logistische regressie kandidaten voor het meervoudig logistisch regressiemodel? Neem hierbij een ruime alfa, bijvoorbeeld α = 0,25.
3. Bouw het model stap voor stap op door eerst te beginnen met de verklarende variabele die univariaat de kleinste P-waarde had. Voeg één voor één de overige kandidaat-variabelen toe aan het model en laat een nieuwe kandidaat-variabele in het model als het model significant beter wordt. Beslis dit laatste met behulp van de Likelihood Ratio toets.
   Je kunt ook “backward” te werk gaan; begin met alle variabelen in je model en verwijder één voor één de minst significante variabele tot je een model overhoudt met alleen maar significante of theoretisch relevante variabelen.
4. Tot slot kijk of er eventuele interactievariabelen aan het model toegevoegd dienen te worden. Doe dit alleen voor variabelen die op grond van theorie een interactie zouden kunnen vertonen en / of erg grote effecten hebben.

Ter illustratie kijken we naar de vergelijking van twee geneste meervoudige logistische regressiemodellen voor laag geboortegewicht: we hebben een model met verklarende variabelen smoke en gewicht en we willen weten of toevoeging van de variabele leeftijd een significant beter model geeft. De -2 log likelihood van het eerstgenoemde model is ongeveer 224,34 (zie tabel 9.5).

Tabel 9.5 *Meervoudige logistische regressie van laag geboortegewicht op rookgedrag en gewicht. In SPSS: :* ***Analyze – Regression – Binary Logistic –*** *Low bij Dependent, Gewicht en smoke bij Covariate(s)*

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	224,341[a]	,053	,075

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

Als we een extra variabele aan het model toevoegen zal het model meer verklaren, en daarmee zal de likelihood (de mate waarin het model bij de data past) toenemen. Minus twee maal de logaritme van die likelihood (m.a.w. de -2 log likelihood) zal dus afnemen. Hoe kleiner de -2 log likelihood, hoe beter het model. Toevoeging van leeftijd leidt tot een -2 log likelihood van ongeveer 222,88 (zie Tabel 9.6).

Tabel 9.6 *Meervoudige logistische regressie van laag geboortegewicht op leeftijd, gewicht en rookgedrag. In SPSS: :* ***Analyze – Regression – Binary Logistic –*** *Low bij Dependent, Gewicht, smoke en leeftijd bij Covariate(s)*

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	222,879[a]	,060	,085

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

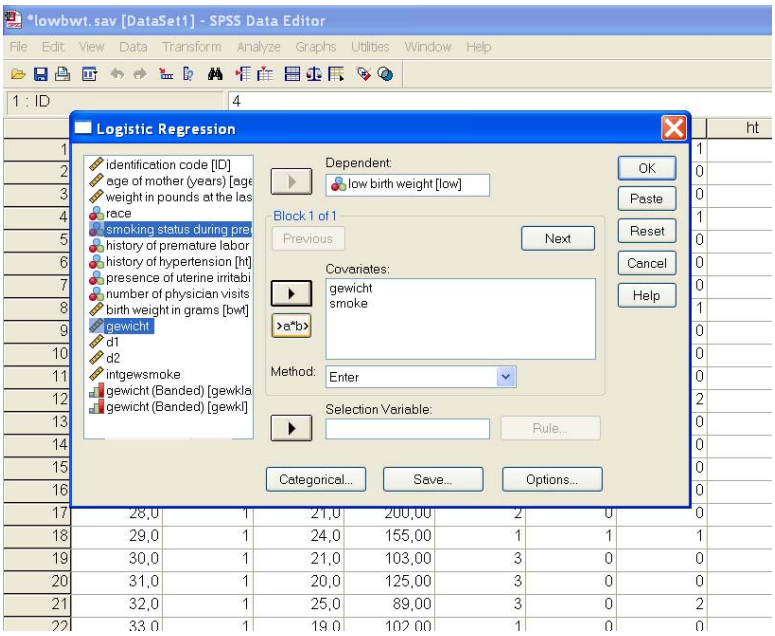
Is dit een significante verbetering? Het verschil tussen twee -2 log likelihoodwaarden heeft, onder de nulhypothese dat de toegevoegde variabele(n) geen invloed heeft (hebben), een chi-kwadraat verdeling met evenveel vrijheidsgraden als er variabelen worden toegevoegd. In ons geval is het verschil 224,34 – 222,88 = 1,46. Dat is niet significant bij een chi-kwadraat toets met één vrijheidsgraad. Het grotere model is dus niet significant beter dan het kleinere model. We geven dan de voorkeur aan het kleinere model: leeftijd van de moeder voegt niets toe aan de voorspelling van laag geboortegewicht als we rookgedrag en gewicht weten.

<u>Logistische regressie in SPPS</u>

Het opbouwen van een model met meerdere verklarende variabelen in een logistische regressie gaat op vergelijkbare wijze als bij een lineaire regressie. Binnen SPSS zijn er echter twee praktische verschillen:

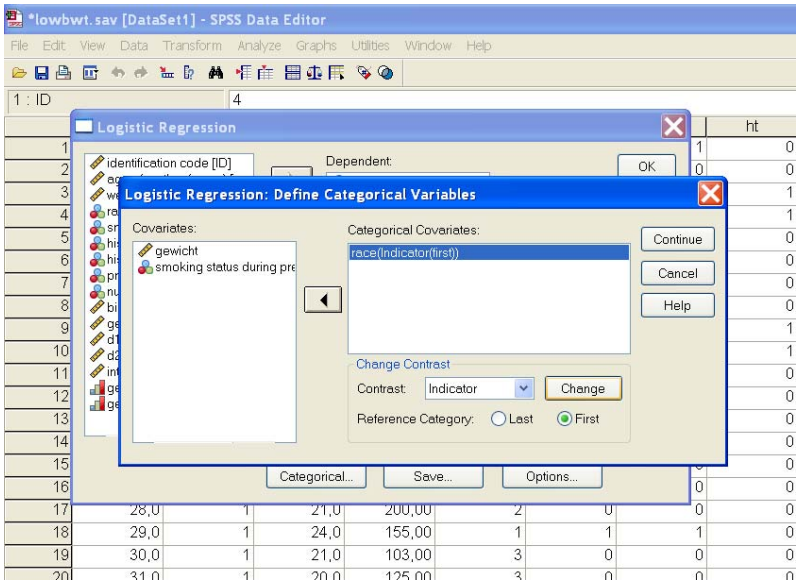
- bij het gebruik van interactietermen in een lineaire regressie moet je zelf de interactievariabelen maken voor je de lineaire regressie uit kunt voeren. Bij logistische regressie kan SPSS dat voor je doen binnen het menuvenster van de logistische regressie, zie Figuur 9.6. Als je twee (verklarende) variabelen markeert wordt het

knopje >a*b> actief. Als je hier op klikt wordt de interactieterm van de gemarkeerde variabelen aan het model toegevoegd.



Figuur 9.6 *Logistische regressie in SPSS, interactietermen*

- bij het gebruik van nominale verklarende variabelen met meer dan twee categorieën in een lineaire regressie moet je zelf dummy variabelen aanmaken. Binnen een logistische regressie wordt het je eenvoudiger gemaakt. In het menuvenster van de logistische regressie kun je met behulp van het knopje “Categorical” aangeven welke variabelen nominaal zijn en SPSS zal automatisch dummyvariabelen aanmaken. Standaard is de referentiegroep codering met de hoogst genummerde groep als referentiegroep. Je kunt dit veranderen in de eerste categorie met “first”gevolgd door “change”, zie Figuur 9.7.



Figuur 9.7 *het maken van dummy’s binnen logistische regressie*

OPDRACHTEN

1.
Toets met behulp van logistische regressie of de variabele “geïrriteerde uterus” van invloed is op de kans op een kind met laag geboortegewicht. Geef de interpretatie van de geschatte regressiecoëfficiënt en geef je conclusie.
Vergelijk je resultaten met een chi-kwadraat toets voor onafhankelijkheid tussen beide variabelen.

2.
Toets met behulp van logistische regressie of de variabele ras van invloed is op de kans op een kind met laag geboortegewicht. Hoe worden de regressiecoëfficiënten geïnterpreteerd? Wat is de conclusie?

3.
Kies op voorhand vijf verklarende variabelen in de datafile lowbwt.sav waaronder ras, gewicht moeder en rookgedrag. Bepaal het model dat zo goed mogelijk de variabele “laag geboortegewicht” voorspelt.

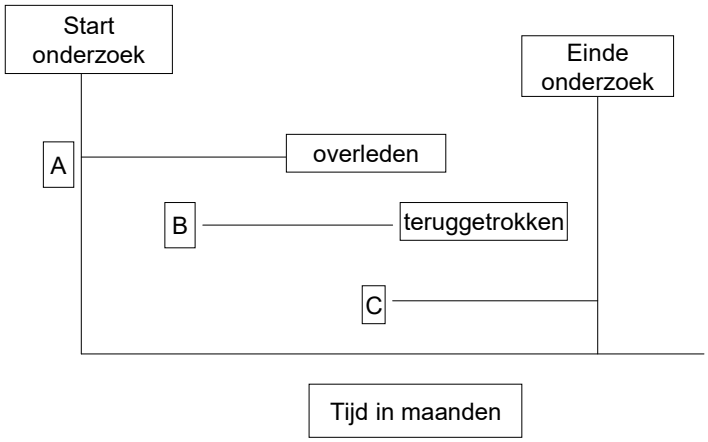
Deel 3

Hoofdstuk 10 Survival analyse

We hebben al kennis gemaakt met verschillende soorten variabelen en statistische technieken. In dit hoofdstuk bekijken we een nieuw type variabele dat niet op een geschikte wijze te analyseren is met de tot op dit moment behandelde methoden. Deze variabele komt naar voren wanneer men geïnteresseerd is in de tijd die verstrijkt tot een bepaalde gebeurtenis plaatsvindt. In de medische context gaat het vaak om tijd tot genezing, tijd tot recidive of tijd tot overlijden (ook wel: overlevingsduur). Analyses met dit soort gegevens worden meestal aangeduid als **survival analyse**. In paragraaf 10.1 zullen we relatief eenvoudige analysetechnieken bekijken: Kaplan-Meier en de logrank toets. In paragraaf 10.2 volgt een complexere, multivariate, methode om survival data te analyseren: de Cox-regressie.

10.1 Kaplan-Meier en de logrank toets

In de dataset Breastcancer.sav komt de variabele Time voor. Dit is de tijd die de vrouwen in het onderzoek betrokken zijn geweest. Wat maakt analyse van deze gegevens anders dan analyse van bijvoorbeeld de bloeddruk? We moeten ons realiseren dat de variabele tijd nog gecombineerd moet worden met de variabele "status" die aangeeft of de desbetreffende vrouw nog in leven is. Ter illustratie drie dames in Figuur 10.1. Mevrouw A was als patiënte geïncludeerd bij de start van het onderzoek en is 15 maanden later overleden. Mevrouw B is pas later geïncludeerd, is 15 maanden als patiënte bij het onderzoek betrokken geweest en heeft zich toen uit het onderzoek teruggetrokken (om onbekende redenen). Mevrouw C ten slotte is vrij laat bij het onderzoek betrokken geraakt en was ten tijde van het sluiten van het onderzoek nog in leven.



Figuur 10.1 *Drie vrouwen die elk 15 maanden in het onderzoek aanwezig waren*

Voor alle drie de dames wordt de waarde op de variabele "Time" 15. Van mevrouw A weten we dat zij, na diagnose, nog 15 maanden heeft geleefd en toen is overleden. Van de dames B en C weten we dat zij, na diagnose, nog ***minstens*** 15 maanden hebben geleefd. We noemen deze waarden **gecensureerd** (naar rechts). Bij een analyse van de overlevingsduur zullen we

met dit verschil rekening moeten houden, domweg een gemiddelde van de overlevingsduur berekenen is hier dus niet correct.
Bij Survival analyse zijn er vaak twee zaken van belang: het schatten van een bepaalde overlevingskans (op zeker tijdstip, bijvoorbeeld vijfjaarsoverleving) of het toetsen van verschillen in overleving tussen verscheidene groepen.
Het schatten van de overlevingskansen gaat stapsgewijs. We zullen dit uitleggen aan de hand van een voorbeeld met een kleine steekproef. Bij een balans experiment proberen 10 deelnemers 5 minuten lang op een smalle balk te staan. Als men van de balk valt stopt de tijd en wordt er een "event" genoteerd. Het is echter ook mogelijk voor de deelnemer om aan te geven te willen stoppen, zonder dat zij / hij van de balk is gevallen. Van deze persoon wordt dan de tijd genoteerd en aangegeven als gecensureerd. Van de 10 deelnemers maakten 6 personen de 300 seconden vol. Één persoon viel na 25 seconden, een ander na 45 seconden, na 110 seconden gaf een deelnemer aan te willen stoppen (gecensureerd) en de tiende persoon viel na 185 seconden.
Hoe worden de "overlevingskansen" berekend? Aangezien alle deelnemers de eerste 25 seconden hebben doorstaan is de overlevingskans voor deze periode gelijk aan 1. De kans om na 25 seconde nog op de balk te staan is 0,9 want 9 van de 10 personen hebben dat volbracht. De overlevingskans blijft 0,9 tot er 45 seconden zijn verstreken. Er zijn nu nog 9 personen "at risk" om een "event" mee te maken. Één van deze 9 personen valt om 45 seconden. De *conditionele kans* om de 45 seconden te overleven, gegeven het feit dat je minstens 25 seconden hebt volbracht, is 8/9 ≈ 0,89. De *cumulatieve kans* om vanaf het begin de 45 seconden te overleven is 9/10 * 8/9 = 0,8. Dit viel natuurlijk ook simpeler te berekenen: 8 van de 10 mensen hebben het langer dan 45 seconden vol gehouden, dus de overlevingskans is 8/10 = 0,8. Zie Tabel 10.1.

Tabel 10.1 *Overlevingskansen in het balansexperiment. In SPSS:* ***Analyze – Survival – Kaplan-Meier*** *- Time en Status invullen, Define Event is 1*

Survival Table

	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	25,000	1,00	,900	,095	1	9
2	45,000	1,00	,800	,126	2	8
3	110,000	,00	.	.	2	7
4	185,000	1,00	,686	,151	3	6
5	300,000	,00	.	.	3	5
6	300,000	,00	.	.	3	4
7	300,000	,00	.	.	3	3
8	300,000	,00	.	.	3	2
9	300,000	,00	.	.	3	1
10	300,000	,00	.	.	3	0

Na 110 seconden vraagt één van de proefpersonen om het experiment te stoppen, zonder dat er een "event" plaats vindt. Er verandert op dat moment niets aan de overlevingskans, maar vanaf 110 seconden verandert wel het aantal personen dat "at risk" is. Op 185 valt één van de zeven overgebleven personen van de balk, dus de kans om langer dan 185 seconden op de balk te blijven, gegeven dat je 110 seconden hebt overleefd is 6/7. De cumulatieve overlevingskans wordt dan 0,8*6/7 = 0,686. (NB: dit is dus geen 7/10). Zie ook Tabel 10.1.

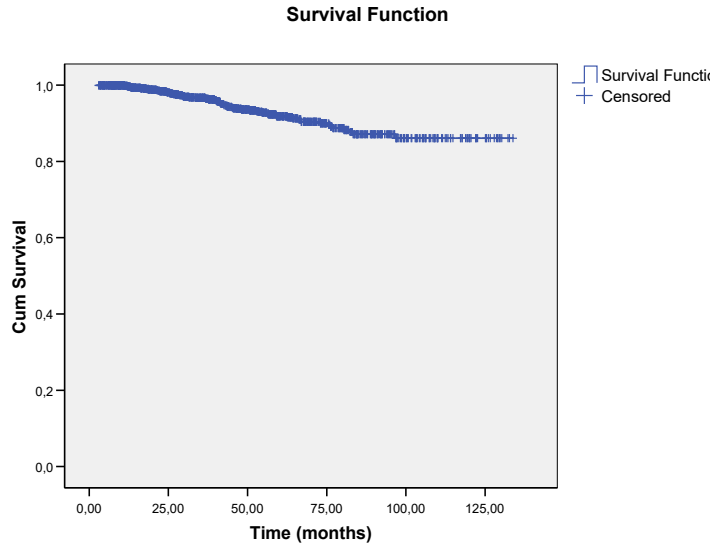
De standaard errors bij deze overlevingskansen *p* worden berekend met de formule van Greenwood, een vrij complexe formule, waarbij r het aantal personen at risk is vanaf dat moment.
Voor de Breastcancer data (n = 1207) wordt een tabel zoals in Tabel 10.1 is afgebeeld vreselijk lang. Het is mogelijk om een "life table" te maken met geclassificeerde data, zie Tabel 10.2.

Tabel 10.2 *Life table van de Breastcancer data. In SPSS:* ***Analyze – Survival – Life Tables –*** *Time en Status invullen, Define Event is 1, Display Time Intervals 0 – 120 by 12*

Life Table[a]

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Number of Terminal Events	Proportion Terminating	Proportion Surviving	Cumulative Proportion Surviving at End of Interval	Std. Error of Cumulative Proportion Surviving at End of Interval
,000	1207	129	1142,500	2	,00	1,00	1,00	,00
12,000	1076	183	984,500	15	,02	,98	,98	,00
24,000	878	147	804,500	14	,02	,98	,97	,01
36,000	717	166	634,000	20	,03	,97	,94	,01
48,000	531	153	454,500	8	,02	,98	,92	,01
60,000	370	121	309,500	5	,02	,98	,90	,01
72,000	244	91	198,500	7	,04	,96	,87	,02
84,000	146	59	116,500	0	,00	1,00	,87	,02
96,000	87	39	67,500	1	,01	,99	,86	,02
108,000	47	25	34,500	0	,00	1,00	,86	,02

Tabel 10.2 laat soortgelijke informatie zien als in Table 10.1, alleen worden de overlevingskansen niet per persoon (event of censurering) berekend, maar (in dit voorbeeld) per jaar. Het aantal personen at risk over een jaar wordt bepaald door het gemiddelde te nemen van het aantal mensen aan het begin van de periode en het aantal aan het eind van de periode. Met andere woorden: er wordt van uit gegaan dat censurering uniform over het jaar heeft plaatsgevonden. Voor het eerste jaar bijvoorbeeld waren er in het begin 1207 vrouwen in het onderzoek en 129 "withdrawels" gedurende dat eerste jaar. Aan het eind van het jaar zijn er dus nog 1207 – 129 = 1078 vrouwen (inclusief de "events"). Gemiddeld waren er dan (1207 + 1078)/2 = 1142,5 vrouwen at risk. Hiervan zijn er 2 overleden, wat een overlevingskans van afgerond 1,00 geeft.



Figuur 10.2 *Kaplan-Meier survival curve voor de Breastcancer data. In SPSS:* ***Analyze – Survival – Kaplan-Meier –*** *Time en Status invullen, Define Event is 1,* ***Options*** *Survival table uitvinken (geeft hier erg veel output) en bij Plots Survival aanvinken.*

In Figuur 10.2 worden de overlevingskansen van de Breastcancer data grafisch weergegeven. Een dergelijke grafiek wordt **Survival curve** of **Kaplan-Meier curve** genoemd.
In het eerste jaar is er nauwelijks sprake van sterfte (zoals ook al uit Tabel 10.2 bleek), dan zakt de grafiek vrij gelijkmatig in ongeveer 6 jaar tijd naar ongeveer 0,87 waar een zekere stabilisatie lijkt op te treden.

In de praktijk van het wetenschappelijk onderzoek gaat het vaak om het vergelijken van verschillende groepen. Zo kan het ook van belang zijn om te kijken naar verschillen in survival van groepen. Als het er om gaat twee of meerdere groepen met elkaar te vergelijken die onderscheiden zijn door één groeperingsvariabele kan dat met behulp van de Log Rank toets.
Als voorbeeld bekijken we in de dataset Breastcancer.sav twee groepen vrouwen, gedefinieerd door "progesterone receptor status", wel (1) of niet (0) gevoelig voor het hormoon progesteron. In het eerste geval kan de tumor worden bestreden met hormoontherapie. Is deze variabele van invloed op de overleving? Zie Tabel 10.3.

Tabel 10.3 *Log Rank toets voor het vergelijken van de overleving in twee groepen vrouwen met borstkanker, ingedeeld op grond van tumor-gevoeligheid voor progesteron. In SPSS:* ***Analyze – Survival - Kaplan-Meier –*** *Time en Status invullen, Define Event is 1,* ***Options*** *Survival table uitvinken (geeft hier erg veel output),* ***Compare Factor*** *Log Rank*

Case Processing Summary

Progesterone Receptor Status	Total N	N of Events	Censored	
			N	Percent
Negative	389	31	358	92,0%
Positive	462	20	442	95,7%
Overall	851	51	800	94,0%

Means and Medians for Survival Time

Progesterone Receptor Status	Mean[a]				Median			
			95% Confidence Interval				95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
Negative	119,192	2,527	114,238	124,145	.	.	.	.
Positive	123,488	2,096	119,379	127,596	.	.	.	.
Overall	122,055	1,638	118,845	125,265	.	.	.	.

a. Estimation is limited to the largest survival time if it is censored.

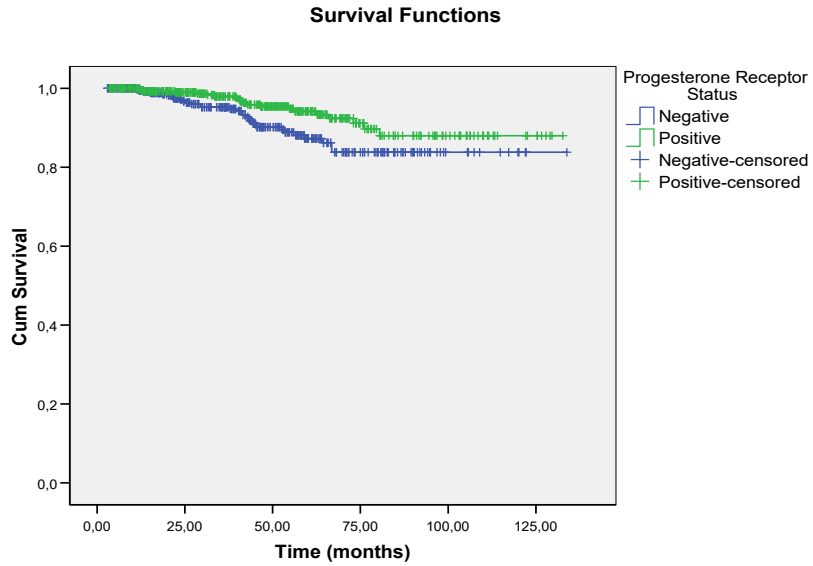
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	5,143	1	,023

Test of equality of survival distributions for the different levels of Progesterone Receptor Status.

In de bovenste tabel van Tabel 10.3 zien we dat het relatieve aantal events in de groep met negatieve receptor status (31 / 389 * 100 % ≈ 8,0 %) hoger is dan in de groep met positieve receptor status (20 / 462 * 100 % ≈ 4,3 %). In de middelste tabel van Tabel 10.3 staan gemiddelde overlevingsduren (in maanden) geschat; deze zijn echter weinig zinvol vanwege de gecensureerde waarden. De mediane overlevingsduren kunnen niet geschat worden omdat meer dan 50 % nog in leven is.
In de onderste tabel van Tabel 10.3 wordt de Log Rank toets uitgevoerd.

De Log Rank toets is een non-parametrische methode die de nulhypothese toetst dat de groepen dezelfde survival curve hebben. De toets is gebaseerd op de gedachte dat, als de nulhypothese juist is, de kans op een event op ieder moment voor elke respondent in beide groepen even groot is. Op ieder tijdstip dat er, in één van de groepen, een event plaats vindt, wordt per groep Observed (wel of geen event) vergeleken met Expected (rekening houdend met het aantal "at risk' in de desbetreffende groep). Deze geobserveerde en verwachte waarden worden gecombineerd tot een chi-kwadraat waarde. Voor een uitgewerkt voorbeeld, zie Altman of Hosmer en Lemeshow (2).
De Log rank toets is in dit geval significant (bij een $\alpha = 0{,}05$) en leidt tot de conclusie dat vrouwen met een positieve progesteron receptor status een significant betere prognose hebben dan vrouwen met een negatieve progesteron receptor status. Zie ook Figuur 10.3.



Figuur 10.3 *Surivival curves voor het vergelijken van de overleving in twee groepen vrouwen met borstkanker, ingedeeld op grond van tumor-gevoeligheid voor progesteron. In SPSS:* ***Analyze – Survival - Kaplan-Meier –*** *Time en Status invullen, Define Event is 1,* ***Options*** *Survival table uitvinken (geeft hier erg veel output), en bij Plots: Survival*

De Log Rank toets is een hypothese toets die geen directe link heeft naar een betrouwbaarheidsinterval. Deze kunnen wel opgesteld worden met behulp van de **hazard ratio** $R = \dfrac{O_1 / E_1}{O_2 / E_2}$, waarin O_i en E_i het geobserveerde en het verwachte aantal events is in groep i. Voor een uitgewerkt voorbeeld, zie Altman of Hosmer en Lemeshow (2).

OPDRACHTEN

1.
Onderzoek of de variabele "oestrogeen receptor" invloed heeft op de survival bij de vrouwen met borstkanker. Laat ook een passende grafiek maken.

10.2 Cox regressie

In de vorige paragraaf zagen we dat de twee groepen vrouwen met verschillende progesteron receptor verschilden in hun prognose. Blijft dat verschil bestaan als we rekening willen houden met andere variabelen zoals bijvoorbeeld leeftijd?
De Log Rank toets stelt ons in staat om verschillende groepen met elkaar te vergelijken met betrekking tot de overlevingsduur. Eventueel kunnen we nog stratificeren naar een tweede categorische variabele, maar rekening houden met een mogelijke confounder die continu is, zoals leeftijd, is niet mogelijk met de Log Rank toets.
Een multivariate techniek die wel die mogelijkheid heeft staat bekend onder de naam **Cox regressie** of ook wel als het **proportional hazards model**. Deze methode is mathematisch complex en in deze paragraaf wordt getracht een beeld te schetsen zonder de pretentie van volledigheid. Een cruciaal begrip in deze benadering is de **hazard functie**. Deze is gerelateerd aan de survival functie, maar geeft het risico op een event gedurende een (zeer kleine) periode. De hazard functie *h(t)* wordt gemodelleerd door middel van de verklarende variabelen volgens formule f9

f9 $$h(t) = h_0(t) * e^{b_1 X_1 + b_2 X_2 + \ldots b_k X_k}$$

$h_0(t)$ in formule f9 wordt de **baseline hazard** genoemd en kan geïnterpreteerd worden als de hazard functie voor een persoon met waarde 0 op alle verklarende variabelen.
Door al deze hazards bij elkaar op te tellen (tot tijdtip t) krijgen we de cumulatieve hazard op tijdstip t:

f10 $$H(t) = H_0(t) * e^{b_1 X_1 + b_2 X_2 + \ldots b_k X_k}$$

De Survival functie S(t) hangt met H(t) samen volgens de formule $S(t) = e^{-H(t)}$.

> Bedenk dat voor een individu dat geen risico loopt, H(t) = 0, en dus de kans op overleven S(t) = e^0 = 1. Omdat e^{-x} een dalende functie is van *x*, geldt: hoe groter H(t), d.w.z. hoe groter het risico, hoe kleiner de kans op overleven.

We zien in formule f10 een multivariaat model waarin op dezelfde wijze als bij lineaire en logistische regressie meerdere variabelen van verschillend meetniveau gebruikt kunnen worden. Er is wederom sprake van een lineaire predictor $b_1 X_1 + \ldots + b_k X_k$, maar die is via een ingewikkelde linkfunctie gekoppeld aan de responsievariabele S(t).
Als we in SPSS een Cox regressie uitvoeren, worden de coëfficiënten van de cumulatieve hazardfunctie geschat. Een positieve coëfficiënt wil dus zeggen een hoger risico en is dus ongunstig voor de overleving.
Laten we om te beginnen een Cox regressie uitvoeren met één verklarende variabele, de progesteron receptor status. In Tabel 10.4 staat een deel van de uitvoer.

Tabel 10.4 *Deel van de uitvoer van Cox regressie van de overlevingsduur op progesteron receptor status. In SPSS:* ***Analyze – Survival – Cox Regression –*** *Time en Status invullen, "pr"bij Covariates*

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	51	4,2%
	Censored	708	58,7%
	Total	759	62,9%
Cases dropped	Cases with missing values	356	29,5%
	Cases with negative time	0	,0%
	Censored cases before the earliest event in a stratum	92	7,6%
	Total	448	37,1%
Total		1207	100,0%

a. Dependent Variable: Time (months)

Block 0: Beginning Block

Omnibus Tests of Model Coefficients

-2 Log Likelihood
610,822

Block 1: Method = Enter

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
605,704	5,141	1	,023	5,118	1	,024	5,118	1	,024

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 610,822

b. Beginning Block Number 1. Method = Enter

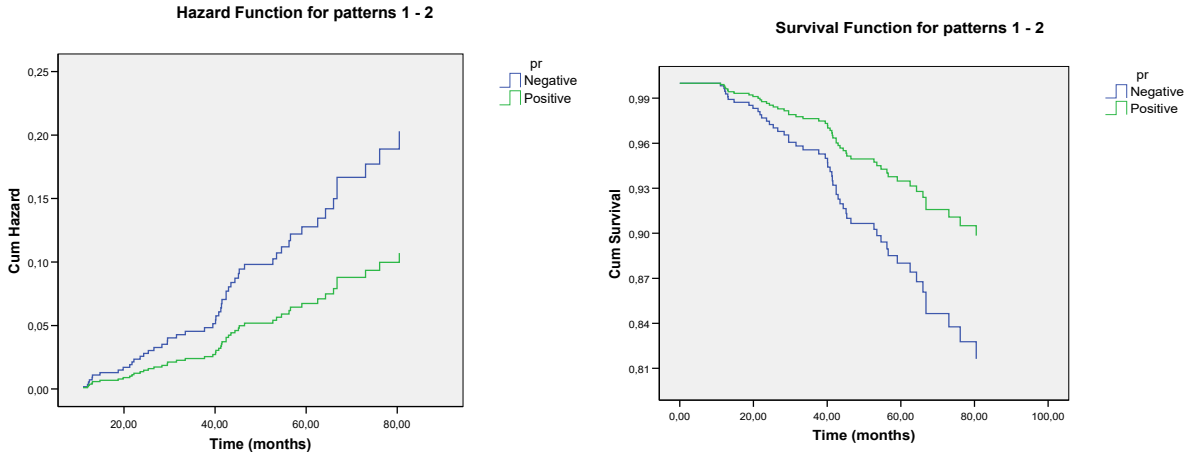
Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
pr	-,639	,287	4,969	1	,026	,528

Een toelichting op de informatie in Tabel 10.4:
- in de bovenste tabel staat beschrijvende informatie, onder andere hoeveel missende cases er zijn (356). Van deze dames was de progesteron receptor status niet bekend. Een verschil met de Log Rank toets is dat de 92 cases die een gecensureerde tijd hebben die korter is dan de tijd tot het eerste event niet in de analyse worden meegenomen. Het risico voor hen is ten slotte toch 0 voor hun observatietijd.
- Het tabelletje onder Block 0, geeft de -2 log Likelihood van het zogenaamde nulmodel, het model zonder verklarende variabelen. Het geeft aan hoe "waarschijnlijk" dat model is. Absoluut gezien zegt dat getal niets, maar het wordt als uitgangspunt gebruikt om de – 2 log

Likelihood van andere modellen mee te vergelijken. Hierbij geldt: hoe kleiner de – 2 log Likelihood, hoe beter het model.
- Daaronder staan toetsresultaten van het volledige model (in dit geval alleen de verklarende variabele progesteron receptor status) ten opzichte van een kleiner model. De nulhypothese luidt dat het volledige model niet meer verklaart dan het nulmodel (linker toets) of dan het model uit de vorige stap (middelste toets) of dan het model zonder het huidige blok met variabelen. In dit geval, nu er maar één verklarende variabele is toegevoegd, vallen deze drie toetsen samen. De berekening van de linkse toets gaat aan de hand van de Score-statistic (die in deze syllabus niet verder besproken wordt), de andere twee met behulp van de -2 log Likelihood. De -2 log Likelihood van dit model is 605,7 wat een verschil is van 610,8 – 605,7 = 5,1. Deze waarde wordt vergeleken met een chi-kwadraat verdeling met k vrijheidsgraden, waarin k het aantal variabelen is dat toegevoegd is ten opzichte van het vorige model; in dit geval k = 1. De P-waarde is hier 0,024, wat bijna hetzelfde is als de P-waarde van de Log Rank toets, en leidt ook hier tot verwerping van de nulhypothese als we alfa = 0,05 nemen.
- In de onderste tabel staat de coëfficiënt van de verklarende variabele geschat: -0,64. Hoe interpreteren we dit getal? In de laatste kolom staat $e^{\beta_1} = e^{-0,64} = 0,53$. Dit is, in het geval van een dichotome verklarende variabele, de **hazard ratio** van de groep die gecodeerd is met 1 (positieve progesteron receptor status) ten opzichte van de groep die 0 gecodeerd is (negatieve progesteron receptor status).

Het is mogelijk in hetzelfde venster van de Cox regressie SPSS grafieken te bestellen, zie Figuur 10.4.

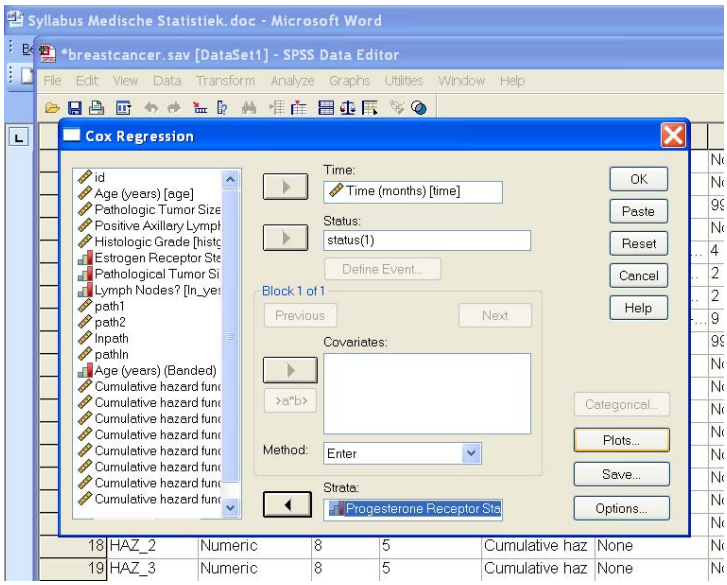


Figuur 10.4 *Grafieken van de geschatte cumulatieve Hazard functie (links) en geschatte survival (rechts) van de groepen vrouwen ingedeeld naar progesteron receptor status, onder aanname van proportional hazards. In SPSS:* ***Analyze – Survival – Cox Regression –*** *Time en Status invullen, "pr"bij Covariates en aangeven dat pr* ***"categorical"*** *is. Vervolgens bij* ***Plots*** *Survival en Hazard aanvinken en "pr"invullen bij* ***seperate lines for***

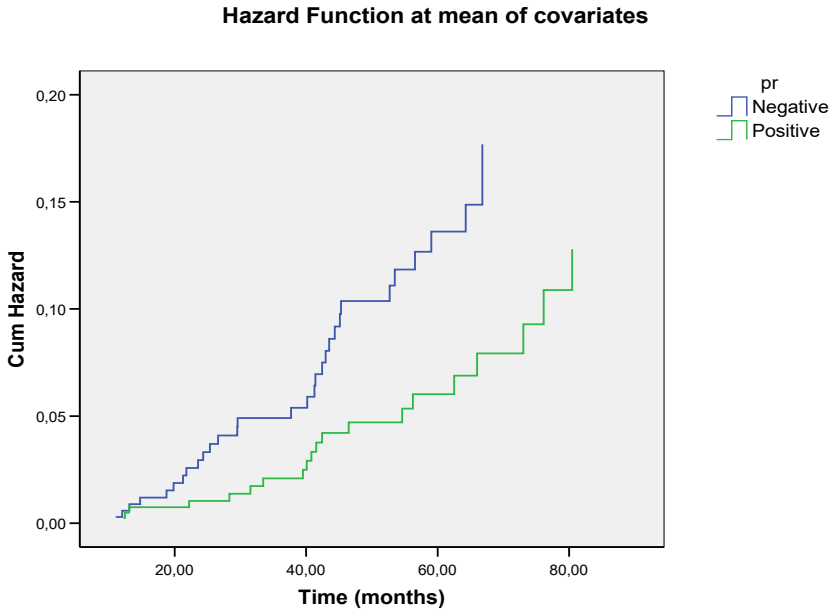
In Figuur 10.4 komen de verschillen in hazard (en dus in overleving) tussen de groepen duidelijk naar voren.

Een voorwaarde voor het uitvoeren van een Cox regressie is dat de hazards van de verschillende groepen proportioneel zijn ten opzichte van elkaar. Als je kijkt naar de grafieken in Figuur 10.4 valt op dat de stappen in de grafieken steeds op hetzelfde tijdstip plaats vinden en de cumulatieve hazard van de negatieve groep steeds (ongeveer) tweemaal zo groot is als van de positieve groep. De afgedrukte grafieken zijn dan ook geschat onder

aanname van de proportionaliteit van de hazards! Voordat we de resultaten van deze analyse als valide kunnen bestempelen zullen we eerst die aanname van proportionaliteit moeten controleren. Als we de variabele Pr niet bij de Covariates invullen, maar bij Strata (zie Figuur 10.5) dan krijgen we aparte cumulatieve hazards per subgroep te zien, zie Figuur 10.6).



Figuur 10.5 *Pr bij strata om de cumulatieve hazard per subgroep te schatten zonder de aanname van proportionaliteit.*



Figuur 10.6 *De cumulatieve hazards per groep zonder aanname van proportionaliteit.*

Uit Figuur 10.6 kunnen we opmaken dat de cumulatieve hazards redelijk proportioneel zijn (het is belangrijk dat de grafieken elkaar niet snijden) en de Cox regressie valide is.

Het opbouwen van een model met meerdere verklarende variabelen gaat bij een Cox regressie op dezelfde wijze als bij een lineaire en logistische regressie. Tot slot een voorbeeld van een

meervoudige Cox regressie: wat gebeurt er als we rekening houden met de leeftijd van de vrouwen? Zie Tabel 10.5.

Tabel 10.5 *Deel van de uitvoer van Cox regressie van de overlevingsduur op progesteron receptor status en leeftijd. In SPSS:* ***Analyze – Survival – Cox Regression –*** *Time en Status invullen, "pr" en "age" bij Covariates*

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
595,895	14,671	2	,001	14,927	2	,001	14,927	2	,001

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 610,822

b. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
pr	-,491	,291	2,850	1	,091	,612
age	-,035	,012	9,230	1	,002	,965

Uit Tabel 10.5 kunnen we het volgende concluderen:

- ten opzichte van het nulmodel is dit een significant verklarend model (chi-kwadraat is 14,9 bij twee vrijheidsgraden)
- ten opzichte van het model met alleen pr als verklarende variabele is de -2 log Likelihood 605,7 – 595,9 = 9,8 afgenomen, zie Tabel 10.4. Dat is significant bij $\alpha = 0,05$ (zie de tabel van de chi-kwadraat verdeling met 1 df).
- leeftijd is signifcant ($p < 0,05$). Oudere vrouwen hebben een lagere hazard!
- progesteron receptor status is niet langer significant op 5 % niveau

OPDRACHTEN

1.
Toets met behulp van de Log Rank toets of de survival voor de verschillende groepen op grond van histologische gradering verschilt.

2.
Dezelfde vraag als 1, maar gebruik nu Cox regressie. Bekijk ook de voorwaarden.

3.
Maak met behulp van Cox regressie het best verklarende model voor de survival in de dataset Breastcancer.sav.

Appendix A.

Het meetkundig gemiddelde van n positieve getallen is gedefinieerd als de n-de machtswortel van het product van deze n getallen. Als voorbeeld nemen we de getallen 4, 9 en 11. Het "gewone" gemiddelde, officieel het ongewogen rekenkundig gemiddelde, wordt bepaald door de getallen op te tellen en te delen door drie: (4 + 9 + 11)/3 = 24 / 3 = 8.
Het meetkundig gemiddelde in dit voorbeeld is de derde machtswortel van de getallen 4, 9 en 11: $\sqrt[3]{4*9*11}$ = 7,34. Dit meetkundig gemiddelde wordt onder andere gebruikt bij het berekenen van de gemiddelde groeifactor van exponentiële groei.

Als voor een statistische techniek een variabele logaritmisch wordt getransformeerd waarna het gemiddelde wordt bepaald op de logaritmische schaal en vervolgens weer terug wordt getransformeerd naar de oorspronkelijke schaal, krijg je het meetkundig gemiddelde van de oorspronkelijke waarnemingen:
noem de oorspronkelijke waarnemingen $X_1, X_2, ..., X_n$. Na transformatie krijg je nieuwe getallen $Y_i = \ln(X_i)$. Voor het gemiddelde van de getransformeerde waarden geldt:

$$\overline{Y} = \sum \frac{Y_i}{n} = \sum \frac{\ln(X_i)}{n} = \frac{1}{n} \sum \ln(X_i) = \frac{1}{n} \ln(\Pi X_i) = \ln \sqrt[n]{\Pi X_i}$$

(Want ln(a) + ln(b) = ln(a*b), n*ln(a) = $\ln(a^n)$ en de n-de machtswortel van een getal kan ook worden geschreven als dat getal tot de macht 1/n).

Terug transformatie van het gemiddelde van Y naar de oorspronkelijke schaal geeft dan $e^{\overline{Y}} = \exp(\ln \sqrt[n]{\Pi X_i}) = \sqrt[n]{\Pi X_i}$, het meetkundig gemiddelde van de oorspronkelijke waarnemingen.

Appendix B

Formeel laten we nog zien dat bij een logistische regressie met responsievariabele Y de coëfficiënt β_1 van een dichotome verklarende variabele X met codes 1 en 0 geïnterpreteerd kan worden met behulp van de odds ratio.

De kans op de gebeurtenis Y = 1, gegeven de variabele X, is gedefinieerd als

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Om de odds te bepalen moeten we ook weten wat 1 - $\pi(x)$, de kans dat Y = 0, is:

$$1-\pi(x) = 1-\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1+e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{\beta_0+\beta_1 x}}$$

De odds wordt dan:

$$\frac{\pi(x)}{1-\pi(x)} = \frac{e^{\beta_0+\beta_1 x} \big/ 1+e^{\beta_0+\beta_1 x}}{1 \big/ 1+e^{\beta_0+\beta_1 x}} = e^{\beta_0+\beta_1 x}$$

Om de odds ratio te kunnen berekenen moeten we de odds van de groep met x = 1 delen door de odds van de groep met x = 0.
Voor de groep met code x = 1 geldt dat de odds gelijk is aan $e^{\beta_0+\beta_1}$,
Voor de groep met code x = 0 geldt dat de odds gelijk is aan e^{β_0}.
Derhalve is de odds ratio: $\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_0+\beta_1-\beta_0} = e^{\beta_1}$.

Referenties

Altman D.G. *Practical statistics for medical research* Chapman & Hall (1991)
Conover W.J. *Practical nonparametric statistics* Wiley (1999)
Hosmer D. and Lemeshow S. (1) *Applied Logistic Regression* Wiley (2000)
Hosmer D. and Lemeshow S. (2) *Applied Survival Analysis* Wiley (1999)
Swinscow T.D.V en Campbell M.J.. *Statistics at square one* BMJ Books (2002)