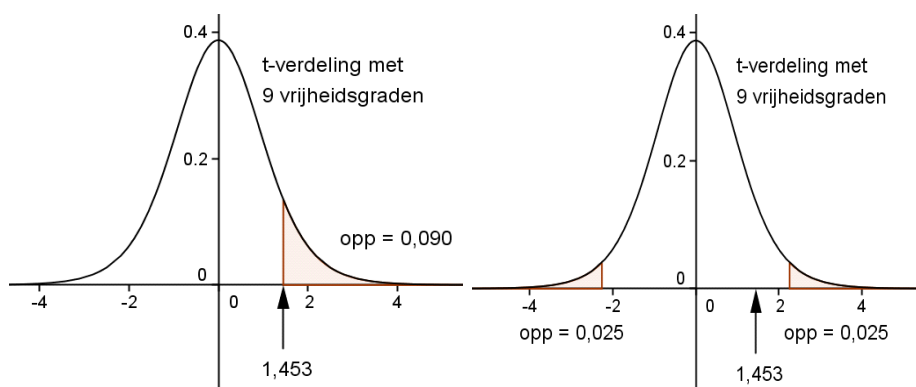


# Levende Statistiek

Een module voor Wiskunde D VWO

Jacob van Eeghen en Liesbeth de Wreede



© Jacob van Eeghen en Liesbeth de Wreede, Leiden 2010  
© cTWO, Utrecht 2010

Dit lesmateriaal kan gebruikt worden voor de invulling van “wiskunde in wetenschap” of van het keuzeonderwerp Wiskunde D op het VWO. De ontwikkeling van het materiaal is financieel mogelijk gemaakt door het Stedelijk Gymnasium Leiden (Van Eeghen), de Afdeling Medische Statistiek en Bio-informatica van het LUMC (De Wreede), cTWO en VVS-OR.

De gebruiker mag het werk kopiëren, verspreiden, doorgeven en remixen (afgeleide werken maken) onder de volgende voorwaarden:

- Naamsvermelding. De gebruiker dient bij het werk de door de maker of de licentiegever aangegeven naam te vermelden (maar niet zodanig dat de indruk gewekt wordt dat zij daarmee instemmen met uw werk of uw gebruik van het werk).
- Niet-commercieel. De gebruiker mag het werk niet voor commerciële doeleinden gebruiken.
- Gelijk delen. Indien de gebruiker het werk bewerkt kan het daaruit ontstane werk uitsluitend krachtens dezelfde licentie als de onderhavige licentie of een gelijksoortige licentie worden verspreid.

Versie 10 september 2010

# Inhoudsopgave

<b>Voorwoord</b>	<b>6</b>
<b>1 Kansverdelingen</b>	<b>9</b>
1.1 Discrete en continue kansverdelingen . . . . .	10
1.2 Verwachtingswaarde, variantie en standaardafwijking . . . . .	13
1.3 Kansvariabelen en rekenregels . . . . .	15
1.4 De centrale limietstelling . . . . .	21
1.5 Van de normale verdeling afgeleide verdelingen . . . . .	23
1.5.1 De $\chi^2$ -verdeling (chi-kwadraatverdeling) . . . . .	23
1.5.2 De t-verdeling (student-t-verdeling) . . . . .	25
<b>2 Schattingen van parameters en betrouwbaarheidsintervallen</b>	<b>27</b>
2.1 Normaal verdeelde populaties . . . . .	28
2.2 Populatiepercentages . . . . .	35
<b>3 t-toetsen voor populatiegemiddeldes</b>	<b>39</b>
3.1 De t-toets : één populatiegemiddelde . . . . .	40
3.2 De t-toets: vergelijking van twee populatiegemiddeldes, ongepaarde waarnemingen . . . . .	43
3.3 De t-toets: vergelijking van twee populatiegemiddeldes, gepaarde waarnemingen . . . . .	48
<b>4 Niet-parametrische toetsen</b>	<b>53</b>
4.1 Gepaarde waarnemingen: de tekentoets . . . . .	53
4.2 Gepaarde waarnemingen: Wilcoxon's rangteken toets . . . . .	55
4.3 Ongepaarde waarnemingen. De toets van Wilcoxon (of van Mann-Whitney) . . . . .	61
4.4 Eén populatie: toetsen betreffende de mediaan . . . . .	66
<b>5 Het vergelijken van populatiepercentages</b>	<b>67</b>
5.1 Vergelijking van twee populatiepercentages (ongepaarde waarnemingen, grote aantallen) . . . . .	67

## Inhoudsopgave

5.2	Vergelijking van twee populatiepercentages (ongepaarde waarnemingen, kleine aantallen): Fishers exacte toets . . . . .	70
5.3	Vergelijking van twee populatiepercentages: betrouwbaar- heidsinterval bij grote aantallen . . . . .	72
5.4	Vergelijking van populatiepercentages bij gepaarde waarne- mingen: McNemars toets . . . . .	75
<b>6</b>	<b>Enkelvoudige lineaire regressie</b>	<b>79</b>
6.1	Criteria voor de best passende lijn . . . . .	80
6.2	Afleiding van de lijn volgens het principe van kleinste kwadraten	83
6.3	Regressieberekeningen met de GR . . . . .	86
6.4	Een stochastisch regressiemodel . . . . .	89
6.5	Eigenschappen van $\hat{\alpha}$ en $\hat{\beta}$ . . . . .	92
6.6	Betrouwbaarheidsintervallen en toetsen voor $\beta$ . . . . .	94
6.7	Betrouwbaarheidsintervallen voor $\mu_0$ en $Y_0$ . . . . .	98
6.8	Als de $X_i$ 's kansvariabelen zijn . . . . .	101
<b>7</b>	<b>Correlatie</b>	<b>103</b>
7.1	De correlatiecoëfficiënt . . . . .	106
7.2	Een stochastisch model . . . . .	108
7.3	Interpretatie van $\hat{\beta}$ en $r$ . . . . .	111
7.4	Kwantitatieve interpretatie van de correlatiecoëfficiënt bij re- gressieanalyse . . . . .	115
<b>8</b>	<b>Meervoudige lineaire regressie</b>	<b>117</b>
8.1	Een model met twee verklarende variabelen . . . . .	118
8.2	Het algemene geval met $k$ verklarende variabelen . . . . .	121
8.3	Een voorbeeld . . . . .	123
8.4	Is het model adequaat? . . . . .	125
8.5	Selectie van verklarende variabelen . . . . .	126
8.6	Categorische variabelen . . . . .	128
8.7	Niet-lineaire regressie . . . . .	130
<b>9</b>	<b>Gemengde opgaven</b>	<b>131</b>
<b>10</b>	<b>R-opgaven</b>	<b>141</b>
10.1	Inleiding . . . . .	141
10.1.1	Installeren en gebruiken . . . . .	141
10.2	Computeropgaven . . . . .	144
10.2.1	Opgaven beschrijvende statistiek . . . . .	144
10.2.2	Opgaven met toetsen en schatten . . . . .	147
	<b>Appendices</b>	<b>152</b>

*Inhoudsopgave*

**Verantwoording**

**157**



# Voorwoord

De titel van deze module is Levende Statistiek. Daar zijn twee redenen voor. Ten eerste biedt de module voldoende stof om je in staat te stellen in veel praktische situaties een correcte statistische analyse uit te voeren. Na bestudering van deze module zul je je kennis niet alleen op schoolvoorbeelden kunnen toepassen, maar ook op situaties uit "het echte leven". Een voorbeeld daarvan kunnen de gegevens zijn die jij, of een medeleerling, hebt verzameld voor een profielwerkstuk. Ten tweede zijn veel voorbeelden uit deze module afkomstig uit de geneeskunde en de biomedische wetenschap, kortom uit de wetenschappen van "het leven".

De noodzakelijke voorkennis voor deze module bestaat uit de voor wiskunde D verplichte onderwerpen uit de kansrekening. Wat statistiek betreft moet je vertrouwd zijn met de beginselen van één- en tweezijdig toetsen, met de Z-toets (gebaseerd op de normale verdeling met gegeven standaardafwijking) en de binomiaaltoets.

In hoofdstuk 1 wordt het begrip kansverdeling opgefrist en maak je kennis met twee nieuwe soorten kansverdeling. In de hoofdstukken 2 t/m 8 worden de belangrijkste statistische technieken gepresenteerd. In hoofdstuk 9 vind je gemengde opgaven, waarin je deze technieken door elkaar moet toepassen. Zolang het aantal waarnemingen niet te groot is, kun je alle noodzakelijke berekeningen uitvoeren met de GR. Bij de verwerking van grotere bestanden heb je een computer nodig. In hoofdstuk 10 leer je hoe je het softwarepakket R daarvoor kunt gebruiken. R kun je gratis downloaden.

De lay-out van dit materiaal maakt beperkt gebruik van kleuren. Ook als het in grijstinten wordt afgedrukt is het goed bruikbaar.

Veel succes en plezier toegewenst.

## *Voorwoord*



# Hoofdstuk 1

## Kansverdelingen

In je eerste kennismaking met statistiek heb je gezien hoe statistiek gebruikt wordt om, op basis van een steekproef, uitspraken te doen over een (grote) populatie. Je hebt ook gezien dat kansrekening, en met name kansverdelingen en hun parameters, een essentieel hulpmiddel voor statistiek zijn. De onderwerpen die in dit hoofdstuk aan de orde komen behoren dan ook niet tot de statistiek, maar tot de kansrekening. In latere hoofdstukken, waar we ons nog uitsluitend met statistiek zullen bezig houden die in de praktijk toepasbaar is, zul je deze onderwerpen uit de kansrekening steeds weer tegenkomen.

In dit hoofdstuk herhalen we eerst het begrip kansverdeling en de daarvan afhankelijke grootheden zoals verwachtingswaarde, variantie en standaardafwijking. We maken daarbij onderscheid tussen discrete en continue kansvariabelen. Bij discrete kansvariabelen kunnen we exact te werk gaan, kunnen we bewijzen wat we beweren en kunnen we zelf alle berekeningen uitvoeren. Om hetzelfde bij continue kansvariabelen te doen, moet je meer van integraalrekening weten dan in het vwo-curriculum wordt aangeboden. Jullie zullen daarom genoeg moeten nemen met zinnen als: “Je kunt bewijzen dat...”. Laat je niet afschrikken door de integralen die je in dit hoofdstuk aantreft. Ze staan er alleen ter illustratie van het feit dat de rekenprincipes voor continue en discrete kansvariabelen hetzelfde zijn, maar je hoeft niet zelf met integralen te kunnen rekenen.

Waarschijnlijk is de normale verdeling de enige continue verdeling die jullie tot dusverre zijn tegengekomen. Aan het eind van dit hoofdstuk wordt een aantal andere continue verdelingen geïntroduceerd, omdat je deze later in deze module nodig zullen hebben. Deze nieuwe verdelingen zijn op de een of andere manier gekoppeld aan de normale verdeling. Net zoals je gewend bent bij de normale verdeling, zul je voor het rekenwerk met deze nieuwe verdelingen gebruik kunnen maken van de GR.

## 1.1 Discrete en continue kansverdelingen

We gooien met een dobbelsteen. Het aantal ogen dat bovenkomt noemen we  $X$ .  $X$  is een *kansvariabele* (of *stochast* of *stochastische variabele*, in het Engels: *random variable*). De mogelijke waarden die  $X$  kan aannemen zijn de gehele getallen 1 t/m 6. Dat is een eindig aantal, daarom noemen we  $X$  een *discrete* stochast. Het is gebruikelijk om de kansvariabele zelf aan te geven met een hoofdletter, en de mogelijke uitkomsten (of realisaties) met een kleine letter. Als de dobbelsteen zuiver is, kunnen we de kans op elk van de mogelijke uitkomsten van  $X$  in een tabel zetten:

$x$	1	2	3	4	5	6
$\Pr(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

We zeggen dat een dergelijke tabel de *kansverdeling van  $X$*  voorstelt (Engels: *probability distribution*). In zo'n tabel moeten alle mogelijke uitkomsten met hun bijbehorende kans voorkomen. De optelsom van alle kansen moet uiteraard gelijk zijn aan 1. Met behulp van de kansverdeling kunnen we de kans op alle gebeurtenissen die door  $X$  worden beschreven uitrekenen. Zo is bijvoorbeeld

$$\Pr(X \text{ is even}) = \Pr(X = 2) + \Pr(X = 4) + \Pr(X = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Bij de kansverdeling hierboven zijn we er vanuit gegaan dat de dobbelsteen zuiver is. Als dat niet het geval is, krijg je een andere kansverdeling. Als het vlak waar de 1 op staat extra zwaar is gemaakt, zal de 1 vaak onder liggen en de 6 vaak boven. De kansverdeling zal er dan bijvoorbeeld zo uit kunnen zien:

$x$	1	2	3	4	5	6
$\Pr(X=x)$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{7}{16}$

### Opgaven

1. Je gooit met twee zuivere dobbelstenen. De som van het aantal ogen dat op beide dobbelstenen bovenkomt noemen we  $Y$ . Stel de kansverdeling op van  $Y$ .
2.  $X$  is binomiaal verdeeld met  $n = 4$  en  $\pi = \frac{1}{4}$ . Stel de kansverdeling op van  $X$ . (In deze module gebruiken we voor de kans op succes niet de letter  $p$ , zoals je misschien gewend bent, maar de Griekse letter  $\pi$ .)

We zullen in het vervolg de volgende notatie gebruiken bij discrete kansvariabelen  $X$ . Het aantal mogelijke uitkomsten van  $X$  noemen we  $n$  en deze mogelijke uitkomsten worden aangeduid met  $x_1, x_2, x_3, \dots, x_n$ . De kans op

### 1.1. Discrete en continue kansverdelingen

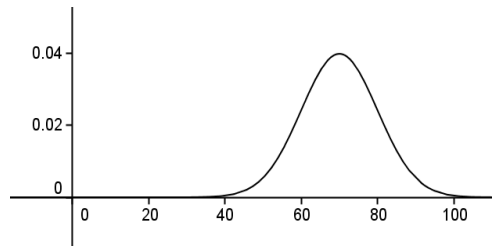
uitkomst  $x_i$  is  $\Pr(X = x_i) = p(x_i)$ . In plaats van  $p(x_i)$  schrijven we ook wel  $p_i$ . Voor alle  $i = 1, 2, \dots, n$  moet dus gelden:  $0 \leq p_i \leq 1$  en  $\sum_{i=1}^n p_i = 1$ . In de meeste gevallen waarin we het sigmateken gebruiken zal het vanzelfsprekend zijn dat we sommeren van  $i=1$  t/m  $n$ . We schrijven dan kortweg  $\sum p_i = 1$  of  $\sum_i p_i = 1$ .

Je gaat op een willekeurige dag naar een willekeurige groenteboer en koopt er een willekeurige appel. Het gewicht in grammen van deze appel noemen we  $Z$ .  $Z$  is een kansvariabele met een oneindig aantal mogelijke uitkomsten: alle reële getallen tussen 0 en 1000 (als je veronderstelt dat appels van meer dan een kilo niet bestaan). In een dergelijk geval zeggen we dat  $Z$  een *continue* kansvariabele is.

Bij een continue kansvariabele kunnen we geen bruikbare tabel maken met de kansen van alle mogelijke waarden van  $Z$ , maar gebruiken we de *kansdichtheid* (Engels: *probability density function*).

Hiernaast zie je een plaatje van een mogelijke kansdichtheid  $f(x)$  van  $Z$ . Deze kansdichtheid is die van een normale verdeling met gemiddelde 70 en standaardafwijking 10. In dat geval geldt:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



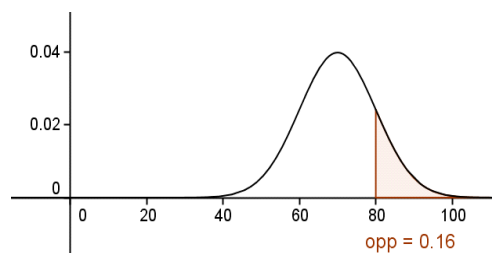
met  $\mu = 70$  en  $\sigma = 10$ .

Uiteraard staat het niet op voorhand vast dat  $Z$  normaal verdeeld is. Als  $Z$  een andere verdeling heeft, dan hebben we te maken met een ander plaatje en een andere functie  $f(x)$  voor de kansdichtheid. De enige algemene eisen die aan een kansdichtheid  $f(x)$  worden gesteld zijn: (i) het is een functie die geen negatieve waarden aanneemt, ofwel  $f(x) \geq 0$  voor alle  $x$  en (ii) de oppervlakte tussen de kromme en de  $x$ -as is gelijk aan 1, ofwel  $\int_{-\infty}^{\infty} f(x)dx = 1$ .

Als de kansdichtheid van een kansvariabele bekend is, kun je de kansen die door deze kansvariabele gedefinieerd worden berekenen met behulp van integreren.

## Kansverdelingen

Zo is de kans dat de appel meer dan 80 gram weegt gelijk aan  $\Pr(Z > 80) = \int_{80}^{\infty} f(x)dx$ . Als  $f$  de kansdichtheid is die hoort bij de genoemde normale verdeling, is deze kans ongeveer gelijk aan 0,16.



De kans dat een appel precies 80 gram weegt ( $\Pr(Z = 80)$ ) is gelijk aan de oppervlakte van een streepje tussen de kromme en de  $x$ -as. Deze oppervlakte (en dus de kans) is gelijk aan nul. Toch betreft het geen onmogelijke gebeurtenis.

Je ziet dat het berekenen van kansen bij een continue verdeling ingewikkelder is dan bij een discrete verdeling. Bij een discrete verdeling maak je een optelsommetje aan de hand van een tabel, bij een continue verdeling moet je een integraal uitrekenen. Dit laatste is lang niet altijd eenvoudig. Zelfs bij de meest voorkomende continue verdeling (de standaardnormale) zijn de meeste integralen niet exact te berekenen, maar kun je de betreffende waarden alleen numeriek benaderen. De GR kan daarbij behulpzaam zijn.

**Opmerking 1.1.1.** Het is van belang om de overeenkomsten te zien tussen een discrete en een continue kansvariabele. Bij een continue stochast wordt de tabel vervangen door een functie. In de tabel komen alleen positieve getallen voor en ook de kansdichtheid mag alleen positieve waarden aannemen. In de tabel moet de optelsom van alle waarden gelijk zijn aan 1, de oppervlakte onder de kansdichtheid moet ook gelijk zijn aan 1. De waarden  $p_i$  uit de kansverdeling van een discrete stochast zijn kansen. De waarden  $f(x)$  uit de kansdichtheid van een continue stochast zijn dat niet. Los geformuleerd (dus intuïtief nuttig maar niet streng wiskundig) kun je de grootheden  $f(x)dx$  (de oppervlakte van oneindig dunne rechthoekjes onder de kromme van de kansverdeling) opvatten als kansen.  $\int_{-\infty}^{\infty} f(x)dx$  stelt dan de oneindige som voor van oneindig kleine kansjes. Deze som moet gelijk zijn aan 1, net zoals  $\sum p_i = 1$ .

## 1.2 Verwachtingswaarde, variantie en standaardafwijking

De *verwachtingswaarde* (Engels: *expectation* of: *mean*) van een kansvariabele  $X$  is de “gemiddelde uitkomst” van deze variabele en wordt genoteerd als  $E(X)$ .

Bij een *discrete* variabele kunnen we, in de notatie van paragraaf 1.1, de verwachtingswaarde als volgt definiëren:

$$E(X) = \sum p_i x_i.$$

In het geval van de stochast  $X$  uit paragraaf 1.1 (de zuivere dobbelsteen) geldt  $E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \dots + \frac{1}{6} \cdot 6 = 3\frac{1}{2}$ .

### Opgaven

3. Bereken  $E(Y)$  voor de stochast  $Y$  uit opgave 1.
4. Bereken  $E(X)$  voor de onzuivere dobbelsteen uit paragraaf 1.1.
5. Bereken  $E(X)$  voor de kansvariabele  $X$  uit opgave 2.

Voor een *continue* kansvariabele  $X$  met kansdichtheid  $f(x)$  definiëren we

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx.$$

**Opmerking 1.2.1.** Let ook hier weer op de overeenkomst tussen het discrete en het continue geval:  $E(X) = \sum p_i x_i$  is een eindige som, waarbij alle mogelijke uitkomsten  $x_i$  het gewicht meekrijgen van de kans  $p_i$  op deze uitkomst.  $E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$  is een oneindige som waarin elke mogelijke uitkomst  $x$  het gewicht meekrijgt van de oneindig kleine kans  $f(x)dx$  op deze uitkomst.

Als  $f(x)$  de kansdichtheid is van een normaal verdeelde variabele  $X$  met parameters  $\mu$  en  $\sigma$ , dan kun je, met behulp van de symmetrie van de kansdichtheid t.o.v. de lijn  $x = \mu$ , bewijzen dat  $E(X) = \mu$ .

Voor een kansvariabele  $X$ , discreet of continu, wordt de *variantie* (Engels: *variance*)  $Var(X)$  gedefinieerd door:

$$Var(X) = E((X - E(X))^2).$$

## Kansverdelingen

Van de variantie wordt de *standaardafwijking* of *standaarddeviatie* (Engels: *standard deviation*)  $\sigma(X)$  afgeleid:

$$\sigma(X) = \sqrt{\text{Var}(X)}.$$

In plaats van  $\sigma(X)$  wordt ook wel  $\sigma_X$  geschreven.

Deze definities, hoewel in principe bekend, behoeven wellicht een korte toelichting aan de hand van het volgende voorbeeld.

**Voorbeeld 1.2.1.** We bekijken de kansvariabele  $X$  uit paragraaf 1.1 ( $X$  is het aantal ogen na een worp met een zuivere dobbelsteen) en we willen  $\text{Var}(X)$  en  $\sigma(X)$  berekenen.

We hebben al berekend dat  $E(X) = 3\frac{1}{2}$ , dus uit de definitie van variantie volgt in dit geval  $\text{Var}(X) = E((X - 3\frac{1}{2})^2)$ . In deze formule kunnen we drie kansvariabelen onderscheiden:  $X$ ,  $Y = X - 3\frac{1}{2}$  en  $Z = Y^2 = (X - 3\frac{1}{2})^2$ . Van elk van deze kansvariabelen kunnen we de kansverdeling opstellen:

Voor  $X$ :

$x$	1	2	3	4	5	6
$\text{Pr}(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Voor  $Y = X - 3\frac{1}{2}$ :

$y$	$-2\frac{1}{2}$	$-1\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$1\frac{1}{2}$	$2\frac{1}{2}$
$\text{Pr}(Y=y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

En voor  $Z = Y^2 = (X - 3\frac{1}{2})^2$

$z$	$6\frac{1}{4}$	$2\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$2\frac{1}{4}$	$6\frac{1}{4}$
$\text{Pr}(Z=z)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

hetgeen we beter kunnen schrijven als:

$z$	$\frac{1}{4}$	$2\frac{1}{4}$	$6\frac{1}{4}$
$\text{Pr}(Z=z)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

De laatste tabel gebruiken we om  $\text{Var}(X)$  te berekenen:

$$\text{Var}(X) = E((X - 3\frac{1}{2})^2) = E(Z) = \frac{1}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot 2\frac{1}{4} + \frac{1}{3} \cdot 6\frac{1}{4} = 2\frac{11}{12}$$

Verder geldt:

$$\sigma(X) = \sqrt{\text{Var}(X)} = \sqrt{2\frac{11}{12}} = \frac{1}{6}\sqrt{105} (\approx 1,71)$$

### 1.3. Kansvariabelen en rekenregels

In bovenstaand voorbeeld heb je gezien hoe je de variantie van een discrete stochast kunt berekenen door eerst de kansverdeling van  $(X - E(X))^2$  op te stellen. Bij continue stochasten heb je integraalrekening nodig en geldt:  $Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$ , waarbij  $\mu = E(X)$ .

Als  $f(x)$  de kansdichtheid is van een normaal verdeelde variabele  $X$  met parameters  $\mu$  en  $\sigma$ , dan kun je bewijzen dat  $Var(X) = \sigma^2$ .

**Opmerking 1.2.2.** In de definitie  $Var(X) = E((X - E(X))^2)$  staan heel wat haakjes. Meestal verkorten we dit tot  $Var(X) = E(X - E(X))^2$ . Hier wordt dan impliciet bedoeld “eerst kwadrateren en dan de verwachtingswaarde nemen”.

Als we het omgekeerde bedoelen (“eerst de verwachtingswaarde nemen en dan kwadrateren”), schrijven we meestal  $E^2(Y)$  i.p.v.  $(E(Y))^2$ .

**Opmerking 1.2.3.** Variantie en standaardafwijking zijn zogenaamde spreidingsmaten. Ze geven een indicatie van de mate waarin de mogelijke uitkomsten van een stochast verspreid liggen rond het gemiddelde.

**Opmerking 1.2.4.** In opgave 11 ga je een formule afleiden waarmee je een berekening als in Voorbeeld 1.2.1 kunt vereenvoudigen.

#### Opgaven

6. Bereken  $Var(X)$  en  $\sigma(X)$  in het voorbeeld van de onzuivere dobbelsteen uit paragraaf 1.1.
7. Bereken  $Var(X)$  en  $\sigma(X)$  als  $X$  de binomiaal verdeelde stochast is uit opgave 2.

### 1.3 Kansvariabelen en rekenregels

We beschouwen twee discrete kansvariabelen  $X$  en  $Y$ .  $X$  kan de waarden 1, 2, 3 en 4 aannemen en  $Y$  de waarden 1, 2 en 3. De kansverdelingen van  $X$  en  $Y$  vind je hieronder.

$x$	1	2	3	4
$\Pr(X=x)$	0,2	0,2	0,3	0,3

$y$	1	2	3
$\Pr(Y=y)$	0,3	0,3	0,4

## Kansverdelingen

We willen nu kansen berekenen waar zowel  $X$  als  $Y$  in voorkomen. Kennis over de kansverdelingen van  $X$  en  $Y$  alleen is hiervoor niet genoeg. We moeten de *samengestelde kansverdeling* (Engels: *joint probability distribution*) van  $X$  en  $Y$  kennen. Dit is een tabel waarin de kansen vermeld staan voor alle *combinaties* van mogelijke waarden van  $X$  en  $Y$ . Hieronder vind je een voorbeeld van een dergelijke tabel.

### Voorbeeld van afhankelijke $X$ en $Y$

	$p_{ij} = \Pr(X = x_i \wedge Y = y_j)$			
$\downarrow x_i \backslash y_j \rightarrow$	1	2	3	$\Pr(X = x_i)$
1	0,09	0,05	0,06	0,2
2	0,07	0,06	0,07	0,2
3	0,08	0,09	0,13	0,3
4	0,06	0,1	0,14	0,3
$\Pr(Y = y_j)$	0,3	0,3	0,4	$\sum_{i,j} p_{ij} = 1$

In deze tabel zie je dat de som van alle kansen gelijk is aan 1. Je ziet ook hoe je uit de samengestelde kansverdeling van  $X$  en  $Y$  de (enkelvoudige) kansverdelingen van  $X$  en  $Y$  afzonderlijk kunt afleiden: zie de randtotalen. De tabel levert ook voorwaardelijke kansen. Zo is de kans dat  $X=4$  gegeven dat  $Y=2$  gelijk aan:

$$\Pr(X = 4|Y = 2) = \frac{\Pr(X = 4 \wedge Y = 2)}{\Pr(Y = 2)} = \frac{0,1}{0,3} = \frac{1}{3}.$$

Deze kans is ongelijk aan  $\Pr(X = 4) = 0,3$ , dus we kunnen concluderen dat de stochasten  $X$  en  $Y$  *afhankelijk* zijn. Als  $X$  en  $Y$  *onafhankelijk* zijn, moet voor alle  $x_i$  en  $y_j$  gelden:

$$p_{ij} = \Pr(X = x_i \wedge Y = y_j) = \Pr(X = x_i) \cdot \Pr(Y = y_j)$$

De tabel voor de samengestelde kansverdeling van de onafhankelijke  $X$  en  $Y$  komt er dan als volgt uit te zien.

### Voorbeeld van onafhankelijke $X$ en $Y$

	$p_{ij} = \Pr(X = x_i \wedge Y = y_j)$			
$\downarrow x_i \backslash y_j \rightarrow$	1	2	3	$\Pr(X = x_i)$
1	0,06	0,06	0,08	0,2
2	0,06	0,06	0,08	0,2
3	0,09	0,09	0,12	0,3
4	0,09	0,09	0,12	0,3
$\Pr(Y = y_j)$	0,3	0,3	0,4	$\sum_{i,j} p_{ij} = 1$



### 1.3. Kansvariabelen en rekenregels

Als de samengestelde kansverdeling van  $X$  en  $Y$  eenmaal bekend is, kunnen allerhande kansen, verwachtingswaarden, varianties die betrekking hebben op  $X$  en  $Y$  worden berekend.

Zo is, in het voorbeeld van de *afhankelijke*  $X$  en  $Y$  hierboven,

$$\Pr(X + Y \leq 3) = 0,09 + 0,05 + 0,07 = 0,21$$

en

$$E(X + Y) = 0,09(1 + 1) + 0,05(1 + 2) + \dots + 0,14(4 + 3) = 4,8$$

#### Opgaven

8. Bereken  $E(X + Y)$  in het voorbeeld van de onafhankelijke  $X$  en  $Y$ . Bereken ook  $E(X)$  en  $E(Y)$ . Wat valt op?
9. Bereken  $E(XY)$  in het voorbeeld van de afhankelijke én in het voorbeeld van de onafhankelijke  $X$  en  $Y$ . Wat valt op?

Opgave 8 doet vermoeden dat voor alle kansvariabelen  $X$  en  $Y$  (of ze nu afhankelijk of onafhankelijk zijn) geldt:  $E(X + Y) = E(X) + E(Y)$ . We zullen, voor discrete kansvariabelen, aantonen dat dit inderdaad altijd geldt.

We gaan daarbij uit van de samengestelde kansverdeling van  $X$  en  $Y$ . Deze wordt geheel bepaald door de waarden van de  $x_i$ 's (de mogelijke uitkomsten van  $X$ ), de waarden van de  $y_j$ 's (mogelijke uitkomsten van  $Y$ ) en de waarden van de  $p_{ij}$ 's. We hebben gezien dat je uit de gezamenlijke kansverdeling van  $X$  en  $Y$  de enkelvoudige kansverdeling van  $X$  en  $Y$  afzonderlijk kunt afleiden door de randtotalen in de tabel te berekenen. We gebruiken daarvoor de volgende notatie: met  $p_{i\bullet} = \sum_j p_{ij}$  en  $p_{\bullet j} = \sum_i p_{ij}$  geven we de randtotalen aan de rechterkant van respectievelijk onder de tabel weer. Ga na dat geldt:

$$E(X) = \sum_i p_{i\bullet} x_i \text{ en } E(Y) = \sum_j p_{\bullet j} y_j.$$

Om  $E(X + Y)$  te berekenen stellen we eerst vast dat de waarde van  $X + Y$  in het vakje op de  $i^e$  rij en de  $j^e$  kolom van de tabel gelijk is aan  $x_i + y_j$ ; de bijbehorende kans is  $p_{ij}$ .  $E(X + Y)$  is nu gelijk aan de optelsom van deze mogelijke uitkomsten vermenigvuldigd met de bijbehorende kans. M.a.w.:  $E(X + Y) = \sum_{i,j} p_{ij}(x_i + y_j)$ .

Je kunt je dit als volgt voorstellen: we hebben alle hokjes van de tabel van de gezamenlijke kansverdeling gevuld met de producten  $p_{ij}(x_i + y_j)$ ;  $E(X + Y)$  is de optelsom van alle hokjes.

Vervolgens splitsen we deze tabel in twee tabellen. In de eerste vullen we de

## Kansverdelingen

hokjes met de producten  $p_{ij}x_i$ , in de tweede met  $p_{ij}y_j$ .  $E(X + Y)$  is nu de optelsom van alle hokjes van beide tabellen. In formule:

$$E(X + Y) = \sum_{i,j} p_{ij}(x_i + y_j) = \sum_{i,j} p_{ij}x_i + \sum_{i,j} p_{ij}y_j.$$

In de eerste tabel berekenen we de randtotalen aan de rechterkant van de tabel. De optelsom van de  $i^e$  rij is gelijk aan  $\sum_j p_{ij}x_i = x_i \sum_j p_{ij} = x_i p_{i\bullet}$ . We

kunnen nu de optelsom van alle hokjes in deze tabel berekenen door eerst de randtotalen aan de rechterkant van de tabel te berekenen en vervolgens deze randtotalen op te tellen. De uitkomst is dan gelijk aan  $E(X)$ . In formule:  $\sum_{i,j} p_{ij}x_i = \sum_i x_i p_{i\bullet} = E(X)$ .

In tweede tabel gaan we op soortgelijke wijze te werk, maar we berekenen nu eerst de randtotalen onder aan de tabel. We krijgen dan:  $\sum_{i,j} p_{ij}y_j = \sum_j y_j p_{\bullet j} = E(Y)$ . Hiermee is het bewijs geleverd. Het valt in formules in één regel samen te vatten:

$$\begin{aligned} E(X + Y) &= \sum_{i,j} p_{ij}(x_i + y_j) = \sum_{i,j} p_{ij}x_i + \sum_{i,j} p_{ij}y_j \\ &= \sum_i p_{i\bullet}x_i + \sum_j p_{\bullet j}y_j = E(X) + E(Y). \end{aligned}$$

Als  $X$  en  $Y$  continue stochasten zijn kun je de eigenschap  $E(X + Y) = E(X) + E(Y)$  op een soortgelijke manier bewijzen. Je hebt daar de samengestelde kansverdeling van  $X$  en  $Y$  voor nodig en je gebruikt dubbele integralen. We zullen dit bewijs hier niet opschrijven, maar het principe ervan lijkt sterk op het discrete geval.

In opgave 9 heb je het vermoeden gekregen dat voor *onafhankelijke*  $X$  en  $Y$  geldt  $E(XY) = E(X) \cdot E(Y)$ . Ook deze stelling gaan we bewijzen. (Je hebt ook gezien dat deze eigenschap niet hoeft te gelden als  $X$  en  $Y$  afhankelijk zijn.)

Er geldt  $E(XY) = \sum_{i,j} p_{ij}x_i y_j$ . Immers, in de tabel van de gezamenlijke kansverdeling van  $X$  en  $Y$  is de waarde van  $XY$  in het hokje op de  $i^e$  rij en de  $j^e$  kolom gelijk aan  $x_i y_j$ ; de bijbehorende kans is  $p_{ij}$ .

Verder geldt  $E(X) \cdot E(Y) = (\sum_i p_{i\bullet}x_i)(\sum_j p_{\bullet j}y_j)$ . Als we in deze laatste uitdrukking de sommen uitschrijven en de haakjes wegwerken krijgen we:

$$(\sum_i p_{i\bullet}x_i)(\sum_j p_{\bullet j}y_j) = \sum_{i,j} p_{i\bullet}p_{\bullet j}x_i y_j.$$

Als  $X$  en  $Y$  onafhankelijk zijn geldt  $p_{ij} = p_{i\bullet}p_{\bullet j}$  en dus kunnen we in dat geval concluderen:  $E(X) \cdot E(Y) = \sum_{i,j} p_{ij}x_i y_j$ , hetgeen tevens de formule is

### 1.3. Kansvariabelen en rekenregels

voor  $E(XY)$ .

Als  $X$  en  $Y$  continu zijn verloopt het bewijs analoog.

#### Opgaven

10.  $X$  is een discrete kansvariabele en  $a$  is een constante. Bewijs de volgende gelijkheden:  
 $E(X + a) = E(X) + a$ ,  $E(aX) = aE(X)$  en  $E(E(X)) = E(X)$ .
11. De resultaten van opgave 10 gelden ook voor continue kansvariabelen. Toon nu aan dat voor elke kansvariabele  $X$  (discreet of continu) geldt:  
 $Var(X) = E(X^2) - E^2(X)$ . (Zie opmerking 1.2.2 voor de betekenis van  $E^2(X)$ .)
12.  $X$  is een kansvariabele en  $a$  is een constante. Bewijs:  $Var(aX) = a^2 Var(X)$ .
13.  $X$  en  $Y$  zijn *onafhankelijke* kansvariabelen. Er geldt dus  $E(XY) = E(X) \cdot E(Y)$ . Gebruik deze gelijkheid en het resultaat van opgave 11 om aan te tonen dat:

$$Var(X + Y) = Var(X) + Var(Y).$$

De resultaten van opgaven 10 en 12 kun je ook als volgt samenvatten: als  $X$  een kansvariabele is en  $a$  en  $b$  constanten, dan geldt :

$$E(aX + b) = aE(X) + b \text{ en } Var(aX + b) = a^2 Var(X).$$

Als bovendien  $X$  normaal verdeeld is, dan is  $aX + b$  ook normaal verdeeld. Dit laatste bewijzen we hier niet. Deze eigenschap maakt het mogelijk om de kansverdeling van een willekeurige normale verdeling af te leiden van de standaardnormale verdeling.

**Voorbeeld 1.3.1.**  $X$  is normaal verdeeld met  $\mu = 20$  en  $\sigma = 10$ . Bereken  $\Pr(X \geq 28)$ . Bereken ook de waarde van  $x$  waarvoor geldt  $\Pr(X \leq x) = 0,15$ . Maak bij je berekeningen alleen gebruik van de standaardnormale verdeling.

*Uitwerking*

Voor  $Z = \frac{X - \mu}{\sigma}$  geldt  $E(Z) = \frac{1}{\sigma} E(X - \mu) = \frac{1}{\sigma} (E(X) - \mu) = 0$  en  $Var(Z) = \frac{1}{\sigma^2} Var(X - \mu) = \frac{\sigma^2}{\sigma^2} = 1$ .

$Z$  is dus standaardnormaal verdeeld.

$\Pr(X \geq 28) = \Pr\left(\frac{X - 20}{10} \geq \frac{28 - 20}{10}\right) = \Pr(Z \geq 0,8) \approx \text{normalcdf}(0,8, 10^99) = 0,211$

$\Pr(X \leq x) = \Pr\left(\frac{X - 20}{10} \leq \frac{x - 20}{10}\right) = \Pr\left(Z \leq \frac{x - 20}{10}\right)$ . Deze kans is gelijk aan 0,15

## Kansverdelingen

als  $\frac{x-20}{10} = \text{invNorm}(0,15) = -1,036$ . Hieruit volgt  $x = 20 - 1,036 \cdot 10 \approx 9,64$ . (Als je bij de GR geen waarden opgeeft voor  $\mu$  en  $\sigma$ , dan worden de “default values” 0 en 1 gebruikt en reken je dus met de standaardnormale verdeling.)

De hierboven en in de opgaven afgeleide rekenregels kun je ook goed toepassen bij steekproeven uit een populatie. Bij een goed genomen steekproef van lengte  $n$  heb je namelijk te maken met  $n$  kansvariabelen  $X_1, \dots, X_n$  die ieder dezelfde verdeling hebben en onderling onafhankelijk zijn. Noem de verwachtingswaarde van de gemeenschappelijke verdeling  $\mu = E(X)$  en de variantie  $\sigma^2 = \text{Var}(X)$  en definieer  $X_{Som} = \sum_{i=1}^n X_i$  en

$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} X_{Som}$ . Dan geldt:

$$E(X_{Som}) = E(X_1) + \dots + E(X_n) = n\mu \text{ en } E(\bar{X}) = \frac{1}{n} E(X_{Som}) = \mu.$$

Merk op dat deze gelijkheden ook gelden als de  $X_i$ 's niet onafhankelijk zijn. De onafhankelijkheid is wel een noodzakelijke voorwaarde voor:

$$\text{Var}(X_{Som}) = \text{Var}(X_1) + \dots + \text{Var}(X_n) = n\sigma^2 \text{ en } \text{Var}(\bar{X}) = \frac{1}{n} \text{Var}(X_{Som}) = \frac{1}{n} \sigma^2.$$

Hieruit volgt de bekende  $\sqrt{n}$ -wet:

$$\sigma(X_{Som}) = \sqrt{\text{Var}(X_{Som})} = \sigma\sqrt{n} \text{ en } \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Bovenstaande eigenschappen zijn geldig ongeacht de verdeling van de  $X_i$ 's. Als verder gegeven is dat de  $X_i$ 's *normaal verdeeld* zijn, dan geldt dat  $X_{Som}$  en  $\bar{X}$  ook normaal verdeeld zijn. Dit bewijzen we hier niet.

Samengevat zijn we in deze paragraaf de volgende rekenregels tegengekomen:

$X$  en  $Y$  zijn kansvariabelen,  $a$  en  $b$  zijn constanten. Dan geldt:

$$E(aX + b) = aE(X) + b \text{ en } \text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$\text{Var}(X) = E(X^2) - E^2(X)$$

$$E(X + Y) = E(X) + E(Y)$$

Als  $X$  en  $Y$  bovendien onafhankelijk zijn, geldt:

$$E(XY) = E(X) \cdot E(Y) \text{ en } \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$X_1, \dots, X_n$  zijn onderling onafhankelijke kansvariabelen met een gemeenschappelijke verwachtingswaarde  $\mu$  en een gemeenschappelijke variantie  $\sigma^2$ . Dan geldt:

$$E(X_{Som}) = n\mu, \quad \text{Var}(X_{Som}) = n\sigma^2 \text{ en } \sigma(X_{Som}) = \sigma\sqrt{n}$$

$$E(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \sigma^2/n \text{ en } \sigma(\bar{X}) = \sigma/\sqrt{n}$$

### 1.4. De centrale limietstelling

#### Opgaven

14.  $X$  en  $Y$  zijn onafhankelijke kansvariabelen met  $E(X) = 5$ ,  $E(Y) = 6$ ,  $Var(X) = 2$  en  $Var(Y) = 3$ . Bereken:
- $E(2X - 3)(Y - 1)$
  - $E(X^2 - 5Y)$
  - $Var(2X - 3)$
  - $Var(3X - Y)$
15. Je mag in deze opgave alleen de kansverdeling van de standaardnormale verdeling gebruiken.  $X$  is normaal verdeeld met verwachting  $\mu$  en standaardafwijking  $\sigma$ .
- $\mu = 2,4$  en  $\sigma = 0,7$ . Bereken  $\Pr(2 \leq X \leq 3)$
  - $\mu = 2,4$  en  $\sigma = 0,7$ . Voor welke waarde van  $x$  geldt  $\Pr(X \geq x) = 0,01$ ?
  - $\sigma = 13$  en  $\Pr(X \geq 50) = 0,07$ . Bereken  $\mu$ .
  - $\mu = 40$  en  $\Pr(X \leq 30) = 0,25$ . Bereken  $\sigma$ .
16. Het gewicht van een willekeurige appel uit een partij appels is normaal verdeeld met gemiddelde 80 gram en een standaarddeviatie van 15 gram. Je neemt een steekproef van 20 appels uit deze partij en je berekent er het gemiddelde gewicht van. Hoe groot is de kans dat dit gemiddelde kleiner is dan 75 gram?
17. Jan moet een stapel van 30 pasfoto's beoordelen. Hij geeft aan iedere foto een score. De score is 1 voor een lelijk, 2 voor een neutraal en 3 voor een knap gezicht. Jan heeft dit werk al eerder gedaan en we kennen de kansverdeling van de scores die Jan geeft: de kans op een 1 is 0,15, de kans op een 2 is 0,55 en de kans op een 3 is 0,3. Bereken het gemiddelde en de standaardafwijking van de gemiddelde score die Jan geeft aan deze 30 foto's.

## 1.4 De centrale limietstelling

In de vorige paragraaf hebben de zogenaamde  $\sqrt{n}$ -wet in herinnering geroepen. Als  $X_1, \dots, X_n$  onderling onafhankelijk en gelijk verdeeld zijn met verwachting  $\mu$  en variantie  $\sigma^2$ , dan geldt voor  $X_{Som} = \sum_{i=1}^n X_i$ :

$$E(X_{Som}) = n\mu \text{ en } Var(X_{Som}) = n\sigma^2.$$

We kunnen dit ook als volgt formuleren:

$$Z = \frac{X_{Som} - n\mu}{\sigma\sqrt{n}} \text{ heeft verwachting 0 en variantie 1.}$$

## Kansverdelingen

Verder hebben we gezien dat, als bovendien de  $X_i$ 's normaal verdeeld zijn,  $X_{Som}$ , en daarmee  $Z$ , eveneens normaal verdeeld zijn. In dat geval zeggen we dat  $Z$  standaardnormaal verdeeld is.

Als de  $X_i$ 's niet normaal verdeeld zijn, is de verdeling van  $Z$  meestal moeilijk te bepalen. De centrale limietstelling zegt nu dat als  $n$  naar oneindig gaat, de verdeling van  $Z$  steeds meer op de standaardnormale verdeling gaat lijken, ook al zijn de  $X_i$ 's zelf niet normaal verdeeld. In de praktijk wordt deze stelling vaak zo toegepast, dat als  $n$  groot genoeg is, de verdeling van  $Z$  bij benadering standaardnormaal is. Je kunt dan ook de verdeling van  $X_{Som}$  benaderen met een normale verdeling met verwachting  $n\mu$  en standaardafwijking  $\sigma\sqrt{n}$ . De vraag is natuurlijk: wanneer is  $n$  "groot genoeg"? Het antwoord op deze vraag is sterk afhankelijk van de verdeling van de  $X_i$ 's.

**Voorbeeld 1.4.1.** Thea en Petrien schieten op de kermis allebei twintig keer met een luchtbuks. De kans dat Thea raak schiet is 0,4. De kans dat Petrien raak schiet is 0,3. Hoe groot is de kans dat ze samen op 17 treffers of meer uitkomen?

*Uitwerking*

Noem het aantal treffers van Thea  $X$ .  $X$  is binomiaal verdeeld met  $n = 20$  en  $\pi_1 = 0,4$ .

Er geldt:  $E(X) = n\pi_1 = 8$  en  $Var(X) = n\pi_1(1 - \pi_1) = 4,8$ .

Noem het aantal treffers van Petrien  $Y$ .  $Y$  is binomiaal verdeeld met  $n = 20$  en  $\pi_2 = 0,3$ . Er geldt  $E(Y) = n\pi_2 = 6$  en  $Var(Y) = n\pi_2(1 - \pi_2) = 4,2$ .

Het aantal treffers van Thea en Petrien samen noemen we  $T (= X + Y)$ . We zoeken  $\Pr(T \geq 17)$ .

We kennen de verdelingen van  $X$  en  $Y$ , maar de verdeling van  $T$  is moeilijk te bepalen. De som van twee normaal verdeelde kansvariabelen is echter wel eenvoudig te bepalen, daarom benaderen we de verdelingen van  $X$  en  $Y$  met behulp van de normale verdeling. De binomiale verdeling ontstaat namelijk als de som van  $n$  onafhankelijke, identiek verdeelde Bernoulli kansvariabelen die de waarde 1 aannemen (succes) of 0 (mislukking). De optelsom van een voldoende groot aantal van dergelijke kansvariabelen zal dus bij benadering normaal verdeeld zijn. (Er is een vuistregel die zegt dat bij de binomiale verdeling deze benadering acceptabel is als  $n\pi$  en  $n(1 - \pi)$  beide groter of gelijk zijn aan 5. Dit is het geval bij de huidige kansvariabelen  $X$  en  $Y$ .)

Als  $X$  en  $Y$  (bij benadering) normaal verdeeld zijn, is  $T = X + Y$  eveneens (bij benadering) normaal verdeeld met verwachting  $8+6 = 14$  en standaardafwijking  $\sqrt{4,8+4,2} = 3$ . Rekening houdend met de continuïteitscorrectie vinden we:

$$\Pr(T \geq 17) = \Pr(T \geq 16,5) \approx \text{normalcdf}(16.5, 10^{99}, 14, 3) = 0,2$$

## 1.5. Van de normale verdeling afgeleide verdelingen

### Opgave

18. Zie opgave 17. Bereken de kans dat de gemiddelde score 2 is of minder.

## 1.5 Van de normale verdeling afgeleide verdelingen

### Opgave

19. Gegeven zijn drie onderling onafhankelijke, discrete kansvariabelen  $X_1$ ,  $X_2$  en  $X_3$ . Deze stochasten hebben dezelfde kansverdeling, weergegeven in onderstaande tabel.
- |                |     |     |     |
|----------------|-----|-----|-----|
| $x$            | -1  | 0   | 2   |
| $\Pr(X_i = x)$ | 0,2 | 0,5 | 0,3 |

We definiëren nu een nieuwe kansvariabele:  $Y = X_1^2 + X_2^2 + X_3^2$ .

Bepaal de kansverdeling van  $Y$ . (Aanwijzing: bepaal eerst de kansverdeling van  $X_i^2$ , vervolgens die van  $X_1^2 + X_2^2$  en tenslotte die van  $Y$ .)

In opgave 19 heb je gezien dat het mogelijk is om de kansverdeling te bepalen van een (redelijk ingewikkelde) functie van andere gegeven, discrete, kansvariabelen. Als we nu de gegeven discrete kansverdeling van de  $X_i$ 's vervangen door de standaardnormale verdeling, wat wordt dan de verdeling van  $Y$ ? Deze vraag is alleen op te lossen met behulp van een flinke dosis integraalrekening. Daar zullen we ons in deze module niet aan wagen, maar we bespreken wel het resultaat, want we zullen de resulterende verdeling van  $Y$  in de volgende hoofdstukken vaak tegenkomen.

### 1.5.1 De $\chi^2$ -verdeling (chi-kwadraatverdeling)

Laat  $X_1, X_2, \dots, X_n$  onderling onafhankelijke kansvariabelen zijn met een standaardnormale verdeling. We definiëren nu een nieuwe kansvariabele

$$Y = \sum_{i=1}^n X_i^2.$$

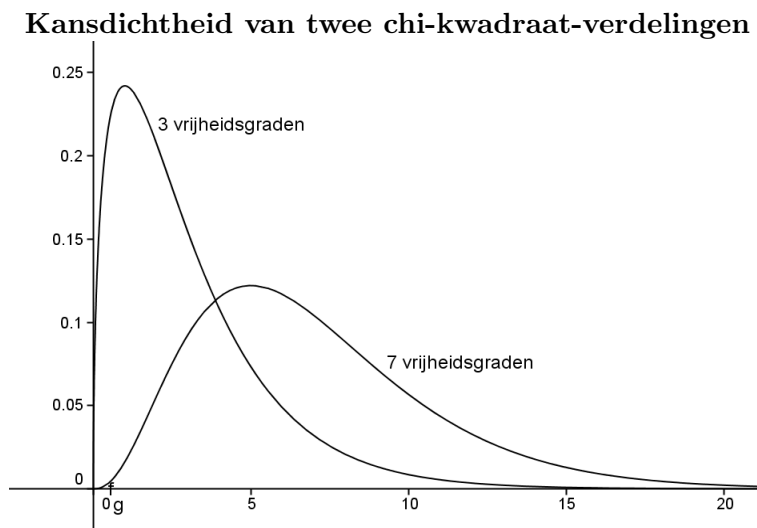
Met behulp van integraalrekening kunnen we de verdeling van  $Y$  bepalen. De verdeling noemen we de  $\chi^2$ -verdeling met  $n$  vrijheidsgraden (*degrees of freedom*), ook wel aangeduid met  $\chi_{[n]}^2$ .

Er geldt:

$$E(Y) = n \text{ en } Var(Y) = 2n.$$

## Kansverdelingen

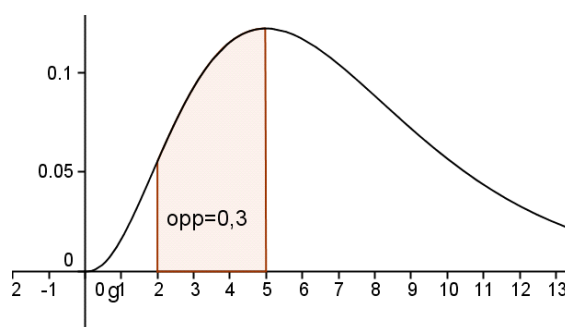
Hieronder zie je een plaatje van de kansdichtheid van een  $\chi^2$ -verdeling met 3 respectievelijk 7 vrijheidsgraden.



Op de GR kun je de *kansdichtheid* van de  $\chi^2$ -verdeling vinden onder “Distr”:  $\chi^2$ pdf. Als argument voer je eerst de  $x$ -waarde in en vervolgens, na een komma, het aantal vrijheidsgraden.

De *cumulatieve kansverdeling* van de  $\chi^2$ -verdeling vind je onder “Distr”:  $\chi^2$ cdf. Het argument is (*linkergrens, rechtergrens, aantal vrijheidsgraden*). Als  $Y$  een chi-kwadraatverdeling heeft met 7 vrijheidsgraden, dan bereken je bijvoorbeeld:

$$\Pr(2 \leq Y \leq 5) \approx \chi^2\text{cdf}(2, 5, 7) = 0,300.$$



## Opgaven

20. Plot de kansdichtheid van de  $\chi^2$ -verdeling met achtereenvolgens 2, 5, 20 en 50 vrijheidsgraden. Let daarbij op het window dat nodig is om de kromme goed in beeld te krijgen en de eventuele symmetrie. Had je deze resultaten kunnen voorspellen zonder te plotten?



### 1.5. Van de normale verdeling afgeleide verdelingen

21.  $X$  heeft een  $\chi^2$ -verdeling met 8 vrijheidsgraden. Bereken  $\Pr(X \leq 10)$ .
22.  $X$  heeft een  $\chi^2$ -verdeling met 15 vrijheidsgraden. Voor welke waarde van  $x$  geldt  $\Pr(X \geq x) = 0,05$ ? (Gebruik de functie Intersect.)
23. Bereken de kans dat  $Y$  minder dan één standaardafwijking afwijkt van het gemiddelde voor het geval dat  $Y$  een  $\chi^2$ -verdeling heeft met 2 respectievelijk 50 vrijheidsgraden en voor het geval  $Y$  (standaard)normaal verdeeld is. Wat valt op? Geef een verklaring.
24.  $Y$  heeft een chi-kwadraatverdeling met  $n$  vrijheidsgraden. Toon aan dat  $E(Y) = n$ .

#### 1.5.2 De t-verdeling (student-t-verdeling)

Laat  $X$  standaard-normaal verdeeld zijn en  $Y$  een  $\chi^2_{[n]}$  verdeling hebben, waarbij  $X$  en  $Y$  onafhankelijk zijn. Definieer

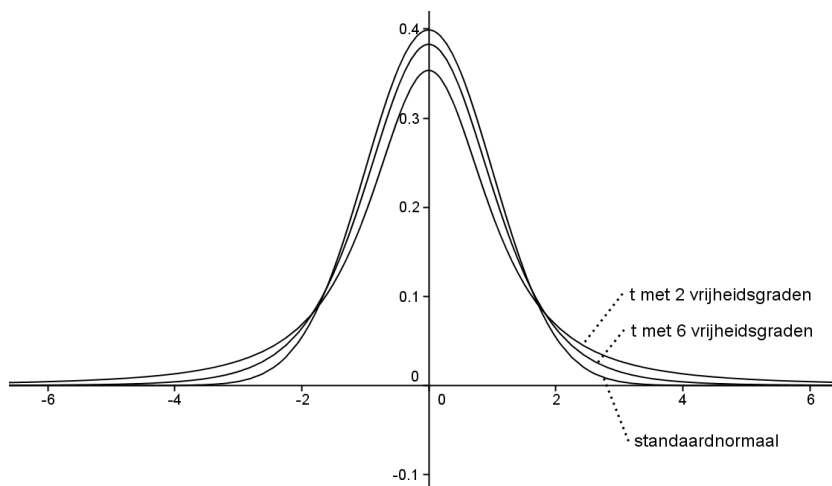
$$T = \frac{X}{\sqrt{\frac{1}{n}Y}}.$$

Dan heeft  $T$  een *t-verdeling met  $n$  vrijheidsgraden*, ook wel aangeduid met  $t_{[n]}$ .

Er geldt:

$$E(T) = 0 \text{ (als } n > 1) \text{ en } Var(T) = \frac{n}{n-2} \text{ (als } n > 2).$$

De kansdichtheid van de t-verdeling lijkt sterk op die van de standaardnormale verdeling. De staarten van de verdeling zijn echter dikker, vooral als het aantal vrijheidsgraden klein is. In de figuur hieronder zie je grafieken van de kansdichtheid van de standaardnormale verdeling en van twee t-verdelingen met 2 en 6 vrijheidsgraden.



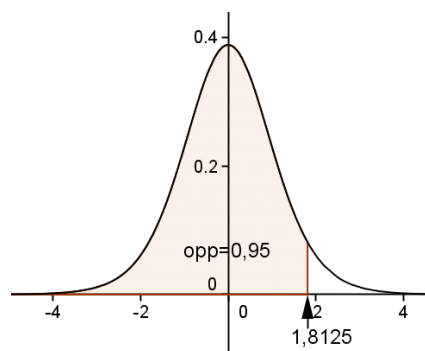
## Kansverdelingen

De *kansdichtheid* van de t-verdeling vind je op de GR onder “Distr”: `tpdf`. Als argument voer je eerst de  $x$ -waarde in en vervolgens, na een komma, het aantal vrijheidsgraden.

De *cumulatieve kansverdeling* van de t-verdeling vind je onder “Distr”: `tcdf`.

Het argument is (*linkergrens*, *rechtergrens*, *aantal vrijheidsgraden*).

Je kunt de functie `tcdf` gebruiken om kansen uit te rekenen, net zoals je dat met `normalcdf` en  `$\chi^2$ cdf` doet. Bovendien kent de GR ook een functie `invT` die de rechtergrens bepaalt van een gebied met een vooraf opgegeven oppervlakte. Bijvoorbeeld: als  $T$  een t-verdeling heeft met 10 vrijheidsgraden en je zoekt de waarde van  $t$  waarvoor geldt  $\Pr(T \leq t) = 0,95$ , dan kun je  $t$  berekenen via `invT(0.95, 10) = 1,8125`.



## Opgaven

25.  $T$  heeft een t-verdeling met 7 vrijheidsgraden. Bereken  $\Pr(T \leq 2)$ .
26.  $T$  heeft een t-verdeling met 4 vrijheidsgraden. Voor welke waarde van  $x$  geldt  $\Pr(-x \leq T \leq x) = 0,95$ ?
27. Een t-verdeling heeft “dikkere staarten” dan de standaardnormale verdeling, vooral als het aantal vrijheidsgraden klein is. Illustreer deze eigenschap door de volgende tabel in te vullen:

	$\Pr(X \geq 1)$	$\Pr(X \geq 2)$	$\Pr(X \geq 3)$
$X$ heeft t-verdeling met 2 vrijheidsgraden			
$X$ heeft t-verdeling met 5 vrijheidsgraden			
$X$ heeft t-verdeling met 10 vrijheidsgraden			
$X$ heeft t-verdeling met 20 vrijheidsgraden			
$X$ is standaardnormaal verdeeld			

## Hoofdstuk 2

# Schattingen van parameters en betrouwbaarheidsintervallen

Stel dat we willen weten hoe lang Leidse jongens van 17 jaar zijn. Daartoe nemen we een aselechte steekproef van 100 van dergelijke jongens en we meten hun lengte. We willen de gegevens van deze steekproef gebruiken om uitspraken te doen over de gehele populatie.

Of: we vragen aan deze 100 jongens of ze meer dan 10 sigaretten per week roken en we willen aan de hand van de antwoorden uitspraken doen over het aantal rokers onder *alle* Leidse 17-jarige jongens.

Bovenstaande twee vragen zijn kenmerkend voor grote delen van de statistiek. We willen iets te weten komen over een grote populatie (in ons voorbeeld alle 17-jarige Leidse jongens) en dat doen we aan de hand van de gegevens uit een aselechte steekproef uit die populatie. De populatie als geheel wordt beschreven door een kansverdeling en deze kansverdeling wordt op haar beurt gekenmerkt door een aantal parameters: het gemiddelde, de variantie etc. Deze parameters liggen vast, maar zijn onbekend. We gaan proberen om deze parameters te schatten met behulp van de steekproef. Zo kun je de waarde van het gemiddelde van de gehele populatie schatten met behulp van het gemiddelde van de steekproef.

In ons voorbeeld: als de gemiddelde lengte van de 100 jongens uit onze steekproef 1,84 m is, dan kunnen we deze 1,84 m (het steekproefgemiddelde) gebruiken als *schatting* van de gemiddelde lengte van *alle* 17-jarige Leidse jongens (het populatiegemiddelde). Het is niet meer dan een schatting, want het werkelijke populatiegemiddelde kennen we niet.

Het is in dit hoofdstuk van belang om onderscheid te maken tussen deze twee grootheden. Het *populatiegemiddelde* (Engels: *population mean*) is een vast getal, een parameter; het *steekproefgemiddelde* (Engels: *sample*

*mean*) is stochastisch d.w.z. een kansvariabele. Als je een nieuwe steekproef neemt, krijg je een ander steekproefgemiddelde.

## 2.1 Normaal verdeelde populaties

We nemen een steekproef  $X_1, X_2, \dots, X_n$  uit een populatie die normaal verdeeld is met gemiddelde  $\mu$  en standaardafwijking  $\sigma$ .  $\mu$  en  $\sigma$  hebben betrekking op de gehele populatie. Het zijn vaste getallen.

Het steekproefgemiddelde is  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . Dit steekproefgemiddelde is een kansvariabele, want het hangt af van de steekproef. We weten dat  $\bar{X}$  normaal verdeeld is met parameters  $\mu$  en  $\sigma/\sqrt{n}$ . De variabele  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  is standaardnormaal verdeeld. Dit maakt het mogelijk om de volgende kansuitspraak te doen:

$$\Pr(-1,96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1,96) = 0,95$$

Of, als we de ongelijkheden tussen haakjes anders opschrijven,

$$\Pr(\bar{X} - \frac{1,96\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{1,96\sigma}{\sqrt{n}}) = 0,95. \quad (*)$$

We concretiseren het bovenstaande aan de hand van een voorbeeld. De kansvariabelen  $X_1, X_2, \dots, X_n$  stellen de lengtes voor van 100 aselekt gekozen Leidse 17-jarige jongens en het gemiddelde van de steekproef is  $\bar{X} = 184,0\text{cm}$ .

Veronderstel nu dat  $\sigma$  bekend is:  $\sigma = 9\text{cm}$ .

Dan kunnen we (\*) herschrijven:  $\Pr(184 - \frac{1,96 \cdot 9}{\sqrt{100}} \leq \mu \leq 184 + \frac{1,96 \cdot 9}{\sqrt{100}}) = 0,95$  ofwel:  $\Pr(182,2 \leq \mu \leq 185,8) = 0,95$ .

We zeggen ook wel dat  $[182,2 ; 185,8]$  een *95%-betrouwbaarheidsinterval* voor  $\mu$  is. (Engels: *confidence interval*). De grenzen van het betrouwbaarheidsinterval worden gegeven door (\*).

**Opmerking 2.1.1.** In bovenstaande situatie zeggen we vaak “ $\mu$  ligt met een kans van 95% tussen 182,2 cm en 185,8 cm”. Deze uitspraak suggereert dat  $\mu$  een kansvariabele is en dat is natuurlijk niet waar, want  $\mu$  is een vast getal. De grenzen van het betrouwbaarheidsinterval, zoals gegeven door (\*) zijn kansvariabelen, want die hangen af van de steekproef. Als je een nieuwe steekproef neemt, krijg je andere grenzen, maar  $\mu$  verandert niet. Als je heel veel nieuwe steekproeven (van gelijke lengte) neemt, krijg je heel veel intervallen. Je kunt dan zeggen dat  $\mu$  in 95% van die intervallen ligt.

## 2.1. Normaal verdeelde populaties

### Opgaven

28. Verklaar bovenstaande afleiding. Laat zien dat het getal 1,96 te maken heeft met de kans van 0,95. Welk getal moet je gebruiken als je een 99%-betrouwbaarheidsinterval wilt hebben?
29. Stel dat de standaardafwijking van de normaal verdeelde lengte van Leidse 17-jarige jongens gelijk is aan 9 cm. Een steekproef van 50 van deze jongens geeft een gemiddelde lengte van 184,0 cm. Geef een 90%- betrouwbaarheidsinterval van de werkelijke gemiddelde lengte van Leidse 17-jarige jongens.

**Opmerking 2.1.2.** Het is verleidelijk om aan de steekproeven van 17-jarige Leidse jongens conclusies te verbinden over de lengtes van alle Nederlandse 17-jarige jongens. Dit zou echter niet correct zijn. Om iets te kunnen zeggen over de populatie van Nederlandse 17-jarige jongens heb je een aselechte steekproef uit die populatie nodig en niet uit een deelpopulatie.

In de praktijk zal het zelden zo zijn dat  $\sigma$  bekend is. We moeten dan de gegevens uit de steekproef gebruiken om  $\sigma$  (of de variantie  $\sigma^2$ ) te schatten.

Veronderstel eerst (hoewel ook dit in de praktijk niet vaak zal voorkomen) dat  $\mu$  bekend is. Het ligt dan voor de hand om  $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$  te gebruiken als schatting van  $\sigma^2$ .

Dit valt ook theoretisch te rechtvaardigen. Voor elke  $i$  geldt  $E(X_i - \mu)^2 = \sigma^2$ , dus  $E(S^2) = \frac{1}{n} \cdot n\sigma^2 = \sigma^2$ . De verwachtingswaarde van de schatter  $S^2$  is dus gelijk aan de parameter die geschat wordt. We zeggen ook wel dat  $S^2$  een *zuivere schatter* is (Engels: *unbiased estimator*). Deze eigenschap van  $S^2$  is niet afhankelijk van de verdeling van de  $X_i$ 's.

Als (realistischer)  $\mu$  onbekend is, is het verleidelijk om in de formule voor  $S^2$ ,  $\mu$  te vervangen door zijn schatter  $\bar{X}$ . We zouden dan  $S^2$  vervangen door  $\frac{1}{n} \sum (X_i - \bar{X})^2$ .

Dit is echter niet onproblematisch, zoals duidelijk moge worden uit de volgende berekening.

$$\begin{aligned} \sum (X_i - \mu)^2 &= \sum ((X_i - \bar{X}) + (\bar{X} - \mu))^2 = \\ &= \sum (X_i - \bar{X})^2 + \sum 2(X_i - \bar{X})(\bar{X} - \mu) + \sum (\bar{X} - \mu)^2 = \\ &= \sum (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \sum (X_i - \bar{X}) + n(\bar{X} - \mu)^2 = \quad (**) \\ &= \sum (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \end{aligned}$$

Hieruit kun je zien dat voor elke steekproef zal gelden dat

$$\frac{1}{n} \sum (X_i - \bar{X})^2 \leq S^2.$$

### Schattingen van parameters en betrouwbaarheidsintervallen

Het is dus verstandig om een schatter voor  $\sigma^2$  te gebruiken die wat groter is dan  $\frac{1}{n} \sum (X_i - \bar{X})^2$ . Dit zullen we nader preciseren.

Uit (\*\*) volgt:

$$\begin{aligned} E(\sum (X_i - \bar{X})^2) &= E(\sum (X_i - \mu)^2) - E(n(\bar{X} - \mu)^2) \\ &= \sum E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2 \\ &= \sum \text{Var}(X_i) - n\text{Var}(\bar{X}) \\ &= n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2. \end{aligned}$$

Of, als we beide zijden delen door  $n-1$ :

$$E\left(\frac{1}{n-1} \sum (X_i - \bar{X})^2\right) = \sigma^2$$

Hieraan zien we dat  $\frac{1}{n-1} \sum (X_i - \bar{X})^2$  een zuivere schatter is van  $\sigma^2$ . Deze schatter wordt aangeduid met  $s^2$ . Dat  $s^2$  een zuivere schatter is, is niet afhankelijk van de verdeling van de  $X_i$ 's.

Als de  $X_i$ 's bovendien normaal verdeeld zijn, kun je met integraalrekening bewijzen dat  $\frac{(n-1)s^2}{\sigma^2}$  een  $\chi^2_{[n-1]}$ -verdeling heeft.

$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  is een zuivere schatter van  $\sigma^2$  en heet de *steekproefvariantie* (Engels: *sample variance*).

Als de  $X_i$ 's bovendien normaal verdeeld zijn, geldt:  $\frac{(n-1)s^2}{\sigma^2}$  heeft een  $\chi^2$ -verdeling met  $n-1$  vrijheidsgraden.

Het voordeel van  $s^2$  boven  $S^2$  is dat je  $s^2$  kunt berekenen op basis van louter de steekproef. (Om  $S^2$  te berekenen moet je de waarde van  $\mu$  kennen en dat zal in de praktijk niet vaak voorkomen.)

Aangezien  $\text{Var}(\bar{X}) = \sigma^2/n$  wordt  $s/\sqrt{n}$  wel de *standaardfout van het gemiddelde* genoemd (Engels: *standard error of the mean, SEM*). Deze *SEM* is een maat voor de nauwkeurigheid waarmee het steekproefgemiddelde het populatiegemiddelde benadert.

Met integraalrekening kun je verder bewijzen dat  $s^2$  onafhankelijk is van  $\bar{X}$  en daarom geldt (zie 1.5.2):

$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\bar{X} - \mu}{SEM}$  heeft een  $t$ -verdeling met  $n-1$  vrijheidsgraden.

Met behulp van deze eigenschap kunnen betrouwbaarheidsintervallen worden afgeleid voor  $\mu$ .

## 2.1. Normaal verdeelde populaties

Stel dat  $T$  een t-verdeling heeft met  $n-1$  vrijheidsgraden en noem  $t_{[n-1],0,05}$  het getal waarvoor geldt:  $\Pr(-t_{[n-1],0,05} \leq T \leq t_{[n-1],0,05}) = 0,95$ . Dit getal kun je berekenen met je GR (of aflezen uit een tabel). Dan geldt dus:

$$\Pr(-t_{[n-1],0,05} \leq \frac{\bar{X} - \mu}{SEM} \leq t_{[n-1],0,05}) = 0,95$$

of

$$\Pr(\bar{X} - t_{[n-1],0,05} \cdot SEM \leq \mu \leq \bar{X} + t_{[n-1],0,05} \cdot SEM) = 0,95.$$

We zeggen daarom:

$\bar{X} \pm t_{[n-1],0,05} \cdot SEM$  is een 95% betrouwbaarheidsinterval voor  $\mu$ .

**Opmerking 2.1.3.** Anders dan aan het begin van deze paragraaf, komt de parameter  $\sigma$  niet voor in de formule van het betrouwbaarheidsinterval hierboven. In plaats daarvan wordt de  $SEM$  gebruikt en deze kun je berekenen aan de hand van je steekproefgegevens. Dat maakt de formule goed bruikbaar in de praktijk.

**Opmerking 2.1.4.** Zoals gezegd is het belangrijk om onderscheid te maken tussen het populatiegemiddelde  $\mu$  en het steekproefgemiddelde  $\bar{X}$ . Op dezelfde manier is er onderscheid tussen de populatievariantie  $\sigma^2$  en de steekproefvariantie  $s^2$ .  $\sigma^2$  is een vast getal,  $s^2$  een kansvariabele.  $\sigma^2$  kun je niet waarnemen,  $s^2$  kun je berekenen aan de hand van de uitkomst van je steekproef en  $s^2$  kan dienen als schatter van  $\sigma^2$ .

$\bar{X}$  is een kansvariabele. De standaardafwijking van  $\bar{X}$  is  $\sigma/\sqrt{n}$ . Deze laatste grootte is weer een niet-waarneembaar vast getal. Dit getal kan worden geschat door  $SEM$ .  $SEM$  is een kansvariabele.

**Voorbeeld 2.1.1.** Het vetgehalte van hotdogs van het merk Tekkel wordt onderzocht. Van een steekproef van 10 willekeurig gekozen hotdogs wordt het vetgehalte (als percentage van het gewicht) geregistreerd: 25,2 21,3 22,8 17,0 29,8 21,0 25,5 16,0 20,9 19,5.

Bepaal een 95% betrouwbaarheidsinterval voor het werkelijke gemiddelde vetgehalte van de hotdogs van Tekkel. Je mag ervan uitgaan dat het vetgehalte normaal verdeeld is.

### *Uitwerking*

We berekenen eerst het gemiddelde vetgehalte van de steekproef:  $\bar{x} = 21,9$   
Vervolgens maken we de volgende tabel:

### Schattingen van parameters en betrouwbaarheidsintervallen

$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
25,2	3,3	10,89
21,3	-0,6	0,36
22,8	0,9	0,81
17,0	-4,9	24,01
29,8	7,9	62,41
21,0	-0,9	0,81
25,5	3,6	12,96
16,0	-5,9	34,81
20,9	-1,0	1,00
19,5	-2,4	5,76

We berekenen verder de steekproefvariantie als schatting van  $\sigma^2$ :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{9} \cdot 153,82 \approx 17,09, \text{ dus } s \approx \sqrt{17,09} \approx 4,13 \text{ en de}$$

standaardfout van het gemiddelde  $SEM \approx \sqrt{17,09/10} \approx 1,307$ .

Met de GR kun je berekenen dat  $t_{[9],0,05} \approx 2,262$ . (Gebruik `invT(0.975, 9)`)

Het 95% betrouwbaarheidsinterval voor het werkelijke gemiddelde vetgehalte  $\mu$  is dus:  $21,9 \pm 2,262 \cdot 1,307$  ofwel  $[18,9 ; 24,9]$ .

*N.B.* Het is belangrijk om de stappen uit bovenstaande berekening goed te doorgronden. Maar in de praktijk valt deze berekening eenvoudig met de GR uit te voeren.

- Voer eerst de waarnemingen uit je steekproef (de  $x_i$ 's) in de lijst L1 in (via STAT -EDIT).
- Via STAT - TESTS - T Interval kun je zelfs rechtstreeks het betrouwbaarheidsinterval berekenen:
- Bij Inpt: wordt gevraagd of je de berekening wilt uitvoeren met de oorspronkelijke gegevens (dan kies je de optie Data) of dat je alleen de kenmerken van je steekproef wilt invoeren (dan kies je de optie Stats). Wij hebben de oorspronkelijke gegevens al ingevoerd, dus we kiezen voor Data.
- Bij List voer je L1 in. Bij Freq voer je het getal 1 in, want elke waarneming komt 1 keer voor in je steekproef. (Als de waarnemingen meerdere malen voorkomen in je steekproef, dan kun je bij Freq een tweede lijst met frequenties invoeren.) Bij C-level voer je 0,95 in. Tenslotte voer je het commando Calculate uit.
- Als uitkomst vind je het betrouwbaarheidsinterval. Verder geeft de GR het steekproefgemiddelde ( $\bar{x}$ ), de steekproefstandaardafwijking ( $s$ ) en het aantal waarnemingen ( $n$ ).



## 2.1. Normaal verdeelde populaties

**Opmerking 2.1.5.** Het kan ook voorkomen dat je niet beschikt over alle steekproefgegevens, maar wel over een aantal samenvattende data. Stel dat je in voorbeeld 2.1.1 weet dat  $n = 10$ ,  $\bar{x} = 21,9$  en  $s = 4,13$ . Je kunt dan op de volgende manier de GR gebruiken om het gevraagde betrouwbaarheidsinterval te bepalen.

- Ga via STAT en TESTS naar T Interval en kies voor de optie Stats.
- Voer de gevraagde gegevens in (met Sx wordt  $s$  bedoeld) en voer het commando Calculate uit.
- De uitkomst is als in het voorbeeld.

**Opmerking 2.1.6.** Je kunt de GR ook gebruiken om zelf een aantal samenvattende data te berekenen op basis van een ingevoerde steekproef. Stel je hebt de gegevens uit Voorbeeld 2.1.1 ingevoerd in L1.

- Ga via STAT en CALC naar 1-Var Stats
- Voer in: 1-VarsStats L1 en druk op Enter. (Als je de frequenties bij de gegevens hebt ingevoerd in L2 voer je in 1-Var Stats L1,L2. In ons voorbeeld is dat niet nodig omdat elke waarneming maar één keer voorkomt.)
- Op het scherm zie je nu talloze samenvattende data van de ingevoerde gegevensverzameling. De meeste spreken voor zich.

Met Sx wordt  $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$  bedoeld en met  $\sigma x$

wordt  $\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$  bedoeld. In deze module zullen we meestal Sx gebruiken.

### Opgaven

30. Op de chocoladerepen van het merk Wonka Extra Puur staat: “Cacaogehalte minstens 70%!”. Eline en Merle willen onderzoeken of dit waar is en kopen, in acht willekeurige winkels, 8 repen Wonka Extra Puur waarvan ze het cacaogehalte laten bepalen. De laboratoriumuitslag is: 70,5 69,5 73,1 72,0 70,2 69,8 72,7 70,2 . Bepaal het 95%-betrouwbaarheidsinterval voor het gemiddelde cacaogehalte van repen Wonka Extra Puur. Je mag ervan uitgaan dat het cacaogehalte per reep normaal verdeeld is. Maak de berekening eerst zonder gebruik

### *Schattingen van parameters en betrouwbaarheidsintervallen*

te maken van één van de STAT functies op de GR en controleer je antwoord daarna door gebruik te maken van STAT-TESTS.

31. Jelle en Pieter doen ook een onderzoek naar het cacao gehalte van chocoladerepen. Ze nemen een steekproef van 30 repen van het merk Dark Willy. Uit hun gegevens volgt  $\bar{x} = 71,5$  en  $s = 3,1$ . Bepaal het 95%-betrouwbaarheidsinterval voor het gemiddelde cacao gehalte van repen Dark Willy.
32. Saskia en Boudewijn doen onderzoek naar de werkingssnelheid van Paracetamol bij brugklassers met hoofdpijn. Gedurende één maand krijgen alle brugklassers die wegens hoofdpijn een paracetamoltablet bij de conciërge komen halen de opdracht om zich te komen melden als de hoofdpijn voorbij is. Zo registreren ze van 48 brugklassers de werkingssnelheid (in minuten) van Paracetamol. Uit hun gegevens volgt  $\sum x_i = 629$  en  $\sum x_i^2 = 9517$ . Geef een schatting van het gemiddelde en de standaardafwijking van de werkingssnelheid van Paracetamol. Geef ook het 95%- betrouwbaarheidsinterval van het gemiddelde.
33. De lengte van het 95%-betrouwbaarheidsinterval uit Voorbeeld 2.1.1 is 6,0 (%punt). Is de lengte van het 90%- betrouwbaarheidsinterval groter of kleiner? En die van het 99%-betrouwbaarheidsinterval?
34. KLM/Air France verzorgen een dagelijkse vlucht van Maastricht naar Bordeaux. Op 12 aselect gekozen dagen is de vliegtijd in minuten gemeten: 57 54 55 51 56 48 52 51 59 59 53 49.
  - a. Geef een 90%-betrouwbaarheidsinterval voor de gemiddelde vluchtduur. Ga daarbij uit van normaal verdeelde vliegtijden.
  - b. De vertrektijd van de vlucht is 10.00 uur. KLM wil een zodanige aankomsttijd publiceren dat 95% van de vluchten niet later dan dat tijdstip zullen arriveren. Bereken, in minuten nauwkeurig, de te publiceren aankomsttijd.

(N.B. Er wordt hier niet gevraagd om een betrouwbaarheidsinterval voor de *gemiddelde* vliegtijd, maar voor de *vliegtijd van een toekomstige vlucht*. Noem deze vliegtijd  $X$ .  $X$  is normaal verdeeld met parameters  $\mu$  en  $\sigma$ . Wij zoeken de waarde van  $x$  waarvoor geldt  $\Pr(X \leq x) = 0,95$ . Ga er bij je berekening van uit dat  $\mu$  en  $\sigma$  gelijk zijn aan hun schatters  $\bar{X}$  en  $s$ .)

De onder b. uitgevoerde berekening is niet strikt correct omdat je geen rekening hebt gehouden met de onzekerheid die het gevolg is van het gelijkstellen van  $\mu$  met  $\bar{X}$  en van  $\sigma$  met  $s$ . In de volgende onderdelen ga je proberen om een theoretisch correcte berekening uit te voeren.

## 2.2. Populatiepercentages

- c. Toon aan dat  $T = \frac{X - \bar{X}}{s\sqrt{13/12}}$  een t-verdeling heeft. Met hoeveel vrijheidsgraden?
- d. Maak de theoretisch correcte berekening van de bij b. bedoelde waarde van  $x$ .

## 2.2 Populatiepercentages

Stel we willen weten welk percentage van de Nederlandse volwassen mannen drager is van een bepaald virus. We nemen daartoe een aselechte steekproef van 50 mannen en stellen vast dat 6 van hen drager zijn van het virus. 12% van de steekproef is dus virusdrager, dus het ligt voor de hand om het percentage virusdragers van gehele populatie ook op 12% te schatten, maar hoe betrouwbaar is deze schatting?

De proportie virusdragers van de totale populatie noemen we  $\pi$ , ( $100\pi\%$  is dus virusdrager). Als we een steekproef van lengte 50 nemen uit de populatie en we noemen het aantal virusdragers uit de steekproef  $X$ , dan is  $X$  binomiaal verdeeld met parameters  $n=50$  en  $\pi$ , waarbij  $\pi$  (de kans dat een willekeurige man virusdrager is) onbekend is. De voor de hand liggende schatting van  $\pi$  is  $p = X/n = 6/50 = 0,12$ .

**Opmerking 2.2.1.** Wij hebben al eerder aangegeven dat we in deze module de Griekse letter  $\pi$  gebruiken voor de kans op succes bij de binomiale verdeling. In eerdere modules gebruikte je hiervoor de letter  $p$ . In deze module gebruiken we de letter  $p$  voor de *schatting van  $\pi$* .  $\pi$  is dus een vast getal, een parameter.  $p$  is een kansvariabele: bij een nieuwe steekproef krijg je een nieuwe realisatie van  $p$ .  $\pi$  blijft echter ongewijzigd.

Het zal duidelijk zijn dat, als de werkelijke waarde van  $\pi$  0,12 is, de steekproefuitkomst  $X = 6$  heel plausibel is. Maar bij een werkelijke waarde van  $\pi$  van 0,5, is deze uitkomst onwaarschijnlijk laag. Immers:  $\Pr(X \leq 6 | \pi = 0,5) \approx \text{binomcdf}(50, 0.5, 6) = 1,6 \cdot 10^{-8}$ .

Omgekeerd, bij een waarde van  $\pi$  van 0,02 is de uitkomst  $X = 6$  onwaarschijnlijk hoog, want  $\Pr(X \geq 6 | \pi = 0,02) \approx 1 - \text{binomcdf}(50, 0.02, 5) = 4,8 \cdot 10^{-4}$ .

### Opgaven

35. Voor welke waarde van  $\pi$  geldt:  $\Pr(X \leq 6) = 0,025$ ?

36. Voor welke waarde van  $\pi$  geldt:  $\Pr(X \geq 6) = 0,025$ ?

Uit de oplossingen van bovenstaande opgaven volgt dat  $[0,045 ; 0,243]$  een 95%-betrouwbaarheidsinterval is voor  $\pi$ , de werkelijke proportie virusdragers in de totale populatie.

## Schattingen van parameters en betrouwbaarheidsintervallen

Samengevat:

Bij een steekproefuitkomst van  $X = x$  wordt een exact 95%-betrouwbaarheidsinterval voor  $\pi$  gevonden door de vergelijkingen  $\Pr(X \leq x) = 0,025$  en  $\Pr(X \geq x) = 0,025$  op te lossen. In deze vergelijkingen is  $x$  gegeven,  $X$  is binomiaal verdeeld met parameters  $n$  (bekend) en  $\pi$  (de onbekende van de vergelijkingen).

In de praktijk worden dergelijke betrouwbaarheidsintervallen dikwijls bepaald door gebruik te maken van het feit dat voor “grote” waarden van  $n$ , de verdeling van  $p$  bij benadering normaal verdeeld is.

Er geldt  $E(X) = n\pi$  en  $Var(X) = n\pi(1 - \pi)$ .

Hieruit volgt:  $E(p) = E(X/n) = \pi$  en  $Var(p) = Var(X/n) = \pi(1 - \pi)/n$ .

We krijgen dan dat  $\frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}}$  bij benadering standaard-normaal verdeeld is.

Als we nu  $\pi$  vervangen door zijn schatter  $p$ , dan wordt, bij “grote” waarden van  $n$ , de standaardafwijking van  $p$  gegeven door  $SEM = \sqrt{p(1 - p)/n}$ .

Definieer nu  $z_{0,05}$  als de waarde waarvoor bij een standaardnormaal verdeelde kansvariabele  $Z$  geldt  $\Pr(-z_{0,05} \leq Z \leq z_{0,05}) = 0,95$  en we hebben:

$p \pm z_{0,05} \cdot SEM$  is een 95%-betrouwbaarheidsinterval voor  $\pi$

In de praktijk wordt de benadering d.m.v. de normale verdeling bevredigend geacht als  $np$  en  $n(1 - p)$  beide groter zijn dan 5.

**Opmerking 2.2.2.** De GR kent een standaardfunctie voor de bepaling van een betrouwbaarheidsinterval voor  $\pi$ :

STAT - TESTS - 1PropZInt.

Aan de toevoeging “Z” kun je zien dat gebruikt wordt gemaakt van de benadering via de normale verdeling. De invoer spreekt voor zich.

## Opgaven

37. Bereken in het voorbeeld van het begin van deze paragraaf het 95%-betrouwbaarheidsinterval voor  $\pi$  met behulp van de normale verdeling en vergelijk dit antwoord met de exacte uitkomst.
38. Van de 64 ondervraagde leerlingen uit klas 4 zeggen er 27 dat ze, buiten de rechtstreekse voorbereiding voor de toetsweek nooit

## 2.2. Populatiepercentages

meer dan 2 uur per week aan huiswerk besteden. Geef het 90%-betrouwbaarheidsinterval voor het werkelijke percentage leerlingen dat op deze manier met huiswerk omgaat. Maak een exacte berekening en een benadering m.b.v. de normale verdeling.

39. De belastingdienst vermoedt dat wiskundeleraren in het vwo dikwijls met bijlessen zwart bijverdienen. De dienst neemt daarom een steekproef van 25 wiskundeleraren die geen neveninkomsten opvoeren bij hun belastingaangifte. Deze 25 worden aan een intensief onderzoek onderworpen: oude agenda's worden gecontroleerd, leerlingen en collega's worden ondervraagd etc. Uiteindelijk wordt bij 1 van deze leraren vastgesteld dat hij inkomsten uit bijlessen ten onrechte niet heeft aangegeven. De proportie wiskundeleraren die zwart bijverdient aan bijlessen (als deel van het aantal dat geen neveninkomsten uit bijlessen aangeeft bij de belastingdienst) noemen we  $\pi$ .
- Geef het 95%-betrouwbaarheidsinterval voor  $\pi$ . Waarom mag je hier de benadering via de normale verdeling niet gebruiken? Wat gebeurt er met de ondergrens van het interval als je dat toch doet?
  - De belastingdienst is niet geïnteresseerd in een ondergrens voor  $\pi$ . Men zoekt een 95%-betrouwbaarheidsinterval voor  $\pi$  van de vorm  $[0, a]$ . Bereken  $a$ .

*Schattingen van parameters en betrouwbaarheidsintervallen*

## Hoofdstuk 3

# t-toetsen voor populatiegemiddeldes

In de eerste kennismaking met het toetsen van hypothesen over het gemiddelde  $\mu$  van een normaal verdeelde populatie gingen we steeds uit van een bekende populatievariantie  $\sigma^2$ . We hanteerden als toetsingsgrootte het steekproefgemiddelde  $\bar{X}$ .  $\bar{X}$  is een kansvariabele (want afhankelijk van de steekproef) waarvan de verdeling bekend is: normaal met gemiddelde  $\mu$  (bepaald door de nulhypothese) en variantie  $\sigma^2/n$ . Op deze manier konden de overschrijdingskansen en kritische grenzen worden berekend met behulp van de normale verdeling. In de literatuur staan dergelijke toetsen bekend als *Z-toetsen*, omdat de letter *Z* vaak gebruikt wordt voor een standaard-normaal verdeelde kansvariabele.

In de praktijk komt het echter zelden voor dat de populatievariantie bekend is. Je zal deze moeten schatten met behulp van de waarnemingen uit je steekproef. De *Z*-toets is dan niet meer bruikbaar en moet vervangen worden door de *t*-toets, die in de praktijk juist heel dikwijls gebruikt wordt. Het is echter van belang om je te realiseren dat we nog steeds uitgaan van normaal verdeelde uitkomsten. In de praktijk zul je je er dus van moeten vergewissen dat deze aanname redelijk is. Een visuele inspectie van de gegevens, met behulp van een histogram, is vaak al voldoende. Je moet daarbij vooral alert zijn op twee zaken:

- De data moeten redelijk symmetrisch zijn. Het gemiddelde en de mediaan van de steekproef moeten niet wezenlijk van elkaar verschillen;
- De data moeten weinig uitschieters (outliers) kennen. Ongeveer 95% van de steekproefgegevens moet binnen 2 standaardafwijkingen van het gemiddelde liggen.

Er bestaan formele statistische toetsen die een uitspraak doen over de normaliteit van een populatie, maar deze vallen buiten het bestek van

### *t*-toetsen voor populatiegemiddeldes

deze module en zijn in de praktijk ook meestal niet nodig. Als je moet concluderen dat de gegevens niet normaal verdeeld zijn, dan kun je proberen om de data te transformeren, zodat de getransformeerde data wél normaal verdeeld zijn. Je kunt bijvoorbeeld kijken naar de logaritme van de data in plaats van naar de data zelf. We zullen dit verder niet behandelen. Een andere mogelijkheid is om over te schakelen op een zogenaamde verdelingsvrije of niet-parametrische toets. Dit onderwerp zal in hoofdstuk 4 aan de orde komen.

In dit hoofdstuk behandelen we drie verschijningsvormen van de *t*-toets. We kunnen te maken hebben met uitspraken over *één* populatiegemiddelde, of we vergelijken *twee* populatiegemiddeldes. In het laatste geval maken we een onderscheid tussen *ongepaarde* en *gepaarde* waarnemingen.

## 3.1 De *t*-toets : één populatiegemiddelde

**Voorbeeld 3.1.1.** De fabrikant van Tekkel hotdogs uit Voorbeeld 2.1.1 beweert dat zijn hotdogs een gemiddeld vetgehalte hebben van 20%. Ga na of, bij een significantieniveau van  $\alpha = 0,05$ , de steekproef uit het voorbeeld aanleiding geeft om aan de bewering van de fabrikant te twijfelen.

#### *Uitwerking*

We toetsen de nulhypothese  $H_0 : \mu = 20,0$  tegen het alternatief  $H_1 : \mu \neq 20,0$ , bij een onbekende  $\sigma$ . We weten (zie paragraaf 2.1) dat onder de nulhypothese de kansvariabele  $T = \frac{\bar{X} - 20}{SEM}$  een *t*-verdeling heeft met  $n - 1 = 9$  vrijheidsgraden. We zullen de hypothese dus verwerpen als  $|T| \geq t_{[9],0,05} = \text{invT}(0.975, 9) = 2,262$ .

In deze steekproef geldt  $T \approx \frac{21,9 - 20}{1,307} \approx 1,453$ .

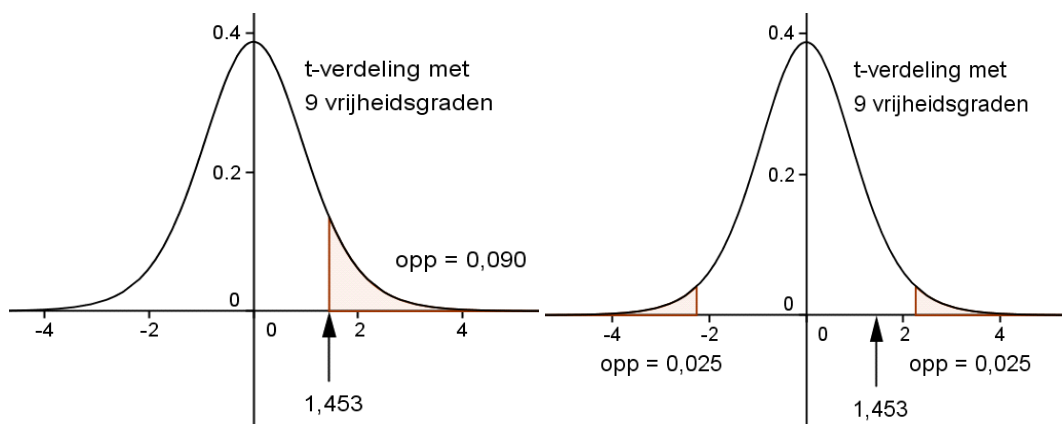
Er is dus geen aanleiding om de hypothese te verwerpen.

We kunnen ook de bij de steekproefuitkomst behorende overschrijdingskans berekenen. Deze is  $\Pr(T \geq 1,453) = \text{tcdf}(1.453, 10^99, 9) = 0,090$ . Omdat  $0,090 > \frac{1}{2}\alpha$  komen we tot dezelfde conclusie: er is geen aanleiding om de hypothese te verwerpen.

Beide aanpakken zijn weergegeven in de figuur hieronder. In het rechterplaatje zie je dat de waargenomen *T* niet in het kritische gebied valt. In het linkerplaatje is de overschrijdingskans weergegeven.



### 3.1. De t-toets : één populatiegemiddelde



We kunnen als volgt samenvatten:

Bij het toetsen van de hypothese  $H_0 : \mu = \mu_0$  met een onbekende  $\sigma$  gebruiken we de toetsingsgrootte  $T = \frac{\bar{X} - \mu_0}{SEM}$ . Onder  $H_0$  heeft  $T$  een t-verdeling met  $n - 1$  vrijheidsgraden.

**Opmerking 3.1.1.** We hebben in het voorbeeld te maken met een tweezijdige toets. Bij een eenzijdige toets zouden we de overschrijdingskans moeten vergelijken met  $\alpha$  in plaats van met  $\frac{1}{2}\alpha$ . Ook bij de berekening van de kritische grens  $t_{[9],0,05} = 2,262$  hebben we rekening gehouden met het tweezijdige karakter van de toets. Reken zelf na dat in het geval van een eenzijdige toets ( $H_1 : \mu > 20,0$ ) we zouden verwerpen als  $T > 1,833$ .

**Opmerking 3.1.2.** De toetsingsgrootte  $T$  wordt ook gebruikt voor de berekening van betrouwbaarheidsintervallen (zie paragraaf 2.1). Er is dan ook een rechtstreeks verband tussen het toetsen van een hypothese over  $\mu$  en het betrouwbaarheidsinterval. Bij een tweezijdige toets kunnen we simpelweg stellen dat we de hypothese kunnen verwerpen bij een significantieniveau van  $\alpha$ , als de waarde voor  $\mu$  uit de nulhypothese niet binnen het  $100(1 - \alpha)\%$  betrouwbaarheidsinterval rondom  $\bar{X}$  ligt.

**Opmerking 3.1.3.** In de literatuur en in statistische software komt naast het begrip “overschrijdingskans” ook de term “*p-waarde*” (Engels: *p-value*) voor. Bij een eenzijdige toets zijn de twee begrippen identiek, maar bij een tweezijdige toets is de *p*-waarde gelijk aan twee maal de overschrijdingskans. Je mag een hypothese dus verwerpen als de *p*-waarde kleiner is dan  $\alpha$ , ongeacht het alternatief. Immers, bij de berekening van de *p*-waarde is al rekening gehouden met het alternatief. Als je een toets uitvoert met de GR, wordt ook de *p*-waarde als uitkomst gegeven. Hoe dat gaat zie je in het volgende voorbeeld.

**Voorbeeld 3.1.2.** We voeren de toets van voorbeeld 3.1.1 uit met de GR en gebruiken daarbij de originele gegevens uit voorbeeld 2.1.1.

- Voer eerst de waarnemingen uit je steekproef (de  $x_i$ 's) in in de lijst L1 (via STAT - EDIT).
- Ga naar STAT - TESTS - T-Test
- Kies bij Inpt voor Data. (Net als bij de berekening van een betrouwbaarheidsinterval kun je ook voor Stats kiezen. De gang van zaken is analoog).
- Bij  $\mu_0$ : voer je 20 in, bij List: L1, bij Freq: 1 en bij  $\mu$ : kies je het alternatief  $\neq \mu_0$
- Calculate levert o.a. een *p*-waarde van 0,180. Hieraan zie je dat je bij een significantieniveau van 0,05 (en ook bij een significantieniveau van 0,1) de nulhypothese niet kunt verwerpen.

## Opgaven

40. Formuleer het verband tussen een eenzijdige toets en een betrouwbaarheidsinterval.
41. Lisa, woordvoerder van de actiegroep “Leef gezond”, beweert dat het gemiddelde vetgehalte in kroketten van een bekend merk tenminste 20% bedraagt, de krokettenfabrikant bestrijdt dit. Ze laten een onderzoeksbureau een steekproef nemen van 15 kroketten uit willekeurige winkels in Nederland. Van elk van deze kroketten wordt het vetgehalte (in procenten) gemeten. Ze vinden:  
19,3 19,9 20,6 18,9 21,1 20,3 18,7 19,6 19,6 18,8 19,6 19,9 20,1 19,2 20,4

### 3.2. Twee populatiegemiddeldes, ongepaarde waarnemingen

- a. Mag je in dit geval aannemen dat de gegevens normaal verdeeld zijn?
  - b. Kan de fabrikant de bewering van Lisa met succes (d.w.z. bij een significantieniveau van 5%) bestrijden? Wees duidelijk in de formulering van je hypothese.
42. Op de verpakking van een diepvriesproduct staat vermeld “bevat 240 kcal”. De wareninspectie voert een controle uit, neemt een aselechte steekproef van 12 pakken en meet de caloriewaarde van de inhoud. De boxplot van de waarnemingen wordt bekeken en lijkt niet in strijd te zijn met de aanname dat de waarnemingen afkomstig zijn uit een normale verdeling. Er geldt (in kcal):  $\bar{x} = 244,3$  en  $s = 12,4$ .
- a. Geef een 95% betrouwbaarheidsinterval voor de gemiddelde hoeveelheid kcal in een pak. Maak een handmatige berekening en controleer je antwoord met de GR
  - b. Toets de geldigheid van hetgeen op de verpakking vermeld staat met  $\alpha = 0,1$ . Maak een handmatige berekening en controleer je antwoord met de GR
  - c. Je neemt een willekeurig pak. Wat is de kans dat dit pak meer dan 250 kcal bevat? Geef een benadering m.b.v. de normale verdeling en voer een exacte berekening uit m.b.v. de t-verdeling. (Zie ook opgave 34.)

## 3.2 De t-toets: vergelijking van twee populatiegemiddeldes, ongepaarde waarnemingen

“Zijn 17-jarige jongens uit Oegstgeest langer dan die uit Leiden?” Dit is typisch een vraag waarin de gemiddelden uit twee verschillende populaties worden vergeleken.

De twee populaties worden weergegeven door de (onafhankelijke) kansvariabelen  $X_1$  (lengte van een willekeurige 17-jarige jongen uit Oegstgeest) en  $X_2$  (idem, maar dan uit Leiden). Veronderstel verder dat beide variabelen onafhankelijk en normaal verdeeld zijn met parameters  $\mu_1$  en  $\sigma_1$  respectievelijk  $\mu_2$  en  $\sigma_2$ . Uit beide populaties hebben we een steekproef van lengte  $n_1$  respectievelijk  $n_2$ .

**Opmerking 3.2.1.** Tot dusver gebruikten we de notatie  $X_1$  en  $X_2$  om de eerste en de tweede waarneming uit een steekproef  $X_i$  ( $i = 1, \dots, n$ ) uit één populatie aan te geven. Nu wordt het subscript gebruikt om de populatie aan te duiden. De steekproef uit de eerste populatie noteren we als  $X_{1i}$  ( $i = 1, \dots, n_1$ ). Uit de context moet duidelijk zijn wat er bedoeld wordt.

### *t-toetsen voor populatiegemiddeldes*

We willen de hypothese  $H_0 : \mu_1 = \mu_2$  toetsen tegen het alternatief  $H_1 : \mu_1 > \mu_2$ .

Het ligt voor de hand dat we de steekproefgemiddeldes  $\bar{X}_1$  en  $\bar{X}_2$  vergelijken. We zullen  $H_0$  verwerpen als  $\bar{X}_1 - \bar{X}_2$  “groot genoeg” is.

Als  $H_0$  waar is, dan heeft  $Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$  een standaardnormale verdeling. Dit ga je aantonen in opgave 43. Om dit resultaat in de praktijk te gebruiken om een toets uit te voeren, is het nodig dat  $\sigma_1$  en  $\sigma_2$  bekend zijn.

Als  $\sigma_1$  en  $\sigma_2$  niet bekend zijn, kunnen alleen exacte resultaten worden verkregen als de veronderstelling  $\sigma_1 = \sigma_2$  gerechtvaardigd is. De gemeenschappelijke waarde van  $\sigma_1$  en  $\sigma_2$  noemen we  $\sigma$ . Met  $s_1^2$  respectievelijk  $s_2^2$  geven we de steekproefvariantie aan van  $X_1$  respectievelijk  $X_2$ . Beide grootheden zijn een zuivere schatter van  $\sigma^2$ , die ieder gebruik maken van slechts een deel van de waarnemingen. Het ligt daarom voor de hand om beide schatters te combineren tot één nieuwe die gebruik maakt van alle waarnemingen. Het gewogen gemiddelde  $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$  is een voor de hand liggende keuze.

In opgave 44 toon je vervolgens aan dat  $T = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{1/n_1 + 1/n_2}}$  een t-verdeling heeft met  $n_1 + n_2 - 2$  vrijheidsgraden.

De toetsingsgrootte  $T$  kun je ook gebruiken om een betrouwbaarheidsinterval voor  $\mu_1 - \mu_2$  op te stellen. Zie opgave 45.

Samenvattend:

Bij het toetsen van  $H_0 : \mu_1 = \mu_2$  (ongepaarde waarnemingen, uit twee onafhankelijke populaties met gemeenschappelijke variantie  $\sigma^2$ ) gebruiken we de toetsingsgrootte  $T = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{1/n_1 + 1/n_2}}$ .  
Onder  $H_0$  heeft  $T$  een t-verdeling met  $n_1 + n_2 - 2$  vrijheidsgraden.  
 $T$  kan ook worden gebruikt voor betrouwbaarheidsintervallen voor  $\mu_1 - \mu_2$ .

### **Opgaven**

43. Laat zien dat, onder  $H_0$ , de kansvariabele  $Z$ , zoals gedefinieerd in bovenstaande paragraaf standaardnormaal verdeeld is.
44.
  - a. Wat is de verdeling van de steekproefvarianties  $s_1^2$  en  $s_2^2$ ? Gebruik paragraaf 2.1.
  - b. Wat is de verdeling van  $s^2$ ? Gebruik paragraaf 1.5.1.
  - c. Laat zien dat  $T$  een t-verdeling heeft met  $n_1 + n_2 - 2$  vrijheidsgraden. Gebruik paragraaf 1.5.2 en 2.1.

### 3.2. Twee populatiegemiddeldes, ongepaarde waarnemingen

45. Gebruik de toetsingsgrootte  $T$  uit de paragraaf hierboven om een betrouwbaarheidsinterval voor  $\mu_1 - \mu_2$  op te stellen.

**Opmerking 3.2.2.** De theorie van deze paragraaf berust (o.a.) op de veronderstelling dat de standaardafwijking van beide populaties even groot is ( $\sigma_1 = \sigma_2 = \sigma$ ). Als dit (duidelijk) niet het geval is, bestaat er een toets die hiermee rekening houdt. De meeste softwarepakketten (waaronder de GR) bieden de mogelijkheid om deze toets te gebruiken. Doorgaans wordt geadviseerd om deze alternatieve toets alleen te gebruiken als de beide standaardafwijkingen sterk (meer dan een factor 4) van elkaar verschillen.

**Opmerking 3.2.3.** De theorie van deze paragraaf berust ook op de veronderstelling dat de waarnemingen afkomstig zijn uit een normale verdeling. Als deze veronderstelling niet juist is, is het echter *bij grote aantallen waarnemingen* nog steeds gerechtvaardigd om de  $t$ -toets toe te passen. Bij duidelijk scheve verdelingen en minder grote aantallen waarnemingen kan worden uitgeweken naar een niet-parametrische toets. Zie hoofdstuk 4.

**Opmerking 3.2.4.** In paragraaf 3.1 gebruikten we een  $t$ -verdeling waarvan het aantal vrijheidsgraden gelijk was aan het aantal waarnemingen min 1. In deze paragraaf is het overeenkomstige aantal vrijheidsgraden gelijk aan het aantal waarnemingen min 2. Kennelijk moet bij het bepalen van het aantal vrijheidsgraden het aantal waarnemingen worden verminderd met het aantal parameters dat we moeten schatten alvorens de populatievariantie kan worden berekend. Dit principe zullen we later nog vaak tegenkomen.

**Voorbeeld 3.2.1.** “Hebben kinderen met ADHD een kleiner hersenvolume dan kinderen zonder ADHD?” Deze vraag werd gesteld in een artikel in de Journal of the American Medical Association uit 2002. Hiertoe werd met behulp van hersenscans de herseninhoud (in ml) gemeten van een groep kinderen met en zonder de diagnose ADHD. (In de praktijk is het niet eenvoudig om ervoor te zorgen dat beide groepen goed vergelijkbaar zijn, d.w.z. dat ADHD het enige systematische verschil is. Zo zul je ervoor willen zorgen dat de samenstelling van beide groepen naar leeftijd, geslacht, sociaal-economische status etc dezelfde is. We gaan er in dit voorbeeld gemakshalve van uit dat dit het geval is.) De verzamelde gegevens kunnen als volgt worden samengevat:

	$n$	$\bar{x}$	$s$
met ADHD	152	1059,4	117,5
zonder ADHD	139	1104,5	111,3

### *t-toetsen voor populatiegemiddeldes*

Geef de populatie kinderen met ADHD aan met de index 1 en die zonder ADHD met 2. Wij toetsen dan  $H_0 : \mu_1 = \mu_2$  tegen  $H_1 : \mu_1 < \mu_2$ .

Er geldt:  $s^2 = \frac{151 \cdot 117,5^2 + 138 \cdot 111,3^2}{152 + 139 - 2} \approx 13129$ .

De toetsingsgrootte  $T$  is in dit geval gelijk aan

$$T = \frac{1059,4 - 1104,5}{\sqrt{13129(152^{-1} + 139^{-1})}} \approx -3,35.$$

De bijbehorende overschrijdingskans is  $\Pr(T \leq -3,35) \approx 5 \cdot 10^{-4}$  (289 vrijheidsgraden). De hypothese moet dus worden verworpen.

Een 95%-betrouwbaarheidsinterval voor  $\mu_1 - \mu_2$  wordt gegeven door  $1059,4 - 1104,5 \pm t_{[289],0,05} \cdot s\sqrt{152^{-1} + 139^{-1}}$ , dus  $[-71,6 ; -18,6]$ .

Ook hier geldt dat het weliswaar belangrijk is om bovenstaande berekening goed te doorgronden, maar dat je hem ook eenvoudig met de GR kunt uitvoeren.

Voor het toetsen van de hypothese:

- Ga naar STAT - TESTS - 2-SampleTTest;
- Kies (in dit geval) voor Stats. Als je alleen beschikt over de oorspronkelijke waarnemingen en deze hebt ingevoerd in L1 en L2, kies je voor Data;
- Voer in:  
 $\bar{x}1$ : 1059.4; Sx1: 117.5; n1: 152;  
 $\bar{x}2$ : 1104.5; Sx2: 111.3; n2: 139;  
 $\mu1$ : <  $\mu2$   
Pooled: Yes  
(Als je No invult, wordt rekening gehouden met verschillende standaardafwijkingen. Dit is alleen nodig als de standaardafwijkingen meer dan een factor 4 van elkaar verschillen. Zie Opmerking 3.2.2);
- Calculate geeft de resultaten.

Voor de berekening van het betrouwbaarheidsinterval:

- Ga naar STAT - TESTS - 2-SampleTInt;
- Invoer als hierboven, met C-level: 0.95;
- Calculate geeft de resultaten.

### 3.2. Twee populatiegemiddeldes, ongepaarde waarnemingen

#### Opgaven

46. “Pilgebruiksters krijgen broze botten.” Om het waarheidsgehalte van deze uitspraak te toetsen is onderzoek gedaan naar de “Bone Mineral Density” (BMD in gram per cc) bij twee groepen vrouwen verdeeld in vrouwen die nooit orale contraceptiva hadden gebruikt (groep 1 “nooit”) en een groep die dit gedurende tenminste 3 maanden wel hadden gedaan (groep 2 “ooit”). Ook hier geldt weer dat het niet eenvoudig zal zijn om de groepen zo samen te stellen dat een goede vergelijking kan worden gemaakt, maar we gaan er nu gemakshalve van uit dat dit in orde is. De verzamelde gegevens zijn als volgt:

Nooit	0,82	0,94	0,96	1,31	0,94	1,21	1,26	1,09	1,13	1,14
Ooit	0,94	1,09	0,97	0,98	1,14	0,85	1,30	0,89	0,87	1,01

- Voer de gegevens in in je GR, maak per groep een boxplot en beoordeel of het redelijk is om aan te nemen dat de gegevens van elke groep normaal verdeeld zijn met gelijke varianties.
  - Stel een toets op die antwoord moet geven op de onderzoeksvraag en voer deze toets uit bij een significantieniveau van 0,05. Maak een handmatige berekening en controleer de uitkomst met de GR,
  - Voer nu met je GR de vergelijkbare toets uit die niet uitgaat van gelijke varianties en vergelijk je antwoord met wat je bij b. gevonden hebt.
  - Bereken het 90%-betrouwbaarheidsinterval voor  $\mu_1 - \mu_2$  met de GR en beschrijf het verband van dit interval met de uitkomst van b.
47. Voor haar profielwerkstuk onderzoekt Amy of mensen met een knap uiterlijk door derden andere eigenschappen toegedicht krijgen dan mensen met een lelijk uiterlijk. Ze heeft daarvoor een verzameling van pasfoto's: 25 foto's van “knappe” en 22 foto's van “lelijke” mensen. (Het oordeel “knap” of “lelijk” is tot stand gekomen in een eerder onderzoek.) Deze 47 foto's worden in een willekeurige volgorde voorgelegd aan een proefpersoon, die bij iedere foto de vraag “Lijkt deze persoon je sympathiek?” moest beantwoorden door aan de foto de score 1 (helemaal niet sympathiek), 2, 3 of 4 (erg sympathiek) toe te kennen. De resultaten van het onderzoek zijn als volgt:

Aantal foto's met score	1	2	3	4
”Knap”	0	4	15	6
”Lelijk”	2	5	14	1

- “Score” is een discrete kansvariabele en kan dus niet normaal ver-

### *t*-toetsen voor populatiegemiddeldes

deeld zijn. Waarom kun je in deze situatie toch een *t*-toets toepassen?

- b. Worden pasfoto's van knappe mensen op het kenmerk "sympathiek" significant anders beoordeeld dan pasfoto's van lelijke mensen? Wees expliciet in de formulering van je hypothese. Geef ook een relevant betrouwbaarheidsinterval.

### 3.3 De *t*-toets: vergelijking van twee populatiegemiddeldes, gepaarde waarnemingen

Gaat je hart sneller kloppen nadat je koffie hebt gedronken? Om antwoord te geven op deze vraag werd de hartslag gemeten van een twaalfstal zesde-klas-leerlingen die een etmaal lang geen koffie hadden gedronken. Direct na deze meting kregen ze een kop sterke koffie te drinken en een half uur later werd hun hartslag opnieuw gemeten. De resultaten (hartslagen per minuut) waren als volgt:

Voor	72	56	48	87	85	71	85	73	81	90	71	65
Na	71	51	46	95	78	69	92	80	85	92	85	69

Ook hier hebben we te maken met de vergelijking van de gemiddeldes uit twee populaties. De eerste is de populatie van hartslagen voor en de andere de populatie van hartslagen na het drinken van koffie. Toch zou het niet correct zijn om de analyse van paragraaf 3.2 voor deze data te gebruiken. De steekproeven uit beide populaties zijn namelijk niet onafhankelijk, omdat de waarnemingen bij dezelfde leerlingen zijn gedaan. We hebben hier te maken met zogenaamde gepaarde waarnemingen. Het is belangrijk om het verschil tussen gepaarde en ongepaarde waarnemingen goed te begrijpen. Zo horen in de tabel hierboven de twee getallen uit eenzelfde kolom bij elkaar, want ze hebben betrekking op dezelfde leerling. In de tabel van opgave 46 daarentegen, hebben de cijfers uit eenzelfde kolom niets met elkaar te maken. Bij gepaarde waarnemingen zal het aantal waarnemingen uit beide populaties noodgedwongen hetzelfde zijn. Bij ongepaarde waarnemingen hoeft dat niet. (In opgave 46 is het toevallig wel zo.)

Bij de analyse van bovenstaande gepaarde waarnemingen gaan we juist gebruik maken van het feit dat er een verband bestaat tussen de waarnemingen uit beide groepen. We richten onze aandacht op het verschil tussen de hartslag voor en na. Definieer  $X$  als het verschil van de hartslag na en de hartslag voor het drinken van koffie. De waarnemingen zien er dan als volgt uit:

$x$	-1	-5	-2	8	-7	-2	7	7	4	2	14	4
-----	----	----	----	---	----	----	---	---	---	---	----	---

In plaats van de hypothese  $H_0 : \mu_1 = \mu_2$  (vergelijking van de gemiddeldes voor en na) toetsen we de hypothese  $H_0 : \mu = 0$ , waarbij  $\mu$  het populatiegemiddelde van  $X$  is. Als alternatieve hypothese nemen we  $H_1 : \mu > 0$ .



### 3.3. Twee populatiegemiddeldes, gepaarde waarnemingen

We kunnen nu de analyse van paragraaf 3.1 uitvoeren. We vinden:  $\bar{x} = 2,42$ ,  $s = 6,08$ ,  $t = 1,376$  (11 vrijheidsgraden) en een overschrijdingkans van 0,098. (Zelf narekenen.) Bij een  $\alpha$  van 0,1 zouden we dus nog net verworpen.

Het 95%-betrouwbaarheidsinterval voor  $\mu = \mu_1 - \mu_2$  wordt gegeven door:

$$\bar{X} \pm t_{[n-1],0.05} \cdot SEM = 2,42 \pm 2,20 \cdot 6,08/\sqrt{12}, \text{ dus } [-1,45 ; 6,28].$$

Samengevat:

Bij het vergelijken van de gemiddeldes van twee populaties met gepaarde waarnemingen, baseren we de analyse op de verschillen van de gepaarde waarnemingen. De analyse verloopt dan zoals beschreven in paragraaf 3.1

### Opgaven

48. Geef in onderstaande situaties aan of er sprake is van gepaarde dan wel ongepaarde waarnemingen.
- Van aselekt gekozen flessen wijn uit een slijterij meten we het alcoholpercentage en vergelijken dit met het op het etiket vermelde percentage.
  - Om het effect van het innemen van gingko-supplementen op het geheugen te onderzoeken worden twee groepen proefpersonen (gebruikers van ginko en gebruikers van een placebo) onderworpen aan een geheugentest.
  - Aan voorverpakt vlees worden soms nitraten toegevoegd om de houdbaarheid te vergroten. Dit staat dan vermeld op de verpakking. Om het effect van deze toevoeging te onderzoeken wordt een aselechte steekproef van vergelijkbare soorten verpakt vlees met en zonder nitraten een dag lang bewaard bij kamertemperatuur en wordt vervolgens het aantal bacteriën gemeten.
  - Is er een verband tussen de kwaliteit van het lange- en het kortetermijngeheugen? Om dit te onderzoeken krijgt een groep aselekt gekozen leerlingen 10 minuten de tijd om een lijst onbekende Griekse woorden te leren. Ze worden een uur later overhoord en een dag later opnieuw.
49. Nikki werkt samen met Amy aan het profielwerkstuk waarvan in opgave 47 een experiment is beschreven. De conclusie uit dit onderzoek was (kort gezegd) dat knappe mensen op basis van hun pasfoto sympathieker overkomen dan lelijke mensen. Nikki beseft dat het onderzoek alleen aantoonde dat dit waar is voor de ene proefpersoon die de foto's heeft beoordeeld. Het is heel goed mogelijk dat een andere proefpersoon andere conclusies aan de pasfoto's verbindt. Daarom besluiten

### *t-toetsen voor populatiegemiddeldes*

Nikki en Amy tot een nieuw experiment, nu met 12 proefpersonen die als aselechte steekproef kunnen gelden uit gymnasiumleerlingen van 16, 17 of 18 jaar. Elk van deze proefpersonen krijgt 6 pasfoto's te zien, 3 knappe en 3 lelijke, maar dat wordt er niet bij gezegd. Gelukkig is de collectie pasfoto's van Amy groot genoeg om ervoor te zorgen dat iedereen verschillende foto's te zien krijgt. Aan elke foto moeten ze een score toekennen net als in opgave 47. De toegekende scores worden geregistreerd en als volgt samengevat:

Knappe foto's:	gemiddelde score: 3,11	standaardfout (s): 1,03
Lelijke foto's:	gemiddelde score: 2,72	standaardfout (s): 0,94

Amy en Nikki brengen deze gegevens naar Merle en vragen haar om een statistische analyse te maken. Merle zegt dat deze gegevens niet goed bruikbaar zijn en dat ze de originele data op een andere manier moeten samenvatten.

- a. Waarom komt Merle tot deze uitspraak? Waarom kan ze geen "2-Sample t-test" met deze gegevens uitvoeren?

Amy en Nikki gaan opnieuw aan het werk. Gelukkig hebben ze de door de proefpersonen ingevulde formulieren nog. Per proefpersoon berekenen ze de som van scores van mooie en lelijke foto's apart. Vervolgens nemen ze (nog steeds per proefpersoon) het verschil van de "mooie" minus de "lelijke" som. Ze vinden de volgende verschillen:

2, 5, -1, 0, 1, -2, 3, 0, 1, 1, 2, 2

- b. Maak een statistische analyse met  $\alpha = 0,05$  van deze gegevens. Wees zo volledig mogelijk in het formuleren en het rechtvaardigen van je aannames.

50. Een onderzoek richt zich op het effect van sporttraining op het niveau van melkzuur in het bloed. Acht mannen en zeven vrouwen deden mee aan het experiment. Het niveau van lactaat in hun bloed werd gemeten voor en na het spelen van drie wedstrijden squash. De resultaten waren:

### 3.3. Twee populatiegemiddeldes, gepaarde waarnemingen

MANNEN			VROUWEN		
Speler	Voor	Na	Speler	Voor	Na
1	13	18	1	11	21
2	20	37	2	16	26
3	17	40	3	13	19
4	13	35	4	18	21
5	13	30	5	14	14
6	16	20	6	11	31
7	15	33	7	13	20
8	16	19			

- Construeer een 90%-betrouwbaarheidsinterval voor de gemiddelde verandering van niveau van lactaat in het bloed voor mannen en vrouwen apart.
- Denk je dat deze gemiddelde verandering verschillend is voor mannen en vrouwen? Kun je de bij a. gevonden betrouwbaarheidsintervallen gebruiken om hier een uitspraak over te doen? Voer een statistische toets uit. Kies  $\alpha = 0,05$ .

*t-toetsen voor populatiegemiddeldes*

## Hoofdstuk 4

# Niet-parametrische toetsen

In hoofdstuk 3 gingen we steeds uit van kansvariabelen die normaal verdeeld zijn. Deze kansvariabelen worden geheel bepaald door twee parameters: het gemiddelde en de standaardafwijking ( $\mu$  en  $\sigma$ ). We veronderstelden doorgaans dat deze parameters onbekend zijn en gebruikten steekproeven om uitspraken te kunnen doen over deze parameters. We hebben ook gezien hoe we de aanname van normaliteit konden rechtvaardigen, bijvoorbeeld door een histogram van de gegevens te bekijken. Bij een grote steekproef zagen we dat, zelfs als de gegevens zelf niet normaal verdeeld zijn, de centrale limietstelling het gebruik van een t-toets kan rechtvaardigen. Dit alles neemt niet weg dat zich situaties kunnen voordoen waarbij het niet redelijk is om te veronderstellen dat de waarnemingen uit een normale verdeling komen en waar niet met succes een beroep op de centrale limietstelling kan worden gedaan. In dergelijke gevallen kan een niet-parametrische of verdelingsvrije toets uitkomst bieden.

In dit hoofdstuk bespreken we de eenvoudige tekentoets en de rangtekentoets van Wilcoxon die je kunt gebruiken bij het vergelijken van gepaarde waarnemingen. Verder komt de toets van Wilcoxon, die ook wel de Mann-Whitney toets wordt genoemd, aan de orde. Deze toets kun je gebruiken bij de analyse van ongepaarde waarnemingen uit twee onafhankelijke populaties. Daarmee krijgen we voor elk type t-toets een niet-parametrisch equivalent. Er is echter geen eenvoudig te berekenen non-parametrisch equivalent van betrouwbaarheidsintervallen.

### 4.1 Gepaarde waarnemingen: de tekentoets

We beginnen met een voorbeeld.

In een ziekenhuis worden twee verschillende soorten apparaten gebruikt om de bloeddruk te meten. Het is bekend dat het meten van de bloeddruk altijd gepaard gaat met (soms aanzienlijke) meetfouten. Men wil nu onderzoeken

### Niet-parametrische toetsen

of er sprake is van een *systematisch* verschil tussen beide apparaten. Daarom wordt bij 22 patiënten de bloeddruk met beide apparaten gemeten. De resultaten zijn weergegeven in de onderstaande tabel.

Gemeten onderdruk				
Patiënt	Apparaat 1	Apparaat 2	Vershil	Teken V
1	75	73	2	+
2	86	94	-8	-
3	91	82	9	+
4	92	82	10	+
5	105	108	-3	-
6	86	65	21	+
7	89	75	14	+
8	80	81	-1	-
9	75	76	-1	-
10	77	77	0	0
11	97	80	17	+
12	73	89	-16	-
13	109	91	18	+
14	80	80	0	0
15	98	76	22	+
16	76	65	11	+
17	82	78	4	+
18	93	100	-7	-
19	91	76	15	+
20	69	63	6	+
21	114	102	12	+
22	69	80	-11	-

Zoals meestal bij gepaarde waarnemingen, concentreren we ons op het verschil  $V$  van de beide metingen. Bij een aantal patiënten is dit verschil positief, bij een aantal andere negatief en bij een restgroep is het verschil 0. In de tabel is dit aangegeven in de laatste kolom. Als er geen systematisch verschil is tussen de metingen van beide apparaten, is de kans op een positief verschil gelijk aan de kans op een negatief verschil. Als we de gevallen waarin dit verschil 0 is buiten beschouwing laten, dan zijn beide kansen gelijk aan  $\frac{1}{2}$ .

Noem het aantal waarnemingen  $n$  en het aantal waarnemingen met een verschil  $\neq 0$ ,  $n'$ . In dit voorbeeld geldt  $n = 22$  en  $n' = 20$ . Noem het aantal “minnetjes”  $X$ .  $X$  is een kansvariabele die binomiaal verdeeld is met parameters  $n'$  en  $\pi$ . We toetsen  $H_0 : \pi = \frac{1}{2}$  (geen systematisch verschil in meetresultaten) tegen  $H_1 : \pi \neq \frac{1}{2}$  met  $\alpha = 0,05$ . Door onze waarnemingen terug te brengen tot het tellen van het aantal “minnetjes” hebben we nu te maken met een eenvoudige binomiaaltoets. Het waargenomen aantal

#### 4.2. Gepaarde waarnemingen: Wilcoxon's rangteken toets

“minnetjes” is  $X = 7$ , dus de overschrijdingskans is  $\Pr(X \leq 7) \approx \text{binomcdf}(20, 0.5, 7) = 0.1316 > \frac{1}{2}\alpha$ .

We kunnen de hypothese niet verwerpen.

Samengevat:

Bij het vergelijken van twee populaties met gepaarde waarnemingen, kan de *tekentoets* (Engels: *sign test*) worden gebruikt. Deze toets is gebaseerd op de waargenomen verschillen uit de steekproef. Van de  $n$ (paren) waarnemingen worden de paren met een verschil gelijk aan nul buiten beschouwing gelaten, zodoende resteren  $n'$  waarnemingen. Het aantal waarnemingen  $X$  met een positief (of negatief) verschil is binomiaal verdeeld met parameters  $n'$  en  $\pi$ . We gebruiken vervolgens de binomiaaltoets om  $H_0 : \pi = \frac{1}{2}$  te toetsen tegen een geschikt alternatief.

#### Opgave

51. Een leerkracht op een basisschool wil de effectiviteit van een nieuw soort oefening voor het onderdeel “optellen” van een rekenmethode toetsen. Voordat ze begint met de oefening laat ze haar 20 leerlingen een rekenblad met optelsommen maken en telt ze per kind het aantal goede antwoorden  $G_0$ . Vervolgens laat ze de kinderen oefenen volgens de nieuwe rekenmethode. Daarna laat ze de kinderen opnieuw een vergelijkbaar rekenblad met optelsommen maken en ze telt het aantal goede antwoorden  $G_1$ . De resultaten vind je in de tabel.

Aantal goede antwoorden per kind

$G_0$	9	10	9	12	10	10	12	8	10	13
$G_1$	10	12	11	12	10	11	13	10	10	13
$G_0$	11	14	12	12	12	11	13	10	10	8
$G_1$	10	12	13	11	13	13	14	12	8	12

Kun je, bij  $\alpha = 5\%$ , concluderen dat de oefening een positief effect heeft op de prestaties van de kinderen?

#### 4.2 Gepaarde waarnemingen: Wilcoxon's rangteken toets

De tekentoets heeft als nadeel dat alleen rekening wordt gehouden met het teken van de verschillen en niet met de omvang ervan. Er bestaat een niet-parametrisch alternatief dat hier wél rekening mee houdt en dat daarom in de meeste praktijkgevallen een groter onderscheidingsvermogen heeft dan de tekentoets. Dit is Wilcoxon's rangtekentoets (Wilcoxon's rank sign test). We

### Niet-parametrische toetsen

zullen deze toets toepassen op het voorbeeld uit paragraaf 4.1. De procedure is als volgt.

- In eerste instantie elimineren we, net als bij de tekentoets, de verschillen die gelijk zijn aan 0 en houden zo  $n'$  van de oorspronkelijke  $n$  waarnemingen over.
- Vervolgens rangschikken we de overblijvende verschillen op volgorde van oplopende *absolute waarde* en kennen we ze een rangnummer toe. Verschillen met dezelfde absolute waarde, de zogenaamde “knopen” of “ties”, krijgen alle het gemiddelde van de bijbehorende rangnummers toegekend.
- Daarna krijgt ieder van de rangnummers hetzelfde teken als het oorspronkelijke verschil en berekenen we de som  $S$  van deze rangnummers met teken.

In ons voorbeeld krijgen we:

Patiënt	Apparaat 1	Apparaat 2	Vershil	Rang $ V $	Rang $ V $ met teken
10	77	77	0	*	
14	80	80	0	*	
8	80	81	-1	1,5	-1,5
9	75	76	-1	1,5	-1,5
1	75	73	2	3	3
5	105	108	-3	4	-4
17	82	78	4	5	5
20	69	63	6	6	6
18	93	100	-7	7	-7
2	86	94	-8	8	-8
3	91	82	9	9	9
4	92	82	10	10	10
22	69	80	-11	11,5	-11,5
16	76	65	11	11,5	11,5
21	114	102	12	13	13
7	89	75	14	14	14
19	91	76	15	15	15
12	73	89	-16	16	-16
11	97	80	17	17	17
13	109	91	18	18	18
6	86	65	21	19	19
15	98	76	22	20	20

en  $S = -1.5 - 1.5 + 3 + \dots + 20 = 111$ .



#### 4.2. Gepaarde waarnemingen: Wilcoxon's rangteken toets

De nulhypothese luidt: “beide populaties hebben dezelfde verdeling”, dus de waargenomen verschillen zijn afkomstig uit een symmetrische verdeling met gemiddelde (=mediaan) gelijk aan 0. Dit betekent bijvoorbeeld dat de kans dat een willekeurige waarneming rangnummer 8 heeft gelijk is aan de kans dat het rangnummer -8 is. De som van de rangnummers heeft daarom een verwachtingswaarde gelijk aan 0, als de nulhypothese waar is. Als daarentegen de nulhypothese niet waar is, bijvoorbeeld omdat één van de twee apparaten systematisch hogere meetresultaten produceert, dan zullen de grote verschillen overwegend hetzelfde teken hebben. We zullen de hypothese dus verwerpen als  $S$  een grote positieve of negatieve waarde aanneemt. Om de kritische grens bij een gegeven significantieniveau  $\alpha$  te bepalen, moeten we de kansverdeling van  $S$  onder de nulhypothese kennen. In opmerking 4.2.1 hieronder zullen we laten zien hoe deze kansverdeling kan worden berekend, maar we zullen ook zien dat deze berekening al snel te omslachtig wordt om met de hand uit te voeren.

**Opmerking 4.2.1.** Gemakshalve gaan we ervan uit dat onze waarnemingen geen knopen bevatten, dus we doen in het voorbeeld hierboven net alsof er waarnemingen zijn met rangnummers 1, 2, 12 en 13 i.p.v. twee keer 1,5 en twee keer 11,5.

De maximale waarde van  $S$  wordt bereikt als alle rangnummers een positief teken hebben. Dan geldt  $S = 1+2+\dots+20 = \frac{1}{2} \cdot 20 \cdot (1+20) = 210$ . Onder de nulhypothese is de kans dat een willekeurig rangnummer het teken “plus” heeft gelijk aan  $\frac{1}{2}$ . Dus  $\Pr(S = 210) = (\frac{1}{2})^{20}$ . Vanwege de symmetrie geldt ook  $\Pr(S = -210) = (\frac{1}{2})^{20}$ .

De op één na grootste waarde van  $S$  wordt verkregen als alleen het rangnummer 1 een negatief teken heeft. Er geldt dan  $S = 208$ . Ook hier kunnen we gebruik maken van symmetrie. Dus  $\Pr(S = 208) = \Pr(S = -208) = (\frac{1}{2})^{20}$ .

Als alleen het rangnummer 2 positief c.q. negatief is vinden we:  $\Pr(S = 206) = \Pr(S = -206) = (\frac{1}{2})^{20}$ .

$S$  kan op twee manieren de waarde 204 aannemen: als alleen het rangnummer 3 negatief is of als alleen de rangnummers 1 en 2 negatief zijn. Zo vinden we:  $\Pr(S = 204) = \Pr(S = -204) = 2 \cdot (\frac{1}{2})^{20}$ . Zo doorgaand worden de berekeningen al snel ingewikkelder, bijvoorbeeld  $\Pr(S = 198) = \Pr(S = -198) = 4 \cdot (\frac{1}{2})^{20}$ , met negatieve rangnummers: 6 of (1 en 5) of (2 en 4) of (1, 2 en 3).

Om een toets te kunnen uitvoeren, hoeft je niet de kans op alle mogelijke uitkomsten van  $S$  te berekenen. Als je bijvoorbeeld eenzijdig wilt toetsen met een significantieniveau van 5%, moet je bovenstaande berekeningen voortzetten totdat de optelsom van alle berekende kansen voor het eerst 0,05 of meer is.

## Niet-parametrische toetsen

### Opgave

52. Bereken  $\Pr(S = 192)$ .

In de praktijk is het gelukkig onnodig om dit monnikenwerk met de hand uit te voeren. Er bestaat software die het werk voor je doet en met dergelijke software is de tabel van Appendix 1 opgesteld. Deze tabel geeft voor een aantal gangbare significantieniveaus de kritieke waarde van  $S$  voor waarden van  $n'$  tot en met 30.

In ons voorbeeld geldt  $n' = 20$ . Als we tweezijdig willen toetsen met  $\alpha = 0,05$ , dan zien we in de tabel dat  $\pm 106$  de bijbehorende kritieke waarden van  $S$  zijn. Onze steekproef gaf  $S = 111$ , dus we kunnen de nulhypothese verwerpen.

Bij grotere waarden van  $n'$  kunnen we, zoals zo vaak, gebruik maken van de centrale limietstelling om de berekeningen te vereenvoudigen. Zoals gezegd,  $E(S) = 0$  en verder kun je berekenen (maar dat zullen we hier achterwege laten) dat  $\text{Var}(S) = \frac{1}{6}n'(n' + 1)(2n' + 1)$ .

Bij grote waarden van  $n'$  zal  $Z = S / \sqrt{\frac{1}{6}n'(n' + 1)(2n' + 1)}$  bij benadering standaardnormaal verdeeld zijn. In de praktijk wordt  $n' > 30$  als voorwaarde voor deze benadering aangehouden.

Als we in dit voorbeeld gebruik maken van de benadering via de normale verdeling (bij een  $n'$  van 20 is dat niet nodig omdat we de exacte grenzen in de tabel kunnen vinden, maar we doen het hier bij wijze van illustratie), dan vinden we

$$Z = S / \sqrt{\frac{1}{6}n'(n' + 1)(2n' + 1)} = 111 / \sqrt{\frac{1}{6} \cdot 20 \cdot 21 \cdot 41} \approx 2,07$$

Voor de standaardnormale verdeling geldt:  $\Pr(|Z| \geq 2,07) \approx 2 \cdot \text{normalcdf}(2,07, 10^99) = 0,04$ , dus ook via deze benadering zouden we de hypothese verwerpen bij  $\alpha = 0,05$ . Volgens deze benadering is de rechter kritieke waarde de uitkomst van  $S$  waarvoor geldt  $Z = \text{invNorm}(0,975) = 1,96$ .

Deze kritieke waarde is  $1,96 \cdot \sqrt{\frac{1}{6} \cdot 20 \cdot 21 \cdot 41} \approx 105$  en deze verschilt niet veel van de waarde die we via de tabel vonden (106).

### Opgaven

53. Geef in bovenstaand voorbeeld de kritieke waarde van  $S$  als we eenzijdig toetsen met  $\alpha = 0,05$ . Gebruik de tabel.

54. In de tabel van Appendix 1 is een aantal vakjes gevuld met streepjes. Verklaar deze situatie met behulp van een berekening.

#### 4.2. Gepaarde waarnemingen: Wilcoxon rangteken toets

55. In de literatuur (en ook in de output van veel statistische software) worden in plaats van  $S$  ook de toetsingsgrootheden  $T_+$  en  $T_-$  gebruikt.  $T_+$  is de som van alle rangnummers die bij positieve verschillen horen en  $T_-$  de som van alle (positieve) rangnummers die bij negatieve verschillen horen. Er geldt dus  $S = T_+ - T_-$ .

- a. Bereken  $E(T_+)$  en  $E(T_-)$  onder de veronderstelling dat de nulhypothese waar is.
- b. Druk  $T_-$  uit in  $T_+$ , zonder in deze uitdrukking  $S$  te gebruiken.
- c. Gebruik het antwoord van b. en de formule voor  $\text{Var}(S)$  om  $\text{Var}(T_+)$  en  $\text{Var}(T_-)$  te berekenen onder  $H_0$ .

**Opmerking 4.2.2.** Bij de berekening van de kritische grenzen in Appendix 1 is (net als in de berekening van opmerking 4.2.1) verondersteld dat de waarnemingen geen knopen bevatten (dus geen waarnemingen met hetzelfde verschil). Als er wel knopen voorkomen, moet je eigenlijk een andere berekening uitvoeren. Bij een klein aantal knopen zijn de verschillen verwaarloosbaar. Bij een groot aantal knopen zijn de grenzen uit Appendix 1 conservatief (dus groter dan strikt noodzakelijk).

**Opmerking 4.2.3.** Als de verdeling van  $S$  wordt benaderd via de normale verdeling, is het theoretisch juist om de continuïteitscorrectie toe te passen. In de formule voor  $Z$  moeten we dan in de teller de *absolute waarde van  $S$*  verminderen met 1, omdat de verschillen tussen opeenvolgende mogelijke uitkomsten van  $S$  steeds gelijk aan 2 zijn. Voor  $n' > 30$  maakt dit echter nauwelijks iets uit.

**Opmerking 4.2.4.** In de paragrafen 4.1 en 4.2 hebben we de tekentoets en Wilcoxon rangtekentoets losgelaten op dezelfde waarnemingen. Bij de tekentoets konden we de nulhypothese niet verwerpen, bij de rangtekentoets wel. In dit geval is het onderscheidingsvermogen van de rangtekentoets groter dan die van de tekentoets. Dit hoeft niet altijd het geval te zijn. Een en ander is afhankelijk van de werkelijke verdeling van de betreffende populaties. In veruit de meeste gevallen uit de praktijk verdient de rangtekentoets de voorkeur. De tekentoets wordt vooral gebruikt omdat het rekenwerk zo eenvoudig is.

## Niet-parametrische toetsen

Samengevat:

Bij het vergelijken van twee populaties met gepaarde waarnemingen ge-  
niet *Wilcoxon's rangtekentoets* (*signed rank test*) in de meeste gevallen de  
voorkeur boven de tekentoets.

Wilcoxon's rangtekentoets werkt als volgt. (1) Bereken de verschillen van  
de  $n$  gepaarde waarnemingen. (2) Geef een rangnummer aan de absolute  
waarde van de verschillen die ongelijk aan 0 zijn. Het aantal van deze  
verschillen noemen we  $n'$ . (3) Voorzie elk rangnummer dat hoort bij een  
negatief verschil van een negatief teken. (4) Bepaal de som  $S$  van de aldus  
van een teken voorziene rangnummers.

Als  $n' \leq 30$ : (5a) Verwerp de nulhypothese dat beide populaties dezelfde  
verdeling hebben als de absolute waarde van  $S$  groter is dan de kritieke  
waarden uit de appendix.

Als  $n' > 30$ : (5b) Onder de nulhypothese dat beide populaties dezelfde  
verdeling hebben is  $Z = S / \sqrt{\frac{1}{6}n'(n'+1)(2n'+1)}$  bij benadering stan-  
daardnormaal verdeeld.

## Opgaven

56. Een statisticus analyseert 37 gepaarde waarnemingen. Hij wil de nul-  
hypothese “de waarnemingen zijn afkomstig uit dezelfde verdeling”  
toetsen tegen het tweezijdig alternatief. Hij constateert dat de 37  
verschillen van deze waarnemingen (waarvan er 2 gelijk aan 0 zijn)  
zeer scheef verdeeld zijn, daarom besluit hij geen t-toets te gebruiken,  
maar Wilcoxon's rangtekentoets. Hij vindt  $S = 316$ . Geef de p-waarde  
van deze toets (uiteraard met continuïteitscorrectie).
57. Door een nieuw onderhoudsprotocol wil men het aantal storingen in  
elektriciteitscentrales verminderen. In 15 centrales heeft men een jaar  
vóór en een jaar ná de ingebruikneming van het nieuwe protocol het  
aantal storingen bijgehouden. De resultaten staan hieronder.

Aantal storingen in 15 elektriciteitscentrales, voor en na het nieuwe  
onderhoudsprotocol

Centrale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
V(oor)	3	6	4	2	1	3	2	5	2	0	1	3	4	2	3
N(a)	0	2	1	2	2	0	3	1	0	1	0	1	2	4	0

Ga na of je, bij  $\alpha = 0,05$ , kunt zeggen dat het aantal storingen signi-  
ficant is afgenomen.

- a. Gebruik de tekentoets. Wat is de p-waarde?

### 4.3. Ongepaarde waarnemingen: De toets van Wilcoxon

- b. Gebruik de Wilcoxon rangtekentoets. Wat kun je zeggen over de p-waarde?
  - c. Welke “slag om de arm” moet je houden bij het resultaat van b.?
58. Zie het voorbeeld uit paragraaf 3.3. Men vraagt zich af of het drinken van koffie de hartslag verhoogt. Formuleer de relevante hypothese en gebruik Wilcoxons rangtekentoets om deze hypothese te toetsen met  $\alpha = 0,05$ .
59. In opgave 49b heb je de t-toets gebruikt om te toetsen of de scoreverschillen van “mooie” minus “lelijke” foto’s gemiddeld van 0 verschilden. Je had geen sterke rechtvaardiging voor het gebruik van de t-toets. Hoewel de data er “netjes” uit zagen, kunnen scoreverschillen (relatief kleine gehele getallen) vanwege hun discrete karakter niet normaal verdeeld zijn. Bovendien was het aantal waarnemingen (12) gering. Voer daarom de toets opnieuw uit met een verdelingsvrije toets. Vergelijk het resultaat met dat van opgave 49b en geef commentaar.

### 4.3 Ongepaarde waarnemingen. De toets van Wilcoxon (of van Mann-Whitney)

We beginnen opnieuw met een voorbeeld. Een onderzoeker vermoedt dat het beoefenen van yoga leidt tot stressvermindering. Stress kan worden gemeten door het invullen van een formulier. Uit de antwoorden wordt een score afgeleid die een maat is van de hoeveelheid stress. Hoe hoger de score, hoe meer stress. Van 17 proefpersonen wordt de stress gemeten. Door middel van loting worden 8 van deze 17 proefpersonen geselecteerd. Hun wordt opgedragen om gedurende één week dagelijks een aantal yoga-oefeningen te doen. De resterende 9 proefpersonen krijgen geen specifieke instructie. Na een week wordt van alle proefpersonen opnieuw de stress gemeten. De verandering van stress (na minus voor) noemen we  $Y_i$  ( $i = 1, \dots, 8$ ) voor de groep die aan yoga heeft gedaan en  $N_i$  ( $i = 1, \dots, 9$ ) voor de groep die niet aan yoga heeft gedaan. De waarnemingen zijn:

$Y$ : 1,60 -1,85 -1,40 -1,20 -0,95 -0,10 -2,40 -1,15

$N$ : -0,90 -0,75 -1,10 -0,85 -0,50 -0,80 -1,50 1,65 2,85.

We rangschikken de waarnemingen zoals in de tabel hieronder. Eerst rangschikken we de  $Y_i$ 's en de  $N_i$ 's van klein naar groot en vervolgens kennen we aan iedere waarneming een rangnummer toe. Waarnemingen met een gelijke waarde (knopen) krijgen, net als bij Wilcoxons rangtekentoets, het gemiddelde rangnummer, maar in dit voorbeeld komen geen knopen voor.

## Niet-parametrische toetsen

$Y$	rangnummer	$N$	rangnummer
-2,40	1	-1,50	3
-1,85	2	-1,10	7
-1,40	4	-0,90	9
-1,20	5	-0,85	10
-1,15	6	-0,80	11
-0,95	8	-0,75	12
-0,10	14	-0,50	13
1,60	15	1,65	16
		2,85	17

Vervolgens bepalen we de som van de rangnummers van de groep met het kleinste aantal waarnemingen. In dit geval de som van de  $Y_i$ 's :  $S_Y = 55$ .

We formuleren nu de nulhypothese:  $Y$  en  $N$  hebben dezelfde verdeling, d.w.z. het doen van yoga-oefeningen heeft geen invloed op de stressverandering. Onder de nulhypothese heeft elke combinatie van 8 van de rangnummers van 1 t/m 17 dezelfde kans om aan de  $Y_i$ 's te worden toebedeeld. De minimale waarde van  $S_Y$  is  $1 + 2 + \dots + 8 = 36$ . De maximale waarde van  $S_Y$  is  $10 + 11 + \dots + 17 = 108$ . De gemiddelde waarde van de rangnummers is 9. Dus de gemiddelde waarde van  $S_Y$  is  $8 \cdot 9 = 72$ . Als de nulhypothese waar is, zal dus gelden  $E(S_Y) = 72$  en verwachten we een uitkomst van  $S_Y$  die “in de buurt” ligt van 72. Als alternatieve hypothese nemen we “yoga resulteert in minder stress”, dus we gaan eenzijdig toetsen en zullen de nulhypothese verwerpen als  $S_Y$  “klein” is. Om de kritische grens bij een gegeven significantieniveau  $\alpha$  te bepalen, moeten we de kansverdeling van  $S_Y$  onder de nulhypothese kennen. In opmerking 4.3.1 zullen we laten zien hoe deze kansverdeling kan worden berekend. Net als in het geval van Wilcoxon's rangtekentoets wordt deze berekening al snel omslachtig.

**Opmerking 4.3.1.** We hebben hierboven gezien dat de minimale waarde van  $S_Y$  gelijk is aan 36. De gebeurtenis  $S_Y = 36$  treedt alleen maar op als de rangnummers 1 t/m 8 alle in de kolom “Y-kolom” terecht komen. Dus  $\Pr(S_Y = 36) = 1 / \binom{17}{8}$ .

De gebeurtenis  $S_Y = 37$  treedt alleen maar op als de rangnummers 1 t/m 7 en 9 in de kolom “Y-kolom” terecht komen. Dus ook  $\Pr(S_Y = 37) = 1 / \binom{17}{8}$ .

De gebeurtenis  $S_Y = 38$  kan zich op twee manieren voordoen: in de “Y-kolom” komen de rangnummers (1 t/m 7 en 10) of (1 t/m 6 en 8 en 9). Dus  $\Pr(S_Y = 38) = 2 / \binom{17}{8}$  enzovoort.

### 4.3. Ongepaarde waarnemingen: De toets van Wilcoxon

#### Opgave

60. Bereken  $\Pr(S_Y = 40)$ .

In de praktijk gebruik je statistische software om deze berekeningen voor je uit te voeren. Je kunt ook een tabel gebruiken zoals in Appendix 2. In deze tabel worden de twee te vergelijken populaties  $X_1$  (met  $n_1$  waarnemingen) en  $X_2$  (met  $n_2$  waarnemingen) genoemd, met  $n_1 \leq n_2$ . In ons voorbeeld geldt  $n_1 = 8$ ,  $n_2 = 9$  en  $S_{X_1} = 55$ . We willen eenzijdig toetsen met  $\alpha = 0,05$ . Omdat de tabel uitgaat van tweezijdig toetsen kijken we onder  $\alpha = 0,1$  en lezen we de kritische waarde 54 af, dus we kunnen de nulhypothese niet verwerpen.

Bij grotere aantallen waarnemingen dan in de tabel zijn opgenomen, geeft een benadering via de normale verdeling bevredigende resultaten. We gebruiken daarbij dat  $S_{X_1}$  bij benadering normaal verdeeld is met:

$$E(S_{X_1}) = \frac{1}{2}n_1(1 + n_1 + n_2)$$

en

$$\text{Var}(S_{X_1}) = \frac{1}{12}n_1n_2(1 + n_1 + n_2).$$

Ter illustratie laten we zien welk resultaat deze benadering zou geven in ons voorbeeld. We krijgen  $E(Y) = \frac{1}{2} \cdot 8 \cdot 18 = 72$  en  $\text{Var}(Y) = \frac{1}{12} \cdot 8 \cdot 9 \cdot 18 = 108$ . Met toepassing van de continuïteitscorrectie geeft dit  $\Pr(Y \leq 55) \approx \text{normalcdf}(-10^99, 55.5, 72, \sqrt{108}) = 0,056$ . Ook via deze benadering zouden we de nulhypothese niet verwerpen. De kritische grens kan als volgt worden berekend:  $\text{invNorm}(0.05, 72, \sqrt{108}) = 54,9$ , dus is 54 de grens (net als in de tabel).

#### Opgaven

61. Toon aan dat  $E(S_{X_1}) = \frac{1}{2}n_1(1 + n_1 + n_2)$

62. Verklaar d.m.v. een berekening waarom in de tabel van Appendix 2 het vakje bij  $n_1 = 2$ ,  $n_2 = 7$  en  $\alpha = 0,05$  leeg is.

**Opmerking 4.3.2.** Bij de berekening van de kritische grenzen in Appendix 2 is (net als bij Appendix 1) verondersteld dat de waarnemingen geen knopen bevatten. Als er wel knopen voorkomen, moet je een andere berekening uitvoeren. Bij een klein aantal knopen zijn de verschillen verwaarloosbaar. Bij een groot aantal knopen zijn de grenzen uit Appendix 2 conservatief, d.w.z. de rechtergrenzen zijn groter en de linkergrenzen kleiner dan strikt noodzakelijk.

## Niet-parametrische toetsen

Samengevat:

$n_1$  en  $n_2$  ( $n_1 \leq n_2$ ) ongepaarde waarnemingen van twee onafhankelijke kansvariabelen  $X_1$  en  $X_2$  kunnen als volgt worden geanalyseerd met de *toets van Wilcoxon (Mann-Whitney)*. (1) Rangschik de waarnemingen uit beide steekproeven en geef iedere waarneming een rangnummer. (2) Bepaal de som  $S_{X_1}$  van de rangnummers van de waarnemingen van  $X_1$ . (3) De nulhypothese “ $X_1$  en  $X_2$  hebben dezelfde verdeling” kan worden getoetst met behulp van de uitkomst van  $S_{X_1}$ . (4a) Bij kleine waarden van  $n_1$  en  $n_2$  gebruik je Appendix 2 (4b) Bij grote waarden van  $n_1$  en  $n_2$  gebruik je dat  $S_{X_1}$  bij benadering normaal verdeeld is met  $E(S_{X_1}) = \frac{1}{2}n_1(1 + n_1 + n_2)$  en  $Var(S_{X_1}) = \frac{1}{12}n_1n_2(1 + n_1 + n_2)$ .

## Opgaven

63. Je analyseert de gegevens uit twee onafhankelijke steekproeven  $A$  (met 12 waarnemingen) en  $B$  (met 7 waarnemingen). Je wilt een tweezijdige toets van Wilcoxon uitvoeren en je vindt  $S_B = 49$ . Bepaal de p-waarde van deze toets.
64. Maaïke onderzoekt de effectiviteit van kunstmest op de groei van een bepaald soort plant. De fabrikant van de kunstmest beweert dat de planten groter worden bij het gebruik van kunstmest. Uit één partij zaaigoed kweekt Maaïke een aantal planten in twee aparte bakken. In bak  $K$  wordt kunstmest gebruikt, in bak  $N$  niet. Verder houdt ze de omstandigheden voor beide bakken zoveel mogelijk gelijk. Na 30 dagen meet ze de hoogte (in cm) van de plantjes. Ze vindt:  
 $K$ : 10,3 11,8 9,8 12,6 11,4 8,3 12,0 10,9  
 $N$ : 7,2 9,5 11,2 8,0 6,9 10,1 9,8  
Hoewel het a priori niet vreemd is om te veronderstellen dat de lengte van plantjes normaal verdeeld is, gebruikt Maaïke voor de zekerheid een verdelingsvrije methode. Kan ze, met  $\alpha = 0,05$ , concluderen dat kunstmest de groei bevordert? Wat kun je zeggen over de p-waarde van de toets?
65. Rik onderzoekt of de prestaties op de 60m sprint bij jongens van groep 7 op de basisschool kunnen worden beïnvloed door het gebruik van een placebo. Daarvoor verdeelt hij de jongens uit één klas d.m.v. het lot in twee groepen  $P$  en  $N$ . De jongens uit beide groepen moeten, onafhankelijk van elkaar, 60m hardlopen waarbij hun tijd wordt geregistreerd. De jongens uit groep  $P$  krijgen voor ze gaan lopen een pilletje met zoetstof, met de mededeling dat het sterk prestatieverhogend werkt. De jongens uit groep  $N$  krijgen niets. De resultaten (in seconden) zijn:  
 $P$ : 12,3 10,8 11,5 11,0 10,7 12,0  
 $N$ : 12,7 13,0 10,9 11,2 13,2 13,4 12,5



### 4.3. Ongepaarde waarnemingen: De toets van Wilcoxon

- a. Bereken de gemiddelde tijd per groep.
  - b. Kun je, met  $\alpha = 0,05$  en gebruik makend van een non-parametrische toets, concluderen dat het placebo de prestatie beïnvloedt?
  - c. Kun je voor Rik een proefopzet bedenken die naar verwachting sneller tot een significant resultaat zou kunnen leiden?
66. Een wiskundelerares krijgt van de schoolleiding twee vragen: (i) Is wiskunde D moeilijker dan wiskunde B, of makkelijker? en (ii) Halen leerlingen die wiskunde D volgen andere cijfers voor wiskunde B dan leerlingen die geen wiskunde D volgen? Om deze vragen te beantwoorden maakt de wiskundelerares een lijstje van haar wiskunde-D-leerlingen en zet er hun cijfers voor wiskunde B en wiskunde D bij. Deze cijfers vind je in onderstaande tabel onder het kopje BmD (B met D) en D. Vervolgens neemt ze een steekproef van eveneens 12 leerlingen die wel wiskunde B, maar geen wiskunde D volgen. Hun cijfers vind je onder het kopje BzD (B zonder D).

<b>BmD</b>	<b>D</b>	<b>BzD</b>
9,0	8,5	6,3
7,0	6,2	7,3
6,0	4,9	6,9
8,5	8,5	7,2
9,7	8,9	7,5
8,0	8,8	7,3
8,1	7,6	6,5
8,4	7,3	6,2
8,4	8,0	7,4
6,8	5,6	4,9
9,6	9,2	9,2
6,1	6,1	5,8

- a. Bereken voor de drie groepen cijfers het gemiddelde.
- b. Toets met  $\alpha = 0,05$  of leerlingen die beide vakken volgen significant verschillend scoren voor beide vakken. Je mag niet aannemen dat de cijfers normaal verdeeld zijn. Gebruik een zo eenvoudig mogelijke toets.
- c. Toets met  $\alpha = 0,05$  of leerlingen met wiskunde D significant anders scoren voor wiskunde B dan leerlingen zonder wiskunde D.
- d. Vergelijk de uitkomsten van a) , b) en c) en geef commentaar.

#### 4.4 Eén populatie: toetsen betreffende de mediaan

In de vorige paragrafen behandelden we niet-parametrische toetsen die als alternatief konden dienen voor een t-toets als de onderliggende populatie niet normaal verdeeld is. We vergeleken twee populaties met gepaarde en ongepaarde waarnemingen.

Tot dusver bleef een alternatief voor de t-toets over de waarde van het gemiddelde van één verdeling onbesproken. Dat is niet verwonderlijk, want het begrip ‘gemiddelde’, dat bij de normale verdeling samenvalt met de parameter  $\mu$ , is moeilijk hanteerbaar in een niet-parametrische context, waar de oorspronkelijke gegevens worden teruggebracht tot rangnummers. Bij de vergelijking van twee populaties konden we dit probleem omzeilen door de nulhypothese “beide populaties hebben dezelfde  $\mu$ ” te vervangen door “beide populaties hebben dezelfde verdeling”.

Als we te maken hebben met één enkele populatie is deze uitweg afgesloten. We kunnen alleen een niet-parametrische toets uitvoeren over de **mediaan** van de verdeling i.p.v. over het gemiddelde. Bij symmetrische verdelingen valt de mediaan samen met het gemiddelde, maar hier hebben we in de praktijk niet zoveel aan, omdat bij een symmetrische verdeling al snel de t-toets kan worden gebruikt. We hebben de niet-parametrische methodes juist nodig als de onderliggende verdeling niet symmetrisch is.

We hebben dan te maken met een steekproef uit een populatie  $X$  met een onbekende verdeling. De mediaan van deze verdeling noemen we  $m$ . We willen de hypothese  $H_0 : m = m_0$  toetsen, waarbij  $m_0$  een vooraf bekende waarde heeft, tegen een één- of tweezijdig alternatief. We gaan daarbij als volgt te werk. We elimineren de waarnemingen die precies gelijk zijn aan  $m_0$ . Voor de overige waarnemingen geldt dat de kansvariabele  $X - m_0$  met kans  $\frac{1}{2}$  een positieve waarde aanneemt en met kans  $\frac{1}{2}$  een negatieve waarde. Op deze grootte ( $X - m_0$ ) kunnen we nu de tekentoets of Wilcoxon's rangtekentoets loslaten.

## Hoofdstuk 5

# Het vergelijken van populatiepercentages

In de vorige twee hoofdstukken hebben we het gehad over de hoogte van de hartslag, het aantal verkeersongelukken, het vetgehalte van hotdogs etc. Dit zijn steeds numerieke variabelen. De uitkomst is een getal, je kunt van de uitkomsten het gemiddelde bepalen en de standaardfout, je kunt de uitkomsten op volgorde zetten en er een rangnummer aan geven.

Als je in aanraking komt met categorische variabelen is de situatie anders. Een behandeling heeft wel of niet het gewenste effect, een scholier rookt wel of niet, een auto is een Renault, een Toyota of een Fiat. Van dit soort uitkomsten is het gemiddelde zinloos en je kunt ze ook niet rangschikken. We concentreren ons dan meestal op de vraag: “hoe vaak komt een bepaald type uitkomst voor?” Zo krijg je te maken met populatiepercentages en gaat de binomiale verdeling een rol spelen. In hoofdstuk twee heb je geleerd hoe je een betrouwbaarheidsinterval voor een populatiepercentage kunt opstellen. Eerder maakte je kennis met de binomiaaltoets. In dit hoofdstuk breiden we deze technieken uit tot het vergelijken van twee populatiepercentages.

### 5.1 Vergelijking van twee populatiepercentages (ongepaarde waarnemingen, grote aantallen)

We beginnen weer met een voorbeeld.

Sommige mensen denken dat je van alles kunt repareren met kleeftape. Je kunt er buizen mee aan elkaar verbinden, elektrische leidingen mee isoleren enzovoorts. Maar kun je er ook wratten mee genezen? Sommige mensen denken van wel. 204 proefpersonen die zich voor hun wratten wilden laten behandelen werden aselekt in twee groepen verdeeld: 100 werden op de gebruikelijke wijze met vloeibaar stikstof behandeld en bij de overige 104 werden de wratten afgeplakt met tape. Na zes dagen werd de tape verwijderd, de wrat geweekt met water en vervolgens afgekrabd. Zowel

### Het vergelijken van populatiepercentages

bij de behandeling met vloeibaar stikstof als met kleeftape werd de behandeling zonodig herhaald. Na twee maanden waren de resultaten als volgt.

Behandeling	$n$	Aantal proefpersonen bij wie de wratten zijn verdwenen
Vloeibaar stikstof	100	60
Kleeftape	104	88

We hebben te maken met twee populaties, gekenmerkt door de behandelingsmethode. De groep die behandeld is met vloeibaar stikstof geven we de index 1, de groep die met kleeftape is behandeld krijgt de index 2. De variabele is categorisch: de wratten zijn wel of niet verdwenen. Het aantal proefpersonen bij wie de wratten zijn verdwenen noemen we  $X_1$  respectievelijk  $X_2$ , afhankelijk van de behandelingsmethode. Deze kansvariabelen zijn binomiaal verdeeld met parameters  $n_1 = 100$  respectievelijk  $n_2 = 104$  en onbekende kansen op succes  $\pi_1$  respectievelijk  $\pi_2$ .

We toetsen de hypothese  $H_0 : \pi_1 = \pi_2$  tegen het alternatief  $H_1 : \pi_1 \neq \pi_2$ . De steekproefuitkomst levert ons een schatting van beide parameters:

$p_1 = X_1/n_1 = 0,6$  en  $p_2 = X_2/n_2 \approx 0,8462$ . Onder de nulhypothese geldt dat beide onbekende parameters een gemeenschappelijke waarde  $\pi$  hebben en de beste schatting van deze parameter maakt gebruik van alle waarnemingen:

$$p = (X_1 + X_2)/(n_1 + n_2) = 148/204 \approx 0,7255.$$

Het ligt voor de hand om onze toetsingsgrootte te baseren op de waarde van  $p_1 - p_2$ . Als de uitkomst van deze toetsingsgrootte te veel van 0 afwijkt, zullen we geneigd zijn de nulhypothese te verwerpen. De verdeling van  $p_1 - p_2$  is eenvoudig te bepalen als we gebruik maken van de benadering via de normale verdeling.

In opgave 67 ga je aantonen dat onder de nulhypothese

$Z = (p_1 - p_2)/\sqrt{p(1-p)(n_1^{-1} + n_2^{-1})}$  bij benadering standaardnormaal verdeeld is. Dit geldt alleen bij “grote” waarden van  $n_1$  en  $n_2$ . In de praktijk wordt hiervoor als voorwaarde gesteld dat  $n_1p$ ,  $n_1(1-p)$ ,  $n_2p$ ,  $n_2(1-p)$  alle vier  $\geq 5$  moeten zijn. In dit voorbeeld is aan deze voorwaarde ruimschoots voldaan. We vinden  $Z \approx -3,938$ . Hierbij hoort een  $p$ -waarde van  $2 \cdot \text{normalcdf}(-10^99, -3,938) = 8,2 \cdot 10^{-5}$ . Er is dus een duidelijk significant verschil tussen  $\pi_1$  en  $\pi_2$ .

### 5.1. Ongepaarde waarnemingen, grote aantallen

Samengevat:

Om  $H_0 : \pi_1 = \pi_2$  te toetsen gebruik je de toetsingsgrootheid

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)(n_1^{-1} + n_2^{-1})}}.$$

Onder  $H_0$  is  $Z$  bij benadering standaardnormaal verdeeld als  $n_1p$ ,  $n_2p$ ,  $n_1(1-p)$  en  $n_2(1-p)$  alle vier  $\geq 5$  zijn.

**Opmerking 5.1.1.** De GR kent een standaardfunctie voor het uitvoeren van deze toets: `2-PropZTest`, die je kunt vinden via de menus `STAT - TESTS`. Aan de aanduiding  $Z$  in de naam kun je zien dat het een benadering via de normale verdeling betreft. In het voorbeeld van deze paragraaf is de invoer als volgt:

`x1:60 n1:100 x2:88 n2:104 p1:≠ p2. Calculate` geeft de waarde van  $Z = -3,938$  en de  $p$ -waarde  $8,2 \cdot 10^{-5}$ , alsmede enkele samenvattende cijfers.

**Opmerking 5.1.2.** De hierboven beschreven toets komt overeen met wat in de statistische literatuur (en in veel statistische softwarepakketten) de *chi-kwadraattoets* (Engels: *chi-square test*) wordt genoemd. De chi-kwadraattoets is gebaseerd op een toetsingsgrootheid die overeenkomt met  $Z^2$  en die onder de nulhypothese bij benadering een chi-kwadraatverdeling met 1 vrijheidsgraad heeft. De chi-kwadraattoets kan eenvoudig worden uitgebreid naar het geval waarin meer dan twee populatiepercentages met elkaar worden vergeleken. Dit onderwerp (*contingency tables*) wordt in deze module niet behandeld. Dus: als je in een softwarepakket op zoek bent naar een functie waarmee je twee (of meer) populatiepercentages wilt vergelijken, zul je deze dikwijls kunnen vinden onder de titel “contingency tables” of “chi-square tests”.

### Opgaven

67. a. Bereken onder  $H_0$ :  $E(p_1)$ ,  $E(p_2)$ ,  $Var(p_1)$  en  $Var(p_2)$ .  
b. Zijn  $p_1$  en  $p_2$  onafhankelijk? Bereken  $E(p_1 - p_2)$  en  $Var(p_1 - p_2)$  onder  $H_0$ .  
c. Laat zien hoe je kunt concluderen dat  $Z$  (bij benadering) standaardnormaal verdeeld is onder  $H_0$ .
68. Voor een cursus Engels kun je je aanmelden via de website van de organisator van de cursus. Wegens klachten over de duidelijkheid van

## Het vergelijken van populatiepercentages

de site roept de directie een tweede mogelijkheid om je aan te melden in het leven: telefonisch. Na verloop van tijd wordt aan 80 (aselect gekozen) cursisten die zich via de website hebben aangemeld gevraagd of ze tevreden zijn met de aanmeldingsprocedure: 59 zijn tevreden. Aan 60 (eveneens aselect gekozen) cursisten die zich telefonisch hebben aangemeld wordt dezelfde vraag gesteld: 48 zijn tevreden. Kun je op grond van deze gegevens bij  $\alpha = 0,05$  concluderen dat de tevredenheid onder telefonische aanmelders hoger is? Maak een berekening met en zonder gebruik te maken van de statistische functies van de GR.

69. In de tabel hieronder zie je hoe het ledenbestand van de leerlingenvereniging van een school is verdeeld tussen onder- en bovenbouw. Zijn de verschillen significant? Gebruik  $\alpha = 0,05$ .

	Lid van de leerlingenvereniging	
	ja	nee
Onderbouw	185	86
Bovenbouw	152	90

## 5.2 Vergelijking van twee populatiepercentages (ongepaarde waarnemingen, kleine aantallen): Fishers exacte toets

Stel dat we het in paragraaf 5.1 beschreven onderzoek naar de effecten van wrattenbestrijding hadden uitgevoerd met minder proefpersonen en daarbij de volgende resultaten hadden gekregen.

Behandeling	$n$	Aantal proefpersonen bij wie de wratten zijn verdwenen
Vloeibaar stikstof	9	2
Kleeftape	12	8

Net als in 5.1 noemen we de kans dat de wratten van iemand die zich met vloeibaar stikstof respectievelijk kleeftape laat behandelen  $\pi_1$  respectievelijk  $\pi_2$ . We toetsen  $H_0 : \pi_1 = \pi_2$  tegen  $H_1 : \pi_1 \neq \pi_2$  met  $\alpha = 0,05$ . De steekproef suggereert dat  $\pi_1 < \pi_2$ , want  $2/9 < 8/12$ . In deze situatie geldt:  $n_1 p = \frac{9 \cdot 10}{21} \approx 4,3 < 5$ , dus we mogen geen gebruik maken van de benadering via de normale verdeling.

In totaal hebben we te maken met 21 proefpersonen. Bij 10 van hen verdwenen de wratten, bij 11 van hen niet. Onder  $H_0$  (d.w.z. als de beide behandelmethodes dezelfde kans op succes hebben) zijn de 9 proefpersonen die zich hebben laten behandelen met vloeibaar stikstof op te vatten als

## 5.2. Fishers exacte toets

een *willekeurige* deelverzameling van deze 21 proefpersonen. Het toetsen van de nulhypothese komt neer op het beantwoorden van de vraag: is twee succesvolle behandelingen in deze willekeurige deelverzameling een “onwaarschijnlijk laag” aantal?

We kunnen het vaasmodel gebruiken om deze situatie te analyseren. Gegeven is een vaas met 21 knikkers (proefpersonen): 10 witte en 11 rode (bij wie de wratten wel respectievelijk niet verdwenen). Uit deze vaas nemen we (zonder terugleggen) 9 willekeurige knikkers (de proefpersonen die met vloeibaar stikstof zijn behandeld). De overschrijdingskans die bij de nulhypothese hoort is gelijk aan de kans op 2 of minder witte knikkers. Deze kans is gelijk aan :  $\Pr(0 \text{ wit, } 9 \text{ rood}) + \Pr(1 \text{ wit, } 8 \text{ rood}) + \Pr(2 \text{ wit, } 7 \text{ rood}) =$

$$\frac{\binom{10}{0} \binom{11}{9} + \binom{10}{1} \binom{11}{8} + \binom{10}{2} \binom{11}{7}}{\binom{21}{9}} = \frac{1 \cdot 55 + 10 \cdot 165 + 45 \cdot 330}{293930} \approx 0,0563$$

We toetsen tweezijdig, dus de  $p$ -waarde van deze toets is gelijk aan  $2 \cdot 0,0563 \approx 0,113$ . De  $p$ -waarde is groter dan  $\alpha (= 0,05)$ , dus we kunnen de nulhypothese niet verwerpen. De twee behandelmethoden leiden niet tot significant verschillende aantallen verdwenen wratten.

Samenvattend:

Om bij kleine aantallen  $H_0 : \pi_1 = \pi_2$  te toetsen gebruiken we Fishers exacte toets. Deze toets gaat uit van het vaasmodel, waarbij de vaas is samengesteld uit de gecombineerde steekproeven uit beide populaties. Onder  $H_0$  is de uitkomst van de steekproef uit één van de twee populaties op te vatten als een willekeurige greep uit de vaas.

**Opmerking 5.2.1.** De GR kent geen standaardfunctie voor Fishers exacte toets. Als we deze toets alleen gebruiken bij kleine aantallen, zijn de berekeningen goed met de GR uit te voeren. Je moet je daarbij richten op de kleinste waarneming. Voor de berekening van  $\binom{n}{p}$  gebruik je de functie `nCr` die je kunt vinden via `MATH` en `PRB`.

### Opgaven

70. Een medewerker van de GGD doet een onderzoek(je) naar de toestand van het gebit van leerlingen uit groep 8 van een basisschool. Hij constateert: van de 24 allochtone leerlingen hebben er 2 één of meer vullingen, van de 102 autochtone leerlingen hebben er 24 één of meer vullingen. Is er een significant verschil in de gebitstoestand tussen beide groepen leerlingen? Neem  $\alpha = 0,05$ .
71. Men vraagt zich af of voetballers (die regelmatig koppen) vaker een hersenschudding hebben dan andere sporters. Daarom wordt een steekproef gehouden onder mannelijke sportende studenten. Hun wordt gevraagd of ze wel eens een hersenschudding hebben gehad en of ze aan voetbal doen. De resultaten zie je in de tabel hieronder. Kun je op grond van deze steekproef concluderen dat voetballers vaker een hersenschudding hebben dan andere sporters? Neem  $\alpha = 0,01$ .

	Aantal sporters	
	zonder hersenschudding	met minstens 1 hersenschudding
Voetbal	46	45
Andere sport	68	28

72. Een onderzoeker vraagt zich af of mensen die lid zijn van een partij die milieubewustzijn hoog in het vaandel heeft staan, hun politieke overtuiging ook persoonlijk in de praktijk brengen. Hij neemt een steekproef uit leden van de PvdA en vraagt hun of ze een auto bezitten: 55 hebben een auto en 4 niet. Hij doet hetzelfde met leden van GroenLinks: 30 hebben een auto en 7 niet. Bij beide steekproeven neemt hij alleen mensen in de steekproef op als hun jaarinkomen 75.000 euro of meer is.
- Waarom, denk je, hanteert de onderzoeker deze inkomensgrens?
  - Verschillen de gevonden aantallen significant van elkaar? Neem  $\alpha = 0,05$ .
  - Vind je dat de onderzoeker, gegeven de hoofdvraag, zijn onderzoek goed heeft opgezet? Geef commentaar.

### 5.3 Vergelijking van twee populatiepercentages: betrouwbaarheidsinterval bij grote aantallen

De in paragraaf 5.1 beschreven toets berust op het feit dat onder  $H_0 : \pi_1 - \pi_2 = 0$  de kansvariabele  $Z = (p_1 - p_2) / \sqrt{p(1-p)(n_1^{-1} + n_2^{-1})}$  bij benadering standaardnormaal verdeeld is. Deze toets heeft een nadeel:



### 5.3. Betrouwbaarheidsinterval bij grote aantallen

als we  $H_0$  kunnen verwerpen, geeft de toets ons geen informatie over de grootte van  $\pi_1 - \pi_2$  in de vorm van een betrouwbaarheidsinterval. Dit wordt veroorzaakt door het feit dat, als  $H_0$  niet waar is, de verdeling van  $Z$  lastig te bepalen valt en afhankelijk is van de individuele waarden van  $\pi_1$  en  $\pi_2$  en niet alleen van het verschil  $\pi_1 - \pi_2$ . In deze paragraaf ontwikkelen we een toetsingsgrootte die dit bezwaar ondervangt.

#### Opgave

73. Als  $n_1$  en  $n_2$  groot zijn, zijn  $p_1$  en  $p_2$  bij benadering normaal verdeeld.

- Geef de verwachting en de variantie van  $p_1$  en  $p_2$ .
- Wat is (bij benadering) de verdeling van  $p_1 - p_2$ ?

Als we in de in opgave 73b gevonden variantie  $\pi_1$  en  $\pi_2$  vervangen door hun schatters  $p_1$  en  $p_2$ , dan krijgen we:  $p_1 - p_2$  is bij benadering normaal verdeeld met verwachting  $\pi_1 - \pi_2$  en standaardafwijking  $\sigma(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ . In opgave 74 ga je dit resultaat gebruiken om een betrouwbaarheidsinterval te maken.

**Opmerking 5.3.1.** De GR kent een standaardfunctie voor de berekening van het hierboven beschreven betrouwbaarheidsinterval: 2-PropZInt, te vinden via STAT - TESTS. De invoer is vergelijkbaar met die van 2-PropZTest.

#### Opgave

74. Geef, in het voorbeeld van paragraaf 5.1 (stikstof versus kleefteape) een 95%-betrouwbaarheidsinterval voor  $\pi_1 - \pi_2$ . Maak een handmatige berekening en controleer deze met de GR.

In het antwoord van opgave 74 zie je dat 0 niet in het 95%-betrouwbaarheidsinterval ligt. Hieruit zou je direct kunnen concluderen dat je bij het toetsen van  $H_0 : \pi_1 = \pi_2$  tegen  $H_1 : \pi_1 \neq \pi_2$  met  $\alpha = 0,05$  de nulhypothese kunt verwerpen. Deze manier van toetsen komt erop neer dat als toetsingsgrootte wordt gehanteerd:

$$Z' = \frac{p_1 - p_2}{\sigma(p_1 - p_2)}.$$

Onder  $H_0$  is  $Z'$  (bij benadering) standaardnormaal verdeeld. Deze  $Z'$  lijkt veel op  $Z$  uit paragraaf 5.1, maar is niet precies hetzelfde. (In ons voorbeeld uit paragraaf 5.1 geldt  $Z' \approx -4,07$ , terwijl  $Z \approx -3,94$ .) We hebben dus een nieuwe toets gemaakt, die als voordeel heeft dat hij tevens rechtstreeks gekoppeld is aan de constructie van een betrouwbaarheidsinterval.

## Het vergelijken van populatiepercentages

**Opmerking 5.3.2.** Beide toetsen (gebaseerd op  $Z$  respectievelijk  $Z'$ ) zijn gebaseerd op een verdeling die bij benadering normaal is. Als  $H_0$  waar is, is de benadering bij de op  $Z$  gebaseerde toets beter dan bij de toets die op  $Z'$  gebaseerd is. De op  $Z$  gebaseerde toets geniet daarom de voorkeur. Je kunt bewijzen dat deze toets conservatiever is dan de  $Z'$ -toets, d.w.z. dat hij minder gemakkelijk de nulhypothese verworpt. In veel statistische software kom je beide varianten tegen. De GR kent alleen de op  $Z$  gebaseerde toets. (2-PropZTest).

Samengevat:

Een betrouwbaarheidsinterval voor  $\pi_1 - \pi_2$  wordt bij benadering gegeven door:  $p_1 - p_2 \pm z_\alpha \cdot \sigma(p_1 - p_2)$ .

$$\text{Hierin is } \sigma(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

De benadering is acceptabel als  $n_1p_1$ ,  $n_2p_2$ ,  $n_1(1-p_1)$  en  $n_2(1-p_2)$  alle vier  $\geq 5$  zijn.

## Opgaven

75. Om het effect van geïsoleerd water op de preventie van cariës te onderzoeken wordt een steekproef genomen van 143 kinderen uit een gemeente zonder fluor in het drinkwater. 106 kinderen hadden tanden met cariës. Uit een gemeente met fluor in het drinkwater werd eveneens een steekproef genomen: 67 van de 119 kinderen hadden last van cariës. Noem  $\pi_1$  de proportie van kinderen met cariës als het drinkwater geen fluor bevat en  $\pi_2$  als het drinkwater wel fluor bevat.
- Geef een 90%- betrouwbaarheidsinterval voor  $\pi_1 - \pi_2$ . Maak een handmatige berekening en controleer je uitkomst met de GR.
  - Verschilt  $\pi_1$  van  $\pi_2$ ? Gebruik  $\alpha = 0,1$  om deze vraag te beantwoorden. Gebruik zowel  $Z$  als  $Z'$ .
  - Hoe veranderen de berekeningen als de vraag bij b. wordt gewijzigd in : “Is  $\pi_1$  groter dan  $\pi_2$ ?”
  - Welke kritiek kun je geven op de aanpak van dit onderzoek?
76. Maak, voor de in opgave 71 beschreven situatie, een 95%-betrouwbaarheidsinterval voor het verschil in de kans op een hersenschudding.

## 5.4 Vergelijking van populatiepercentages bij gepaarde waarnemingen: McNemars toets

Weer beginnen we met een voorbeeld.

Bij 20 patiënten met chronische hoofdpijnlachten worden twee pijnbestrijders  $A$  en  $B$  uitgetoetst. Alle patiënten gebruiken in twee niet te dicht op elkaar liggende periodes het middel  $A$  en  $B$ . Het lot bepaalt welk middel gedurende de eerste periode wordt gebruikt, het andere middel wordt in de tweede periode gebruikt. Elke patiënt geeft per periode aan of het middel wel (+) of niet (-) werkt. De uitkomsten van het experiment zijn als volgt:

patiënt	1	2	3	4	5	6	7	8	9	10
$A$	+	-	-	-	-	-	+	+	-	-
$B$	-	+	-	+	+	+	+	+	-	-
patiënt	11	12	13	14	15	16	17	18	19	20
$A$	-	-	-	-	-	+	-	-	-	-
$B$	-	+	+	+	+	+	+	-	-	+

Net als in de vorige paragrafen noemen we de kans dat middel  $A$  respectievelijk  $B$  bij een willekeurige patiënt werkt  $\pi_1$  respectievelijk  $\pi_2$  en willen we  $H_0 : \pi_1 = \pi_2$  toetsen.

Toch is de situatie hier wezenlijk anders dan in het wrattenvoorbeeld van paragraaf 5.1. Daar werd iedere persoon behandeld op één van de twee manieren. Hier gebruikt iedere persoon beide middelen. Het gaat dus om gepaarde waarnemingen. We kunnen de analyse uit 5.1 en 5.2 hier niet toepassen, want deze is gebaseerd op de onafhankelijkheid van alle waarnemingen en hier zijn de twee waarnemingen bij één persoon afhankelijk.

We moeten dus iets anders verzinnen en dat doen we als volgt. We splitsen de waarnemingen paarsgewijs in het aantal personen dat positief reageerde op beide middelen ( $N_{++}$ ), het aantal personen dat positief reageerde op  $A$ , maar negatief op  $B$  ( $N_{+-}$ ) enzovoorts. In ons voorbeeld krijgen we:  $N_{++} = 3$ ,  $N_{+-} = 1$ ,  $N_{-+} = 10$  en  $N_{--} = 6$ , opgeteld tot  $n = 20$ . Voor onze vraagstelling zijn de getallen  $N_{++}$  en  $N_{--}$  niet interessant. Voor deze personen geven  $A$  en  $B$  hetzelfde resultaat en dat zegt niets over  $\pi_1 - \pi_2$ . Onze aandacht gaat uit naar  $N_{+-}$  en  $N_{-+}$ . We beperken ons dus tot  $n' = 11$  waarnemingen. Onder  $H_0$  is de kans dat een proefpersoon in de categorie  $+-$  valt gelijk aan de kans dat hij in de categorie  $-+$  valt en beide kansen zijn gelijk aan  $\frac{1}{2}$ . Onder  $H_0$  is  $N_{+-}$  dus binomiaal verdeeld met parameters  $n' = 11$  en  $\pi = \frac{1}{2}$ . (Hetzelfde geldt voor  $N_{-+}$ .) Dit resultaat kunnen we gebruiken voor onze toets.

## Het vergelijken van populatiepercentages

**Opmerking 5.4.1.** Bij een voldoende groot aantal waarnemingen kun je situaties als hierboven de binomiale verdeling benaderen met een normale verdeling. De resulterende toets staat bekend onder de naam Mc Nemars toets.  
Het gebruik van de (exacte) binomiaaltoets geniet uiteraard de voorkeur.

Samengevat:

Bij het vergelijken van twee populatiepercentages met gepaarde waarnemingen concentreren we ons op het aantal waarnemingsparen ( $n'$ ) met verschillende uitkomsten, te weten  $N_{+-}$  en  $N_{-+}$ . Onder de nulhypothese van geen verschil in percentages zijn beide kansvariabelen binomiaal verdeeld met parameters  $n'$  en  $\frac{1}{2}$ . Pas vervolgens een binomiaaltoets toe. Mc Nemars toets is een benadering van deze toets die gebruik maakt van de normale verdeling.

## Opgaven

77. Zie bovenstaand voorbeeld. Toets  $H_0$  tegen  $H_1 : \pi_1 \neq \pi_2$  met  $\alpha = 0,05$ .
78. Liesbeth vermoedt dat er meer studenten van witte dan van rode wijn houden. Van 30 proefpersonen zeggen er 15 zowel van rode als van witte wijn te houden, 3 houden helemaal niet van wijn, 3 houden wel van rood en niet van wit en 9 houden niet van rood en wel van wit. Kun je op grond van deze gegevens concluderen dat Liesbeth gelijk heeft? Gebruik  $\alpha = 0,05$ .
79. Meneer Fier is docent Frans aan het Pearson College. Hij is trots op het feit dat in de bovenbouw meer leerlingen voor het vak Frans kiezen dan voor Duits. Zijn collega Stols, docent Duits, vindt dat meneer Fier overdrijft en zegt dat deze situatie net zo goed door toeval zou kunnen zijn veroorzaakt. Meneer Fier weigert dit te geloven en besluit hun meningsverschil voor te leggen aan hun beider wiskundecollega, mevrouw Zeker. Mw. Zeker constateert dat 12 leerlingen zowel Frans als Duits volgen, 45 volgen Frans en geen Duits, 31 volgen Duits en geen Frans en 51 leerlingen volgen geen van beide moderne talen.
- a. Wie krijgt gelijk? Mw. Zeker gebruikt  $\alpha = 0,05$ .

Meneer Fier beweert ook dat bovenbouwleerlingen van het Pearson College significant vaker Frans kiezen dan bovenbouwleerlingen van het Fisher Gymnasium, waar 59 van de 190 bovenbouwleerlingen Frans volgen.

#### 5.4. McNemars toets

- b. Ga na of meneer Fier hierin gelijk krijgt van mw. Zeker, die nog steeds  $\alpha = 0,05$  gebruikt.

Meneer Fier denkt vaak met weemoed terug aan zijn eigen schooljaren. Hij zat op een kleine school. Zijn jaarlaag bestond uit 23 leerlingen, waarvan er 19 Frans volgden. Hij vraagt zich af of dit significant meer is dan nu op het Pearson College.

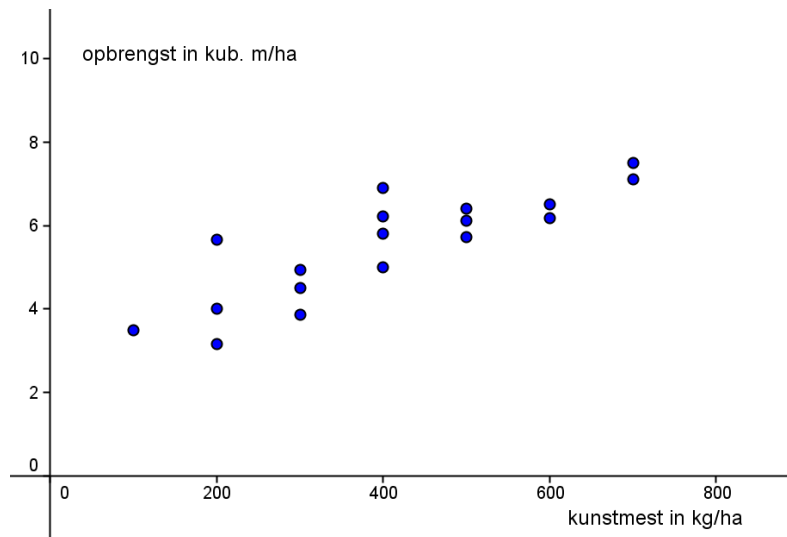
- c. Is het nodig om in deze situatie Fishers exacte toets te gebruiken?
- d. Geef de p-waarde van Fishers exacte toets en vergelijk deze met de p-waarde van de toets die gebruik maakt van de normale verdeling.

*Het vergelijken van populatiepercentages*

## Hoofdstuk 6

# Enkelvoudige lineaire regressie

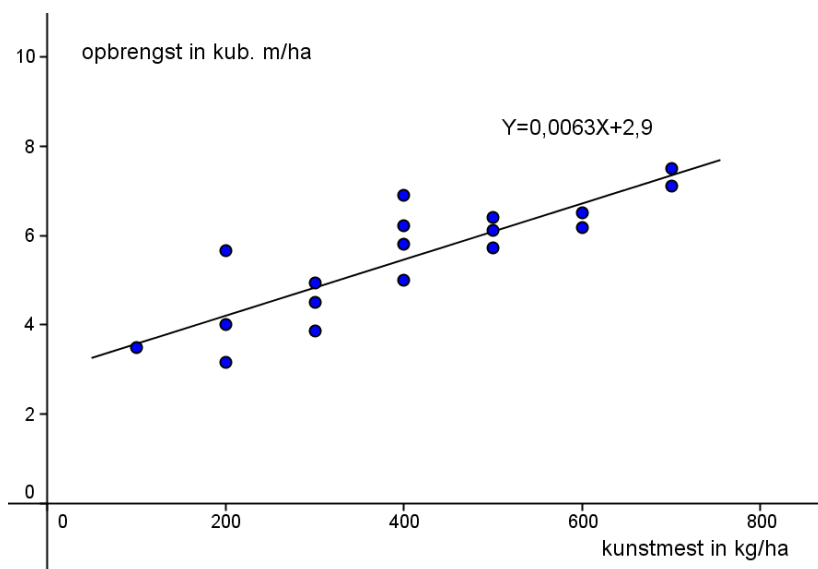
Op een aantal percelen akkerland wordt een bepaald gewas verbouwd. Per perceel wordt kunstmest gebruikt, maar niet steeds in dezelfde hoeveelheden. Per perceel is bijgehouden wat de opbrengst is. De resultaten zijn weergegeven in onderstaande grafiek. Elk punt van de grafiek hoort bij een perceel. De  $x$ -coördinaat stelt de gebruikte hoeveelheid kunstmest voor (in kg per ha) en de  $y$ -coördinaat staat voor de opbrengst (in  $m^3$  per ha).



Het plaatje suggereert dat er een verband is tussen de hoeveelheid gebruikte kunstmest ( $X$ ) en de opbrengst ( $Y$ ): de opbrengst wordt hoger als je meer kunstmest gebruikt. Sterker: het verband lijkt lineair te zijn, want het plaatje suggereert dat de punten evenwichtig gespreid liggen rond een rechte lijn. Verder zie je dat gelijke hoeveelheden kunstmest niet leiden

## Enkelvoudige lineaire regressie

tot precies gelijke opbrengsten. Er zijn kennelijk nog andere factoren in het spel die de opbrengst beïnvloeden. Dit maakt het moeilijk om het (lineaire) verband tussen hoeveelheid kunstmest en opbrengst exact te beschrijven. Voorlopig negeren we deze moeilijkheid en doen we in het volgende plaatje een eerste poging om het verband tussen  $X$  en  $Y$  te kwantificeren.



We hebben “zo goed mogelijk” op het oog een rechte lijn getrokken door de gegeven puntenwol. De formule van die lijn is ongeveer  $Y = 0,006X + 3$ . Aangezien bij gegeven  $X$  de waarde van  $Y$  niet vastligt, moeten we  $Y$  interpreteren als een *benadering* van de opbrengst, of als de *verwachte* of *gemiddelde* opbrengst bij de gegeven waarde van  $X$ . Later in dit hoofdstuk zullen we dit preciezer formuleren. Eerst zullen we onze aandacht richten op de manier waarop je “zo goed mogelijk” een lijn kunt trekken door een verzameling punten.

### 6.1 Criteria voor de best passende lijn

In het voorbeeld uit de vorige paragraaf beschikten we over 17 waarnemingen. Elke waarneming bestond uit een getallenpaar  $(X, Y)$  dat kon worden weergegeven als een punt in een assenstelsel. Als we de punten nummeren van 1 t/m 17, beschikken we over de waarnemingen  $(X_i, Y_i)$  met  $i = 1, 2, \dots, 17$ . Om de discussie algemeen te houden, vervangen we het getal 17 door de letter  $n$ . Er moet natuurlijk gelden  $n \geq 3$ , want het is geen kunst om een lijn te trekken door 2 punten en door één punt gaan oneindig veel lijnen die allemaal even goed zijn.

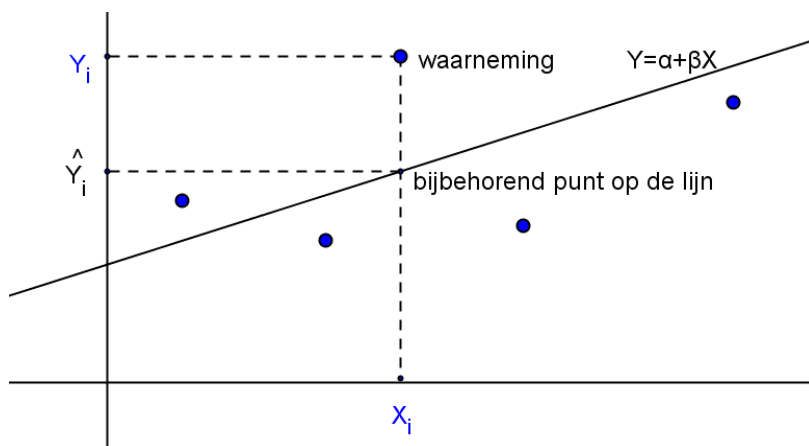
Door deze “puntenwol” trokken we een “zo goed mogelijk passende” rechte lijn, beschreven door de formule  $Y = \alpha + \beta X$ . In het plaatje uit ons



### 6.1. Criteria voor de best passende lijn

voorbeeld kwamen we uit op  $\alpha = 3$  en  $\beta = 0,006$ . We zullen de waarden van  $\alpha$  en  $\beta$  nu open laten, want we zoeken juist die waarden die de “best passende lijn” opleveren.

Over het algemeen zal niet voor alle  $i$  gelden dat  $Y_i = \alpha + \beta X_i$ . Als dat wel zo zou zijn, zouden alle punten precies op een rechte lijn liggen en dat zal in de praktijk niet dikwijls voorkomen. Daarom definiëren we, bij gegeven waarden van  $\alpha$  en  $\beta$ ,  $\hat{Y}_i = \alpha + \beta X_i$ . Het dakje boven  $Y_i$  geeft aan dat we niet te maken hebben met de feitelijke waarneming  $Y_i$ , maar met een waarde die op de lijn ligt. Deze waarde is uiteraard afhankelijk van de keuze van  $\alpha$  en  $\beta$ .



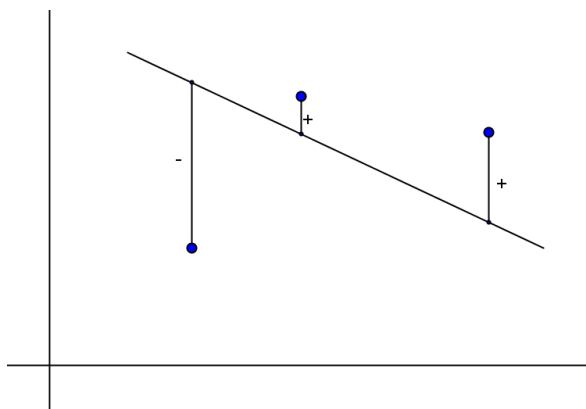
De verschillen  $Y_i - \hat{Y}_i$  noemen we de *residuen* (Engels: *residuals*). Het zal duidelijk zijn dat bij de “best passende lijn” de residuen “zo dicht mogelijk bij 0” moeten liggen. Een mogelijk criterium om dit te bereiken zou kunnen zijn:

1. Kies de lijn zó dat de optelsom  $\sum_{i=1}^n (Y_i - \hat{Y}_i)$  zo dicht mogelijk bij 0 ligt.

(In het vervolg zullen we dit korter schrijven als  $\sum (Y_i - \hat{Y}_i)$  omdat het wel duidelijk is dat de som genomen wordt over alle  $n$  waarnemingen.)

Als je hier even over nadenkt, zul je snel inzien dat dit criterium niet geschikt is. In de optelsom kunnen relatief grote positieve en negatieve residuen elkaar opheffen. Zo kan een lijn die bij elke waarneming een groot (positief of negatief) residu oplevert en die dus duidelijk niet geschikt is, tóch een optelsom geven die dicht bij 0 ligt. Het volgende plaatje geeft een voorbeeld.

## Enkelvoudige lineaire regressie



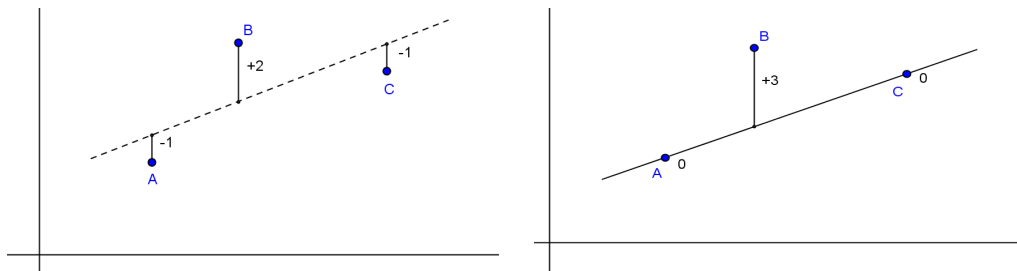
Bij deze lijn, die duidelijk “niet passend” is, ligt de som van de residuen dicht bij nul.

Het lijkt dus verstandig om ons te concentreren op de absolute waarden van de residuen. Zo komen we bij een tweede criterium:

2. Kies de lijn zó dat de optelsom  $\sum |Y_i - \hat{Y}_i|$  zo klein mogelijk is.

(In dit geval betekent “zo klein mogelijk” hetzelfde als “zo dicht mogelijk bij 0”.)

Dit criterium is duidelijk beter dan het eerste. Toch is het mogelijk om een situatie te bedenken waarin dit tweede criterium een uitkomst geeft die verschilt van wat je “redelijkerwijs” mag verwachten.



In bovenstaande plaatjes zie je twee pogingen om de “best passende lijn” te trekken door de punten  $A$ ,  $B$  en  $C$ . Welke van deze twee lijnen is het “best passend”? Volgens criterium 2 is de doorgetrokken lijn het “best passend”, want daar geldt  $\sum |Y_i - \hat{Y}_i| = 3$ , terwijl de gestippelde lijn  $\sum |Y_i - \hat{Y}_i| = 4$  geeft.

Toch zullen de meeste mensen vinden dat de stippellijn “beter past” dan de doorgetrokken lijn. De doorgetrokken lijn houdt geen rekening met het punt  $B$ , terwijl de stippellijn dat wel doet. Je mag er vanuit gaan dat elke waarneming evenveel informatie bevat, dus kan het niet goed zijn om één

## 6.2. Afleiding van de lijn volgens het principe van kleinste kwadraten

punt te negeren. Bovendien lijkt het ook redelijk om van de “best passende lijn” te verlangen dat de waarnemingen zich aan beide zijden van de lijn bevinden. Bij de stippellijn is dat het geval, maar bij de doorgetrokken lijn niet. Kennelijk moet criterium 2 zó worden aangepast dat de grote fout die de doorgetrokken lijn heeft t.o.v. punt  $B$  zwaarder weegt dan de twee kleine fouten die de gestippelde lijn heeft t.o.v. de punten  $A$  en  $C$ . Zo komen we bij een derde criterium:

3. Kies de lijn zó dat de optelsom  $\sum(Y_i - \hat{Y}_i)^2$  zo klein mogelijk is.

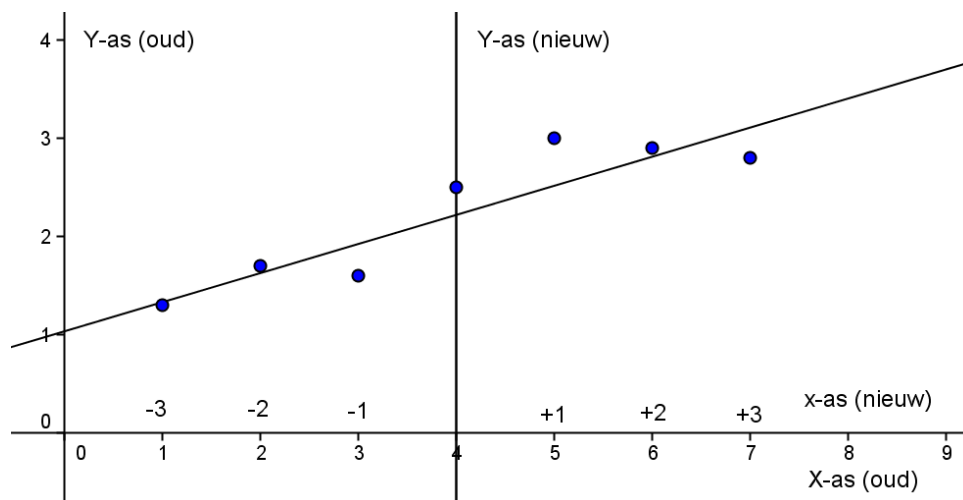
Dit criterium komt tegemoet aan het bezwaar van het eerste criterium, want alle fouten worden gekwadraterd. Positieve en negatieve fouten kunnen elkaar daardoor niet opheffen. Ook aan het bezwaar van het tweede criterium wordt tegemoet gekomen. Door het kwadrateren krijgen grote fouten extra veel gewicht. In bovenstaande plaatjes geeft criterium 3 de voorkeur aan de stippellijn ( $(-1)^2 + 2^2 + (-1)^2 = 8$ ) boven de doorgetrokken lijn ( $0^2 + 3^2 + 0^2 = 9$ ). Criterium 3 heet het criterium van de “(gewone) kleinste kwadraten” (Engels: “(ordinary) least squares”, afgekort *OLS*) en wij zullen voortaan uitsluitend met dit criterium werken. Wij zullen later in dit hoofdstuk zien dat dit criterium nog meer voordelen heeft.

## 6.2 Afleiding van de lijn volgens het principe van kleinste kwadraten

Gegeven zijn de waarnemingen  $(X_i, Y_i)$  ( $i = 1, 2, \dots, n$ ). Wij zoeken nu waarden voor de coëfficiënten  $\alpha$  en  $\beta$  zó dat de som  $S = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \alpha - \beta X_i)^2$  minimaal is.

Dit probleem gaan we eerst enigszins wijzigen, omdat de berekeningen dan eenvoudiger worden. We vervangen de waarden  $X_i$  door hun afwijkingen t.o.v. het gemiddelde  $\bar{X}$ :  $x_i = X_i - \bar{X}$ . Dit komt neer op een verplaatsing van de  $Y$ -as, zodat deze door de gemiddelde  $X$ -coördinaat van de waarnemingen komt te lopen. In de figuur hieronder zie je hoe dit werkt. Het gemiddelde van de  $X_i$ 's is  $\bar{X} = 4$ . Voor het punt geheel links geldt dus  $X_1 = 1$  en  $x_1 = 1 - 4 = -3$

## Enkelvoudige lineaire regressie



We constateren dat de richtingscoëfficiënt  $\beta$  van de rechte lijn niet wordt beïnvloed door deze wijziging van het assenstelsel. Dat geldt niet voor  $\alpha$ , want dat is het snijpunt met de  $Y$ -as. We hebben dus twee equivalenten vergelijkingen van de lijn:

$$Y = \alpha + \beta X \text{ en } Y = \alpha' + \beta x.$$

We herformuleren ons probleem: we zoeken coëfficiënten  $\alpha'$  en  $\beta$  zó dat de som  $S = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \alpha' - \beta x_i)^2$  minimaal is. De coëfficiënt  $\alpha'$  correspondeert met het snijpunt van de lijn en de nieuwe  $Y$ -as. Als we  $\alpha'$  hebben gevonden, zullen we  $\alpha$  berekenen, het snijpunt met de oorspronkelijke  $Y$ -as.

In opgave 80 wordt gevraagd om toe te lichten dat geldt:

$$\begin{aligned} S &= \sum (Y_i^2 + \alpha'^2 + \beta^2 x_i^2 - 2\alpha' Y_i - 2\beta x_i Y_i + 2\alpha' \beta x_i) \\ &= \sum Y_i^2 + n\alpha'^2 + \beta^2 \sum x_i^2 - 2n\alpha' \bar{Y} - 2\beta \sum x_i Y_i \end{aligned} \quad (1)$$

Bij een vaste waarde van  $\beta$  is  $S$  op te vatten als een kwadratische functie van  $\alpha'$ , die we als volgt kunnen schrijven:

$$S(\alpha') = n\alpha'^2 - 2n\bar{Y}\alpha' + C, \text{ waarbij } C \text{ een constante is die niet afhangt van } \alpha'.$$

We hebben te maken met een dalparabool ( $n > 0$ ), die zijn minimum bereikt als  $\frac{dS}{d\alpha'} = 0$ .

$$\text{Uit } \frac{dS}{d\alpha'} = 2n\alpha' - 2n\bar{Y} = 0 \text{ volgt } \alpha' = \bar{Y}.$$

We weten nu, ongeacht de waarde van  $\beta$ , hoe we  $\alpha'$  moeten kiezen om ervoor te zorgen dat  $S$  minimaal is.

## 6.2. Afleiding van de lijn volgens het principe van kleinste kwadraten

We kunnen, bij deze vaste waarde van  $\alpha'$ ,  $S$  ook opvatten als een kwadratische functie van  $\beta$  en deze als volgt schrijven:

$S(\beta) = \beta^2 \sum x_i^2 - 2\beta \sum x_i Y_i + C'$ , waarbij  $C'$  een constante is die niet afhangt van  $\beta$ .

Ook hier hebben we te maken met een dalparabool die zijn minimum bereikt als  $\frac{dS}{d\beta} = 0$ .

Uit  $\frac{dS}{d\beta} = 2\beta \sum x_i^2 - 2 \sum x_i Y_i = 0$  volgt  $\beta = \frac{\sum x_i Y_i}{\sum x_i^2}$ .

We hebben nu de richtingscoëfficiënt bepaald van de lijn die volgens het principe van de kleinste kwadraten het best past bij onze waarnemingen  $(X_i, Y_i)$ . Deze lijn heet de regressielijn van  $Y$  op  $X$ . Deze lijn snijdt de nieuwe  $Y$ -as in het punt  $\bar{Y}$ . Het snijpunt met de oorspronkelijke  $Y$ -as is dus  $\alpha = \bar{Y} - \beta \bar{X}$ . In opgave 81 wordt aan je gevraagd om dit toe te lichten. Het is de gewoonte om de coëfficiënten van de regressielijn aan te duiden als  $\hat{\alpha}$  en  $\hat{\beta}$ .

Samenvattend:

De regressielijn die volgens het principe van de kleinste kwadraten het best past bij de punten  $(X_i, Y_i)$  ( $i = 1, 2, \dots, n$ ) heeft als formule  $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ , met

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} = \frac{\sum X_i Y_i - n\bar{X} \cdot \bar{Y}}{\sum X_i^2 - n\bar{X}^2} \text{ en } \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}.$$

$\hat{\alpha}$  en  $\hat{\beta}$  worden de OLS-schatters van  $\alpha$  en  $\beta$  genoemd.

De OLS-regressielijn gaat door het punt  $(\bar{X}, \bar{Y})$ .

In bovenstaand kader wordt ook een formule voor  $\hat{\beta}$  gegeven die uitsluitend gebruik maakt van de oorspronkelijke gegevens en niet van de  $x_i$ 's. In opgave 82 ga je deze formule afleiden.

**Opmerking 6.2.1.** De lijn waarvan we hierboven de formule hebben afgeleid heet de regressielijn van  $Y$  op  $X$ . Dat betekent dat  $Y$  de verklaarde en  $X$  de verklarende variabele is. Als we spreken van regressie van  $X$  op  $Y$ , worden de rollen omgedraaid. We gaan dan  $X$  verklaren met behulp van  $Y$ . In het voorbeeld aan het begin van dit hoofdstuk was duidelijk dat we de opbrengst ( $Y$ ) wilden verklaren aan de hand van de gebruikte hoeveelheid kunstmest ( $X$ ). Maar niet in alle praktijkvoorbeelden staat op voorhand vast welke variabele de verklaarde moet zijn en welke de verklarende. In opgave 84 komen een situatie tegen, waarin je “beide kanten opkunt”.

## Enkelvoudige lineaire regressie

### Opgaven

80. In deze opgave ga je de afleiding (1) nader toelichten.

a. Licht toe:  $(Y_i - \alpha' - \beta x_i)^2 = Y_i^2 + \alpha'^2 + \beta^2 x_i^2 - 2\alpha' Y_i - 2\beta x_i Y_i + 2\alpha' \beta x_i$

b. Waarom geldt:

$$\begin{aligned} \sum (Y_i^2 + \alpha'^2 + \beta^2 x_i^2 - 2\alpha' Y_i - 2\beta x_i Y_i + 2\alpha' \beta x_i) = \\ \sum Y_i^2 + n\alpha'^2 + \beta^2 \sum x_i^2 - 2\alpha' \sum Y_i - 2\beta \sum x_i Y_i + 2\alpha' \beta \sum x_i \end{aligned}$$

c. Waarom geldt:

$$\sum Y_i = n\bar{Y} \text{ en } \sum x_i = 0$$

d. Controleer nu dat je afleiding (1) volledig hebt verklaard.

81. Licht toe waarom het snijpunt van de regressielijn met de oude  $y$ -as gelijk is aan  $\bar{Y} - \hat{\beta}\bar{X}$ . Gebruik de figuur uit paragraaf 6.2.

82. Laat zien hoe je uit  $x_i = X_i - \bar{X}$  de gelijkheid voor beide uitdrukkingen voor  $\hat{\beta}$  uit het kader hierboven kunt afleiden.

## 6.3 Regressieberekeningen met de GR

Aan de in vorige paragraaf afgeleide formules kun je zien dat het berekenen van  $\hat{\alpha}$  en  $\hat{\beta}$  in concrete situaties flink wat rekenwerk vraagt, zeker als het aantal waarnemingen groot is. We zullen dus snel onze toevlucht nemen tot de GR en later tot de computer. In deze paragraaf laten we zien hoe je de GR kunt gebruiken. We gebruiken de volgende gegevens als voorbeeld

$X_i$	2	6	10	16	20	24
$Y_i$	33	24	28	18	14	20

- Eerst moet je de gegevens **invoeren** in lijsten. Dat gaat het eenvoudigst via **STAT - EDIT**. Het is gebruikelijk om de  $X_i$ 's in te voeren in L1 en de  $Y_i$ 's in L2.
- Om de gegevens te **plotten** ga je naar **STAT PLOT (2nd Y=)**  
Kies (bijvoorbeeld) **Plot1** door er met de cursor op te gaan staan en op **ENTER** te drukken.  
Kies m.b.v. de cursor:  
**On**;  
Bij **Type**: het 1e icoontje (losse punten)  
Bij **Xlist** en **Ylist** : L1 en L2  
Bij **Mark**: één van de drie icoontjes

### 6.3. Regressieberekeningen met de GR

Kies een geschikt WINDOW: (zó dat alle  $X_i$ 's en  $Y_i$ 's in het venster passen.)

GRAPH levert de gewenste plot. (Zorg ervoor dat eventuele functies in het Y=scherm uit staan of verwijderd zijn, anders krijg je de bijbehorende grafiek ook te zien.)

- Om de **formule van de regressielijn** te vinden ga je naar STAT - CALC  
Kies: **LinReg(a+bx)** (Let op: er is ook een optie **LinReg(ax+b)**. Deze doet precies hetzelfde als **LinReg(a+bx)**, alleen zijn de namen van de parameters a en b omgewisseld. De notatie a+bx sluit het best aan bij wat in de statistiek gebruikelijk is.)  
Je ziet nu op je gewone scherm **LinReg(a+bx)**, tik hierachter L1,L2 in en druk op ENTER. Zo verschijnt de formule van de regressielijn ( $Y = 31,7 - 0,68X$ ). (Het is niet nodig om L1, L2 in te tikken, omdat het zg. defaultwaardes zijn. Als je gegevens in L3 en L4 staan is het wel nodig.)
- Om de **regressielijn te plotten** zorg je er opnieuw voor dat **LinReg(a+bx)** op je gewone scherm staat. Tik hier achter in L1,L2,Y1 en druk op ENTER. (L1,L2 is eigenlijk niet nodig; alleen Y1 is voldoende; Y1 krijg je via de VARS.) De formule van de regressielijn staat nu in het functiescherm Y=. Via GRAPH krijg je de regressielijn te zien samen met de oorspronkelijke waarnemingen, als tenminste **Plot1** nog aan staat.
- Om een **lijst met residuen** te krijgen, voer je eerst het commando **LinReg(a+bx)** L1,L2,Y1 uit. Vervolgens ga je naar het lijsteninvoerscherm via STAT - EDIT. Je zet de cursor op L3. Je gaat naar LIST - NAMES - RESID en drukt op ENTER. Je bent nu terug in het lijsteninvoerscherm en drukt nogmaals op ENTER. De lijst L3 is nu gevuld met residuen. Bij de X-waarde van 10 (en de bijbehorende Y-waarde van 28) vind je een residu van 3,1304. Controleer of dit klopt, via 28-Y1(10).

## Enkelvoudige lineaire regressie

### Opgaven

83. Van een vijftal gezinnen is in een bepaald jaar het netto inkomen ( $Y$ ) en het bedrag dat het gezin in dat jaar gespaard heeft ( $S$ ) geregistreerd. De gegevens in euro zijn:

$Y$	25.000	33.000	28.000	18.000	18.000
$S$	1.700	3.600	3.100	2.100	900

- Maak op de GR een plot van deze gegevens en geef de formule van de regressielijn van  $S$  op  $Y$ . (D.w.z.: in de formule moet je  $S$  uitdrukken als functie van  $Y$ .)
  - Bij welk gezin is het verschil tussen het werkelijk gespaarde bedrag en het bedrag dat door de regressievergelijking wordt voorspeld het grootst?
  - Stel opnieuw de formule van de regressievergelijking op, gebruikmakend van de formules uit paragraaf 6.2.
  - Van een zesde gezin is bekend dat het netto inkomen 30.000 bedroeg. Geef een schatting van het door dat gezin gespaarde bedrag.
84. Van elf biatleten zijn de volgende twee gegevens geregistreerd:  
 $X$  : het aantal minuten dat ze het op de loopband in een vast (hoog) tempo konden volhouden;  
 $Y$ : hun persoonlijk record (in minuten) op de 20 km langlaufen.

$X$	7,7	8,4	8,7	9,0	9,6	9,6	10,0	10,2	10,4	11,0	11,7
$Y$	71,0	71,4	65,0	68,7	64,4	69,4	63,0	64,6	66,9	62,9	61,7

- Maak een plot van de gegevens. Lijkt het redelijk om een lineair verband op te stellen?
- Stel (met de GR) de regressievergelijking op van  $Y$  op  $X$ .
- Een twaalfde atleet houdt het 10,4 minuten vol op de loopband. Geef een voorspelling van zijn tijd op de 20 km langlaufen.
- Herschrijf de bij b. gevonden regressievergelijking in de vorm  $X = \alpha + \beta Y$ .
- Stel (met de GR) de regressievergelijking op van  $X$  op  $Y$ .
- De bij d. en e. gevonden vergelijkingen zijn niet identiek. Kun je verklaren waarom niet?
- Van een andere biatleet is het persoonlijk record op de 20 km langlaufen 70,0 minuten. Je wilt zijn tijd op de loopband voorspellen. Moet je hiervoor de formule bij d. of bij e. gebruiken? Geef de voorspelling.



## 6.4. Een stochastisch regressiemodel

- h. Je wiskundeleraar houdt het nog geen minuut vol op de loopband. Kun je zijn tijd op de 20 km langlaufen voorspellen?

85. Gegeven een aantal waarnemingen  $X_i$  ( $i = 1, \dots, n$ ). Wij willen deze waarnemingen vervangen door één getal  $\alpha$  op basis van het principe van de kleinste kwadraten. Met andere woorden: voor welke  $\alpha$  is de som  $S = \sum (X_i - \alpha)^2$  minimaal?

(Je moet uiteraard dezelfde techniek gebruiken als bij het minimaliseren van de  $S$  uit paragraaf 6.2, maar het is niet nodig om de  $X_i$ 's eerst om te schrijven tot  $x_i$ 's.) Geef commentaar op de formule die je als antwoord vindt.

## 6.4 Een stochastisch regressiemodel

Tot dusver hebben we min of meer mechanisch een “zo goed mogelijk passende” rechte lijn getrokken door een gegeven puntenwolk. We hebben hiervoor het criterium van de kleinste kwadraten gebruikt en we hebben een aantal redenen gevonden waarom dit criterium redelijk lijkt. Tegelijkertijd blijven veel vragen onbeantwoord. In ons voorbeeld van de invloed van kunstmest op de opbrengst van een stuk land zou je je kunnen afvragen wat er gebeurt als je het experiment herhaalt. Er worden nieuwe gegevens geproduceerd en door deze nieuwe puntenwolk trekken we net als eerst de OLS-regressielijn. Naar alle waarschijnlijkheid zal deze tweede lijn verschillen van de eerste. Er zit dus nog onzekerheid in de “best passende lijn”, zelfs als het onderliggende verband (invloed van kunstmest op opbrengst) ongewijzigd is. We willen uiteindelijk in staat zijn om over dit onderliggende verband en over de precisie van de OLS-schatters  $\hat{\alpha}$ ,  $\hat{\beta}$  en  $\hat{Y}_i$  (kans)uitspraken te doen. Dat kan alleen als we gebruik maken van een geschikt model.

Laten we hiervoor nog even terugkeren naar het voorbeeld uit het begin van dit hoofdstuk, waarin de opbrengst  $Y$  van een stuk land werd gerelateerd aan de hoeveelheid gebruikte kunstmest  $X$ . Je kunt je voorstellen dat we voor één vaste hoeveelheid kunstmest  $X_i$  het experiment een groot aantal keren herhalen. Ondanks het feit dat we steeds dezelfde hoeveelheid kunstmest gebruiken, zal de opbrengst  $Y_i$  niet steeds precies dezelfde zijn. Daarvoor zijn er teveel oncontroleerbare (toevallige) omstandigheden: kwaliteit van het zaaigoed, hoeveelheid zon en regen enzovoorts. De verschillende opbrengsten  $Y_i$  zijn daarom goed op te vatten als een steekproef uit een populatie. Met andere woorden we vatten  $Y_i$  op als een stochast (kansvariabele) met een eigen kansverdeling en bijbehorende verwachtingswaarde  $E(Y_i) = \mu_i$  en variantie  $Var(Y_i) = \sigma_i$ . Deze verwachtingswaarde en variantie hebben beide een index  $i$  om aan te geven dat er een verband bestaat met de hoeveelheid gebruikte kunstmest  $X_i$ . (Merk op dat  $X_i$  geen stochast is, maar (per  $i$ ) een vast getal. We kunnen de

### Enkelvoudige lineaire regressie

hoeveelheid kunstmest per experiment immers zelf bepalen. Later zullen we terugkomen op de situatie waarin de waarden van  $X_i$  wél door het toeval worden bepaald.) We hebben dus te maken met een aantal verschillende kansverdelingen met bijbehorende verwachtingswaarde en variantie, per waarde van  $X_i$  één.

Om tot een bruikbaar model te komen maken we de volgende *veronderstellingen* over deze kansverdelingen.

1. De verwachtingswaarden  $E(Y_i)$  liggen op een rechte lijn:

$$E(Y_i) = \mu_i = \alpha + \beta X_i$$

2. Alle kansverdelingen hebben dezelfde variantie  $\sigma^2$ .
3. De stochasten  $Y_i$  zijn onderling onafhankelijk.

Alle parameters uit deze veronderstellingen ( $\alpha$ ,  $\beta$  en  $\sigma^2$ ) zijn niet-waarneembare grootheden. Wat wij waarnemen is alleen de puntenwolk  $(X_i, Y_i)$ . We zullen m.b.v. die puntenwolk een zo goed mogelijke schatting proberen te maken van de parameters.

Dikwijls worden deze veronderstellingen ook op een andere manier geformuleerd:

$$Y_i = \alpha + \beta X_i + e_i$$

waarin de  $e_i$ 's onderling onafhankelijke kansvariabelen zijn met verwachtingswaarde 0 en variantie  $\sigma^2$ .

In deze formulering (die inhoudelijk natuurlijk identiek is aan de veronderstellingen 1, 2 en 3) worden de storingstermen  $e_i$  (e van error) apart benoemd. Deze kansvariabelen zorgen ervoor dat de werkelijke regressielijn  $y = \alpha + \beta x$  onzichtbaar blijft. Als we eenmaal de OLS-schatters  $\hat{\alpha}$  en  $\hat{\beta}$  hebben berekend dan zijn de residuen schattingen van de  $e_i$ 's. We noteren dan  $\hat{e}_i = Y_i - \hat{Y}_i (= Y_i - (\hat{\alpha} + \hat{\beta}X_i))$ . Het is van belang dat je voor jezelf een duidelijk onderscheid maakt tussen de theoretische, niet waarneembare, grootheden en de schattingen die we van deze grootheden maken op basis van de waarnemingen. Als regel worden dergelijke schattingen aangeduid door toevoeging van een dakje.

Het is je wellicht opgevallen dat we geen veronderstelling maken over de vorm van de verdeling van de  $e_i$ 's. Later zullen we ook nog veronderstellen dat de  $e_i$ 's normaal verdeeld zijn. Maar eerst, in paragraaf 6.5, zullen we een aantal eigenschappen van OLS-schatters bespreken waarbij deze veronderstelling niet nodig is.

#### 6.4. Een stochastisch regressiemodel

**Opmerking 6.4.1.** Het is belangrijk om te beseffen dat we met de introductie van een stochastisch regressiemodel een wezenlijke stap hebben gezet in vergelijking met hetgeen in de eerste drie paragrafen is besproken. De parameters  $\alpha$  en  $\beta$  zijn niet langer variabele getallen die we zó proberen vast te stellen dat een “goed passende” lijn ontstaat. Ze zijn nu *vaste (maar onbekende, want niet-waarneembare) parameters* die een onderliggende werkelijkheid beschrijven waarvan we zoveel mogelijk te weten willen komen.  $\hat{\alpha}$ ,  $\hat{\beta}$  zijn nu *schaters* geworden van deze onderliggende werkelijkheid. Ze zijn op te vatten als kansvariabelen, want afhankelijk van de (toevallige) waarnemingen  $Y_i$ .

**Opmerking 6.4.2.** Het is goed om stil te staan bij de stochastische component in ons model, de storingstermen  $e_i$ . Waarom wordt de waarde van  $Y_i$  niet geheel vastgelegd door  $X_i$ ? Er kan natuurlijk sprake zijn van meetfouten. In ons eerste voorbeeld zou het zo kunnen zijn dat een deel van de opbrengst van het land bij het oogsten verloren gaat of dat het volume van de opbrengst slordig wordt opgemeten. Kun je dan nog wel aannemen dat aan de voorwaarde  $E(e_i) = 0$  is voldaan? In het geval van slordig oogsten niet. Je moet je dan realiseren dat je waarnemingen  $Y_i$  over de geoogste opbrengst en niet over de werkelijke opbrengst gaan. Bij het meten van het volume is het veel redelijker om te veronderstellen dat de meting soms te hoog en dan weer te laag zal uitvallen.

Daarnaast kan er sprake zijn van allerlei externe factoren die we niet in de hand hebben, maar die wel van invloed zijn op de verklaarde variabele. In ons landbouwvoorbeeld wordt de opbrengst natuurlijk ook bepaald door de manier van zaaien, door de hoeveelheid regen, zon enzovoorts. Als het mogelijk is, moet je proberen om bij de proefopzet dergelijke externe factoren zoveel mogelijk gelijk te houden. Een andere mogelijkheid is om in het model expliciet met andere, waarneembare, factoren rekening te houden. We krijgen dan meer dan één verklarende variabele. Hierop zullen we in hoofdstuk 8 nader ingaan.

**Opmerking 6.4.3.** Als je in de praktijk een regressie-analyse uitvoert, is het altijd van belang om na te gaan of de veronderstellingen die ten grondslag liggen aan het model redelijk zijn.

De veronderstelling van een onderliggend lineair verband tussen de variabelen is essentieel. Je moet daarom altijd een plaatje maken van de gegevens. Daaraan kun je meestal goed zien of de punten rond een rechte lijn liggen of dat een kromme beter zou passen.

Daarnaast moet je jezelf afvragen of de storingstermen onafhankelijk zijn en of ze dezelfde variantie hebben. Een belangrijk hulpmiddel bij het beantwoorden van deze vragen zijn de residuen. Het is dan ook verstandig om deze goed te bekijken. Als de residuen een duidelijk patroon vertonen, zal dat een indicatie zijn van het feit dat niet alle aannames reëel zijn.

## 6.5 Eigenschappen van $\hat{\alpha}$ en $\hat{\beta}$

We reproduceren hier de OLS-schatter die we in paragraaf 6.2 hebben afgeleid:

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} = \frac{\sum X_i Y_i - n \bar{X} \cdot \bar{Y}}{\sum X_i^2 - n \bar{X}^2} \text{ en } \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}, \text{ met } x_i = X_i - \bar{X}.$$

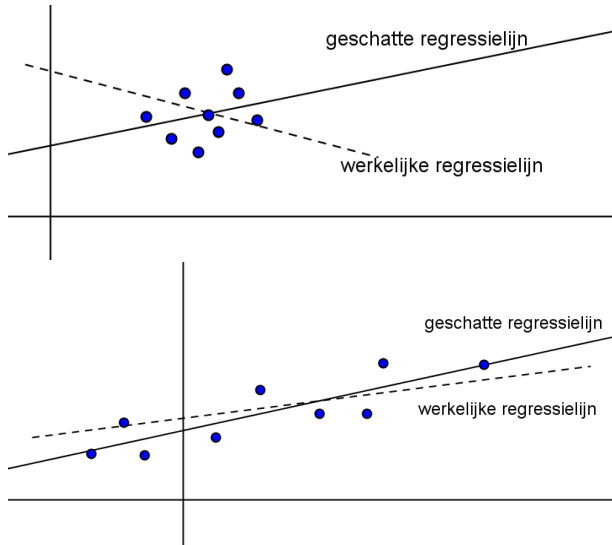
$\hat{\beta}$  en  $\hat{\alpha}$  zijn functies van de  $X_i$ 's en  $Y_i$ 's. De  $X_i$ 's zijn vaste getallen, maar de  $Y_i$ 's zijn kansvariabelen. Dus  $\hat{\beta}$  en  $\hat{\alpha}$  zijn ook kansvariabelen, ze hebben een kansverdeling, een verwachting en een variantie. Er geldt (we zullen het verderop bewijzen):

$$\begin{aligned} E(\hat{\beta}) &= \beta & \text{en} & & E(\hat{\alpha}) &= \alpha \\ \text{Var}(\hat{\beta}) &= \frac{\sigma^2}{\sum x_i^2} & \text{en} & & \text{Var}(\hat{\alpha}) &= \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) \end{aligned}$$

**Opmerking 6.5.1.** De eerste twee gelijkheden zijn van speciaal belang. Ze zeggen dat  $\hat{\beta}$  en  $\hat{\alpha}$  *zuivere schatters* zijn. Als we op basis van een puntenwolk  $\hat{\beta}$  en  $\hat{\alpha}$  schatten en we herhalen dit proces heel vaak, met steeds een nieuwe puntenwolk, dan krijgen we per keer nieuwe waarden van  $\hat{\beta}$  en  $\hat{\alpha}$ . Als we het gemiddelde nemen van al die waarden dan krijgen we “op den duur” de werkelijke  $\beta$  en  $\alpha$ .

## 6.5. Eigenschappen van $\hat{\alpha}$ en $\hat{\beta}$

**Opmerking 6.5.2.** Wat betekent de formule voor  $Var(\hat{\beta})$  in praktijk? In de eerste plaats zien we dat deze variantie evenredig is met  $\sigma^2$ . Dit ligt voor hand: als de variantie van storingstermen groot is, zal het moeilijk zijn om een accurate schatting van  $\beta$  te maken en dat zie je terug in een grote variantie van  $\hat{\beta}$ . Verder is  $Var(\hat{\beta})$  omgekeerd evenredig met  $\sum x_i^2$ . In het linkerplaatje hierna zie je een situatie waarin de  $X_i$ 's dicht rond het gemiddelde  $\bar{X}$  liggen, met als gevolg dat  $\sum x_i^2$  klein is. In het rechterplaatje zijn de  $X_i$ 's breed gespreid rond het gemiddelde  $\bar{X}$ , met als gevolg dat  $\sum x_i^2$  groot is. Het zal duidelijk zijn dat in het linkerplaatje de richtingscoëfficiënt van de werkelijke regressielijn veel minder nauwkeurig kan worden geschat dan in het rechterplaatje. Dit zie je terug in de formule voor  $Var(\hat{\beta})$ .



Om de formules voor  $\hat{\beta}$  te bewijzen merken we eerst op dat  $\hat{\beta}$  een *lineaire combinatie* is van de  $Y_i$ 's. Immers:  $\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum w_i Y_i$  met  $w_i = \frac{x_i}{\sum x_i^2}$ . Daarbij zijn de  $w_i$ 's vaste getallen die niet van de  $Y_i$ 's afhangen.

Er geldt dus:

$$\begin{aligned} E(\hat{\beta}) &= E\left(\sum w_i Y_i\right) = \sum w_i E(Y_i) = \sum w_i (\alpha + \beta X_i) = \sum w_i (\alpha + \beta(\bar{X} + x_i)) \\ &= \sum w_i (\alpha + \beta \bar{X} + \beta x_i) = (\alpha + \beta \bar{X}) \sum w_i + \beta \sum w_i x_i \end{aligned}$$

Verder geldt:

$$\sum w_i = \sum \frac{x_i}{\sum x_i^2} = \frac{1}{\sum x_i^2} \sum x_i = 0 \text{ en } \sum w_i x_i = \sum \frac{x_i^2}{\sum x_i^2} = \frac{1}{\sum x_i^2} \sum x_i^2 = 1$$

Dus  $E(\hat{\beta}) = \beta$ .

Om de formule voor  $Var(\hat{\beta})$  te bewijzen, maken we gebruik van het feit dat

## Enkelvoudige lineaire regressie

de  $Y_i$ 's onafhankelijk zijn. Daarom geldt:

$$Var(\hat{\beta}) = Var\left(\sum w_i Y_i\right) = \sum w_i^2 Var(Y_i) = \sigma^2 \sum w_i^2 = \frac{\sigma^2}{\left(\sum x_i^2\right)^2} \sum x_i^2 = \frac{\sigma^2}{\sum x_i^2}.$$

Hiermee is het bewijs voor de formules van  $E(\hat{\beta})$  en  $Var(\hat{\beta})$  geleverd. De overeenkomstige formules voor  $\hat{\alpha}$  ga je (deels) afleiden in de opgaven.

### Opgave

86. Is  $\hat{\alpha}$  een lineaire combinatie van de  $Y_i$ 's? Onderbouw je antwoord.

87. Bewijs  $E(\hat{\alpha}) = \alpha$ .

88. Gegeven is  $\bar{X} = 0$ . Bewijs  $Var(\hat{\alpha}) = \frac{\sigma^2}{n}$ .

## 6.6 Betrouwbaarheidsintervallen en toetsen voor $\beta$

We beschikken nu over formules voor de verwachting en de variantie van  $\hat{\beta}$ . Om betrouwbaarheidsintervallen te kunnen maken hebben we als *extra* veronderstelling nodig dat de  $Y_i$ 's (en dus ook de  $e_i$ 's) normaal verdeeld zijn. Voor alle duidelijkheid herhalen we hier alle gemaakte veronderstellingen:

$$Y_i = \alpha + \beta X_i + e_i$$

waarin de  $e_i$ 's onderling onafhankelijke kansvariabelen zijn, normaal verdeeld met verwachtingswaarde 0 en variantie  $\sigma^2$ .

We weten dat  $\hat{\alpha}$  en  $\hat{\beta}$  lineaire combinaties zijn van de  $Y_i$ 's. De  $Y_i$ 's zijn normaal verdeeld, dus  $\hat{\alpha}$  en  $\hat{\beta}$  zijn ook normaal verdeeld en hun parameters hebben we in de vorige paragraaf afgeleid. We kunnen dus concluderen dat

$$Z = \frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / \sum x_i^2}} \text{ standaardnormaal verdeeld is.}$$

Met deze formule kunnen we nog geen betrouwbaarheidsinterval voor  $\beta$  afleiden, omdat de parameter  $\sigma^2$  onbekend is. Net zoals we in hoofdstuk 2 hebben gedaan, zullen we eerst een schatting moeten maken van  $\sigma^2$ . Een voor de hand liggende schatter van  $\sigma^2$  is:

$$\frac{1}{n} \sum e_i^2 = \frac{1}{n} \sum (Y_i - \alpha - \beta X_i)^2 \quad (*)$$

maar in de praktijk is deze niet bruikbaar, omdat we  $\alpha$  en  $\beta$  niet kennen. Het is verleidelijk om nu  $\alpha$  en  $\beta$  te vervangen door hun schatters  $\hat{\alpha}$  en  $\hat{\beta}$ ,

## 6.6. Betrouwbaarheidsintervallen en toetsen voor $\beta$

maar je weet zeker dat je dan een waarde zult vinden die kleiner is dan (\*), omdat we  $\hat{\alpha}$  en  $\hat{\beta}$  zó gekozen hebben dat (\*) minimaal is. We kiezen daarom een schatter die iets groter is dan (\*):

$$s^2 = \frac{1}{n-2} \sum (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 = \frac{1}{n-2} \sum (Y_i - \hat{Y}_i)^2 = \frac{1}{n-2} \sum \hat{e}_i^2$$

Je kunt bewijzen dat (maar we zullen dat achterwege laten):

- $s^2$  een zuivere schatter is (m.a.w.  $E(s^2) = \sigma^2$ )
- $\frac{(n-2)s^2}{\sigma^2}$  een  $\chi^2$ -verdeling heeft met  $n-2$  vrijheidsgraden
- $s^2$  onafhankelijk is van  $\hat{\alpha}$  en  $\hat{\beta}$ .

Hieruit volgt:

$$T = \frac{\hat{\beta} - \beta}{\sqrt{s^2 / \sum x_i^2}} \text{ heeft een } t\text{-verdeling met } n-2 \text{ vrijheidsgraden.}$$

en

het 95%-betrouwbaarheidsinterval voor  $\beta$  wordt gegeven door:

$$\beta = \hat{\beta} \pm t_{[n-2], 0.05} \frac{s}{\sqrt{\sum x_i^2}}$$

**Opmerking 6.6.1.** Het bovenstaande lijkt sterk op de theorie voor het schatten van de parameters van een normaal verdeelde populatie. (Zie paragraaf 2.1.) Daar was de schatter voor  $\sigma^2$   $s^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$  en  $\frac{(n-1)s^2}{\sigma^2}$  had een  $\chi^2$ -verdeling met  $n-1$  vrijheidsgraden.

In ons regressiemodel delen we de kwadraatsom door  $n-2$  en heeft de resulterende  $\chi^2$ -verdeling  $n-2$  vrijheidsgraden. Het verschil zit hem in het aantal parameters dat wordt geschat. Bij een normaal verdeelde populatie moeten we eerst de verwachtingswaarde  $\mu$  schatten (door  $\bar{X}$ ) voordat we  $\sigma^2$  kunnen schatten. In het regressiemodel moeten we eerst  $\alpha$  en  $\beta$  schatten. De (informele) vuistregel is: voor elke parameter die je moet schatten voordat je  $\sigma^2$  kunt schatten, verlies je een vrijheidsgraad (en wordt de spreiding van je schatter groter).

## *Enkelvoudige lineaire regressie*

### **Opgave**

89. Geef de formule voor het 95%-betrouwbaarheidsinterval voor  $\alpha$ .



## 6.6. Betrouwbaarheidsintervallen en toetsen voor $\beta$

**Voorbeeld 6.6.1.** In het voorbeeld uit paragraaf 6.3 willen we een 95%-betrouwbaarheidsinterval voor  $\hat{\beta}$  berekenen.

Reeds berekend is  $\hat{\beta} = -0,68$ . Met behulp van de lijst met residuen berekenen we

$\sum \hat{e}_i^2 = 75,89$ , dus  $s = \sqrt{\frac{1}{4} \cdot 75,89} = 4,36$ . M.b.v. 2-Var Stats vinden we verder

$\sum x_i^2 = \sum X_i^2 - n\bar{X}^2 = 1372 - 6 \cdot 13^2 = 358$  en  $t_{[4],0.05} \approx \text{invT}(0.975, 4) \approx 2,776$ .

Het betrouwbaarheidsinterval wordt dus:  $\beta = -0,68 \pm 2,776 \cdot 4,36/\sqrt{358}$ , d.w.z.  $[-1,32; -0,04]$ .

Je kunt ook de GR alle berekeningen laten uitvoeren: STAT - TESTS - LinRegTInt. De invoer spreekt verder voor zich: het veld RegEQ kun je leeg laten. Als je het vult met bijvoorbeeld Y1, dan wordt de geschatte regressielijn opgeslagen onder de functie Y1.

**Voorbeeld 6.6.2.** In het voorbeeld van paragraaf 6.3 willen we  $H_0 : \beta = 0$  toetsen tegen  $H_1 : \beta \neq 0$  bij een significantieniveau van 5%. We constateren dat 0 niet in 95%-betrouwbaarheidsinterval ligt en concluderen direct dat de nulhypothese wordt verworpen.

Als we niet over het betrouwbaarheidsinterval beschikken, berekenen we de toetsgrootte

$$T = \frac{\hat{\beta} - \beta}{\sqrt{s^2 / \sum x_i^2}}. \quad T \text{ heeft een t-verdeling met } 6 - 2 = 4 \text{ vrijheidsgraden.}$$

Onder  $H_0$  geldt  $\beta = 0$ , dus berekenen we uit de steekproef:

$$T = \frac{-0,68 - 0}{4,36\sqrt{358}} = -2,95$$

De bijbehorende (tweezijdige)  $p$ -waarde is  $2 \cdot \text{tcdf}(-10^{99}, -2.95, 4) = 0.04$ . Dus we verwerpen de hypothese.

De GR kent een functie waarmee je deze toets kunt uitvoeren: STAT - TEST - LinRegTTest. Aan het invoerscherm zie je dat je met de GR alleen  $H_0 : \beta = 0$  kunt toetsen en niet bijvoorbeeld  $H_0 : \beta = -2$ . De uitvoer spreekt voor zich.

**Opmerking 6.6.2.** De hypothese  $H_0 : \beta = 0$  zul je in de praktijk vaak tegen komen. Als  $\beta = 0$ , wordt de regressielijn  $Y_i = \alpha + e_i$ . Dit betekent dat de waarde van  $X_i$  niet van invloed is op de waarde van  $Y_i$ . Als je  $H_0 : \beta = 0$  kunt verwerpen, heb je statistisch aangetoond dat er een verband bestaat tussen  $Y$  en  $X$ .

## Enkelvoudige lineaire regressie

### Opgaven

90. a. Toets in de situatie van opgave 83  $H_0 : \beta = 0$  tegen  $H_1 : \beta > 0$  bij een significantieniveau van 0,05.  
b. Wat is de uitkomst van de toets als je twee- in plaats van eenzijdig toetst? Geef je antwoord zonder opnieuw een berekening te maken.
91. Ultraviolet licht kan schadelijk zijn voor planten. De toegebrachte schade wordt gemeten met de zg. “sunburn index”. In een proef met een bepaald type plant werden de volgende gegevens verzameld. Hierbij is  $X$  de afstand in cm tot een ultraviolette lichtbron en  $Y$  de waarde van de sunburn index.

$x$	0,18	0,21	0,25	0,26	0,30	0,32	0,36	0,40
$y$	4,0	3,7	3,0	2,9	2,6	2,5	2,2	2,0
$x$	0,40	0,50	0,51	0,54	0,61	0,62	0,63	
$y$	2,1	1,5	1,5	1,5	1,3	1,2	1,1	

- a. Geef de OLS-schatting van de regressievergelijking en geef het 95%-betrouwbaarheidsinterval voor  $\beta$ .
- b. Bereken de  $p$ -waarde van de toets van de hypothese  $H_0 : \beta = -4$  tegen  $H_1 : \beta < -4$ .  
(Tip: de GR kan deze toets niet rechtstreeks uitvoeren, dus je moet “handig” te werk gaan met de onderliggende formules en de informatie die de GR wél geeft.)
- c. Maak een plot van de residuen en geef commentaar.

## 6.7 Betrouwbaarheidsintervallen voor $\mu_0$ en $Y_0$

We hebben ons tot dusver beziggehouden met betrouwbaarheidsintervallen (en toetsen) voor  $\beta$ . Het nut van een betrouwbaarheidsinterval voor  $\beta$  lijkt duidelijk, omdat  $\beta$  iets zegt over de wijze waarop  $Y$  wordt verklaard door  $X$ . In termen van het voorbeeld aan het begin van dit hoofdstuk:  $\beta$  vertelt ons hoeveel kubieke meter extra opbrengst we mogen verwachten als we 1 kg extra kunstmest op onze akker gebruiken. De betekenis van  $\alpha$  is lang niet in alle gevallen relevant. In ons voorbeeld is  $\alpha$  de verwachte opbrengst als je helemaal geen kunstmest gebruikt, maar in veel andere situaties is de verwachte  $Y$  bij een  $X$  van 0 op zichzelf niet van belang. Wat je eigenlijk zou willen is zelf de relevante waarde van  $X$  kunnen kiezen. Met andere woorden: wat kunnen we zeggen over de opbrengst  $Y_0$  (of over de verwachte opbrengst, die we  $\mu_0$  zullen noemen) als we een (door onszelf gekozen) hoeveelheid  $X_0$  aan kunstmest gebruiken?

## 6.7. Betrouwbaarheidsintervallen voor $\mu_0$ en $Y_0$

Laat  $X_0$  een willekeurig gekozen waarde zijn van de verklarende variabele die al dan niet kan samenvallen met één van de waargenomen  $X_i$  ( $i = 1, \dots, n$ ). De bijbehorende  $Y$ -coördinaat op de regressielijn is  $\alpha + \beta X_0$ . De bijbehorende waarde van de verklaarde variabele is  $Y_0 = \alpha + \beta X_0 + e_0$ , waarbij  $e_0$  een nieuwe waarde van de storingsterm is.  $Y_0$  en  $e_0$  zijn kansvariabelen, dus voor dezelfde waarde van  $X_0$  zijn meerdere uitkomsten van  $Y_0$  en  $e_0$  mogelijk. Er geldt  $E(Y_0) = \alpha + \beta X_0 = \mu_0$ .

De OLS-schatter van  $\mu_0$  is  $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta}X_0$ . (Bedenk dat  $\hat{\mu}_0$  een kansvariabele is, omdat  $\hat{\mu}_0$  afhangt van de kansvariabelen  $\hat{\alpha}$  en  $\hat{\beta}$  die op hun beurt afhangen van de kansvariabelen  $Y_i$ .  $\mu_0, \alpha$  en  $\beta$  zijn geen kansvariabelen, maar constante (onbekende) parameters.) Er geldt:

$$E(\hat{\mu}_0) = E(\hat{\alpha} + \hat{\beta}X_0) = E(\hat{\alpha}) + X_0 E(\hat{\beta}) = \alpha + \beta X_0 = \mu_0$$

Dus  $\hat{\mu}_0$  is een zuivere schatter.

Het bepalen van  $Var(\hat{\mu}_0) = Var(\hat{\alpha} + \hat{\beta}X_0)$  is moeilijker. We kunnen *niet* schrijven

$Var(\hat{\alpha} + \hat{\beta}X_0) = Var(\hat{\alpha}) + Var(\hat{\beta}X_0)$ , want  $\hat{\alpha}$  en  $\hat{\beta}$  zijn niet onafhankelijk. Je kunt echter aantonen dat  $\hat{\alpha}$  en  $\hat{\beta}$  wél onafhankelijk zijn als  $\bar{X} = 0$ , m.a.w. als we werken met een verschoven  $y$ -as, zoals in paragraaf 6.2. en opgave 88. Via deze omweg kun je  $Var(\hat{\mu}_0)$  berekenen. We laten dit nu achterwege en vermelden alleen het resultaat:

$$\begin{aligned} E(\hat{\mu}_0) &= E(\hat{\alpha} + \hat{\beta}X_0) = \alpha + \beta X_0 = \mu_0 \\ &\text{en} \\ Var(\hat{\mu}_0) &= Var(\hat{\alpha} + \hat{\beta}X_0) = \sigma^2 \left( \frac{1}{n} + \frac{x_0^2}{\sum x_i^2} \right) \end{aligned}$$

### Opgaven

92. Gebruik bovenstaande formule om  $Var(\hat{\alpha})$  te berekenen.
93.
  - a. Leg uit waarom  $\hat{\mu}_0$  een lineaire combinatie is van de  $Y_i$ 's.
  - b. Uit a. volgt dat  $\hat{\mu}_0$  normaal verdeeld is. En  $Z = (\hat{\mu}_0 - \mu_0) / \sqrt{Var(\hat{\mu}_0)}$  is standaardnormaal verdeeld. Geef de formule voor het 95%-betrouwbaarheidsinterval voor  $\mu_0$ .
94.
  - a. Geef in de situatie van opgave 83 de 90%-betrouwbaarheidsintervallen voor de verwachte besparingen van een gezin met een inkomen van 24.000 euro en van 30.000 euro.
  - b. Waarom is het interval voor 30.000 euro groter dan voor 24.000 euro?

### Enkelvoudige lineaire regressie

De hierboven afgeleide betrouwbaarheidsintervallen gelden voor de *verwachte waarde*  $\mu_0$  van  $Y_0$ . In termen van ons oorspronkelijke voorbeeld: als we een groot aantal keren een hectare land bemesten met  $X_0$  kg kunstmest, dan zal de *gemiddelde* opbrengst  $\mu_0$  zijn.

Het is natuurlijk zeker zo interessant om een betrouwbaarheidsinterval af te leiden voor  $Y_0$  zelf. In de volgende opgave gaan we dat doen.

### Opgave

95. Er geldt  $Y_0 = \mu_0 + e_0$ . Aangezien  $E(e_0) = 0$ , is de (punt)schatting voor  $Y_0$  gelijk aan de (punt)schatting voor  $\mu_0$ , dus  $\hat{Y}_0 = \hat{\mu}_0$ .
- a. Licht toe dat  $Y_0$  en  $\hat{\mu}_0$  onafhankelijk zijn. Gebruik dit voor de berekening van  $Var(Y_0 - \hat{\mu}_0)$ .
- b. Geef het 95%-betrouwbaarheidsinterval voor  $Y_0$ .

We vatten samen:

Het 95%-betrouwbaarheidsinterval voor  $\mu_0$  is:

$$\mu_0 = \hat{\mu}_0 \pm t_{[n-2],0.05} \cdot s \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2}}$$

Het 95%-betrouwbaarheidsinterval voor  $Y_0$  is:

$$Y_0 = \hat{\mu}_0 \pm t_{[n-2],0.05} \cdot s \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2} + 1}$$

### Opgaven

96. Zie opgaven 83 en 94. Geef het 90%-betrouwbaarheidsinterval voor de besparingen van een gezin met een inkomen van 30.000 euro.
97. Hieronder zie je voor de jaren 1947 t/m 1957 het aantal auto's in tienduizenden ( $X$ ) en het aantal verkeersongevallen in duizenden ( $Y$ ) in het Verenigd Koninkrijk.

### 6.8. Als de $X_i$ 's kansvariabelen zijn

$X$	$Y$
352	166
373	153
411	177
441	201
462	216
490	208
529	227
577	238
641	268
692	268
743	274

- Bereken de OLS regressievergelijking. Bekijk een plot om na te gaan of het redelijk is de gangbare veronderstellingen te maken.
  - Geef het 95%-betrouwbaarheidsinterval voor  $\beta$ .
  - In 1958 zijn er 8 miljoen auto's. Geef een 95%-betrouwbaarheidsinterval voor het aantal verkeersongelukken.
  - Wat is de kans dat, bij 8 miljoen auto's, het aantal verkeersongelukken groter is dan 325.000?
  - Geef een schatting van het aantal verkeersongelukken bij 15 miljoen auto's. Welke kanttekening maak je bij dit resultaat?
98. In hoofdstuk 2 hebben we de formule  $\mu = \bar{X} \pm t_{[n-1],0.05} \cdot \frac{s}{\sqrt{n}}$  afgeleid voor het 95%-betrouwbaarheidsinterval voor het gemiddelde van een normaal verdeelde populatie. Deze formule is verwant aan de eerste formule uit het laatste kader van paragraaf 6.7. Bedenk welke formule (over een normaal verdeelde populatie) verwant is aan de tweede formule uit dit kader en geef een interpretatie van die formule.

## 6.8 Als de $X_i$ 's kansvariabelen zijn

Tot dusver hebben we de verklarende variabele  $X$  als een constante beschouwd, waarvan de niveaus niet door het toeval worden bepaald, maar naar believen kunnen worden vastgesteld. Het voorbeeld aan het begin van dit hoofdstuk voldoet goed aan deze veronderstelling. De hoeveelheid kunstmest die wordt gebruikt kun je zelf bepalen. In de praktijk voldoen lang niet alle situaties aan deze eis. Opgave 84 is een mooi voorbeeld. De verklarende variabele  $X$  was daar het aantal minuten dat de atleet het kon volhouden op de loopband. Dit is duidelijk een voorbeeld van een kansvariabele en niet van een instelbare constante. Betekent dit dat de theorie van dit hoofdstuk

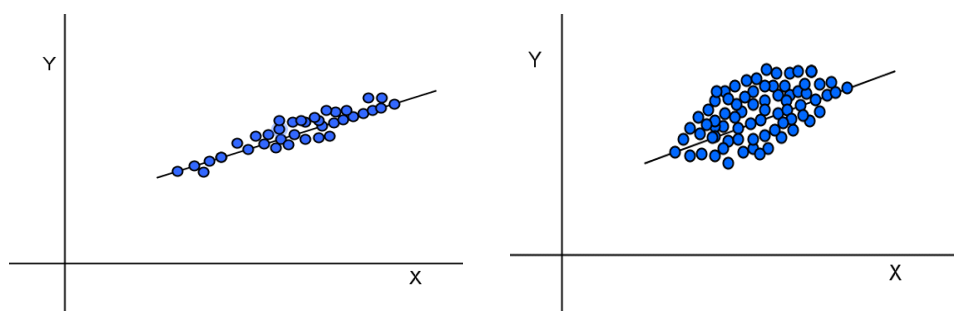
### *Enkelvoudige lineaire regressie*

op dergelijke situaties niet van toepassing is? Gelukkig niet. Gemakshalve kunnen we zeggen dat de resultaten van dit hoofdstuk geldig blijven mits aan de volgende additionele voorwaarden is voldaan:

- De verdeling van  $X$  is wordt niet bepaald door de parameters  $\alpha, \beta$  of  $\sigma^2$
- De  $e_i$ 's zijn onafhankelijk van de  $X_i$ 's.

## Hoofdstuk 7

# Correlatie

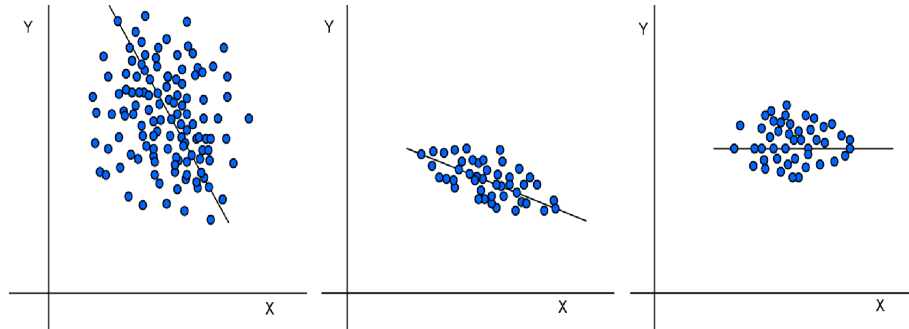


In de plaatjes hierboven zie je twee puntenwolken van de kansvariabelen  $X$  en  $Y$ . (Bijvoorbeeld:  $X$  is de lengte in cm en  $Y$  het gewicht in kg van een groep Nederlandse mannen. Elk punt is de weergave van lengte en gewicht van één man.) Ook de regressielijn van  $Y$  op  $X$  is weergegeven. De formules van de regressielijnen zijn gelijk, toch zijn de plaatjes duidelijk verschillend. In het linkerplaatje liggen de punten veel dichter op de regressielijn dan in het rechterplaatje. Als je in het linkerplaatje de  $X$ -waarde van een punt kent, kun je de bijbehorende  $Y$ -waarde met grote zekerheid voorspellen, in het rechterplaatje is de zekerheid veel minder groot en resteert er, bij een gegeven waarde van  $X$ , een behoorlijke spreiding in de mogelijke waarden van  $Y$ . We vatten het verschil in beide plaatjes samen door te zeggen dat in het linkerplaatje de *correlatie* tussen  $X$  en  $Y$  groter is dan in het rechterplaatje. Later in dit hoofdstuk zullen we exact aangeven hoe je correlatie kunt meten, op dit moment gaat het alleen om de globale betekenis van het woord. In beide plaatjes zeggen we dat  $X$  en  $Y$  *positief gecorreleerd* zijn. Bij grotere waarden van  $X$  horen gemiddeld grotere waarden van  $Y$  en omgekeerd.

In de twee linkerplaatjes hierna is sprake van *negatieve correlatie*. Bij grotere waarden van  $X$  horen gemiddeld kleinere waarden van  $Y$ . Je kunt ook zeggen (het komt namelijk op hetzelfde neer): als de helling van de regressielijn ( $\hat{\beta}$ ) negatief is, spreken we van negatieve correlatie, als de helling positief is, dan spreken we van positieve correlatie. Is de helling

## Correlatie

nul, zoals in het rechterplaatje hieronder, dan is de correlatie ook nul. We zeggen dan ook wel dat er geen correlatie is. Als je de  $X$ -waarde van een punt kent, geeft dat geen enkele informatie over de  $Y$ -waarde.



Waar is de correlatie nu het sterkst (d.w.z. het meest negatief), in het linker of in het middelste plaatje hierboven? Het antwoord op deze vraag is: in het middelste plaatje. Correlatie zegt iets over de mate van lineaire samenhang tussen twee variabelen. De correlatie is sterk als de punten dicht op de regressielijn liggen en zwak als dat niet zo is. Maar de grootte (positief of negatief) van de helling van de regressielijn heeft geen invloed op de grootte van de correlatie. Liggen alle punten exact op de regressielijn, dan is de correlatie *perfect*, gelijk aan  $+1$ , bij positieve en gelijk aan  $-1$  bij negatieve correlatie.

## Opgaven

99. Hieronder zie je een lijst van steeds twee variabelen. Zijn ze volgens jou positief, negatief of niet gecorreleerd?
- Cataloguswaarde en gewicht van personenauto's.
  - Lengte van vader en zoon.
  - Voor middelbare scholieren: tijd per week besteed aan huiswerk en tijd per week besteed aan uitgaan
  - Hoeveelheid zakgeld en leeftijd van middelbare scholieren.
  - Voor een student: prestatie op de 200m sprint en aantal alcoholische consumpties op de dag ervoor.
  - Cijfer voor Nederlands en voor wiskunde van gymnasiumleerlingen op het centraal eindexamen.
  - Leeftijd en bloeddruk van Nederlandse mannen.
  - Gemiddelde temperatuur te De Bilt in januari en juli van eenzelfde jaar.
  - Prijs en vloeroppervlakte van een woning in Amsterdam.
  - Gewicht en lengte van jongens op de basisschool.

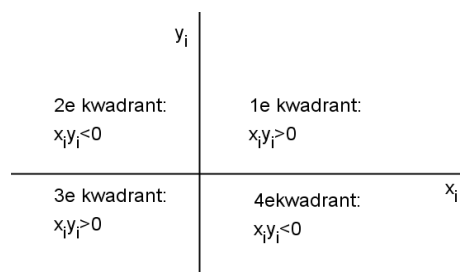


- k. Aantal keren de klas uitgestuurd (in een schooljaar) en gemiddeld cijfer op het eindrapport (van datzelfde schooljaar) van middelbare scholieren.
  - l. Gemiddelde temperatuur op dezelfde dag in Amsterdam en Sydney.
  - m. Leeftijd van man en vrouw in een echtpaar.
  - n. Lengte van de stam en lengte van een blad van eikenbomen.
100. Geef van de volgende uitspraken aan of ze goed of fout zijn:
- a. Als je de eenheid van één van de twee variabelen verandert (bijv. lengte in m in plaats van in cm), dan verandert de helling van de regressielijn ( $\hat{\beta}$ ).
  - b. Als je de eenheid van één van de twee variabelen verandert (bijv. lengte in m in plaats van in cm), dan verandert de mate van correlatie.
  - c. Als twee variabelen onafhankelijk zijn, is er geen correlatie.
  - d. Als twee variabelen perfect gecorreleerd zijn, is de ene een lineaire functie van de andere.
  - e. Als  $Y$  een kwadratische functie is van  $X$ , is de correlatie perfect.

## 7.1 De correlatiecoëfficiënt

Er bestaan verschillende definities van correlatiecoëfficiënt. Wij behandelen hier alleen de correlatiecoëfficiënt van Pearson, die ook wel Pearsons productmoment-correlatiecoëfficiënt genoemd wordt. Als gesproken wordt van correlatiecoëfficiënt zonder verdere toevoeging, wordt altijd deze bedoeld.

Gegeven is een puntenwolk  $(X_i, Y_i)$  ( $i = 1, 2, \dots, n$ ). We definiëren, net als in hoofdstuk 6,  $x_i = X_i - \bar{X}$  en analoog  $y_i = Y_i - \bar{Y}$ . De OLS-regressielijn gaat door het punt  $(\bar{X}, \bar{Y})$ , dus als we naar de puntenwolk  $(x_i, y_i)$  kijken, gaat de regressielijn door de oorsprong.



Als er sprake is van positieve correlatie, gaat de regressielijn door het eerste en derde kwadrant. De meeste punten van de puntenwolk zullen zich dus in die twee kwadranten bevinden. M.a.w. voor de meeste punten zal gelden:  $x_i y_i > 0$ . Als er sprake is van negatieve correlatie zullen de meeste punten van de puntenwolk zich in het 2e en 4e kwadrant bevinden en zal voor de meeste punten gelden:  $x_i y_i < 0$ .

Algemeen: als we voor onze puntenwolk  $\sum x_i y_i$  berekenen dan krijgen we een uitkomst die positief zal zijn als er sprake is van positieve correlatie en omgekeerd.

Toch is  $\sum x_i y_i$  niet de formule voor de correlatiecoëfficiënt waarnaar we op zoek zijn. Om te beginnen is  $\sum x_i y_i$  afhankelijk van het aantal waarnemingen. Als de puntenwolk veel waarnemingen bevat zal deze som groot (positief of negatief) zijn, maar dat zegt niets over de correlatie. We kunnen dit verhelpen door te delen door  $n$ , of, wat gebruikelijker is, door  $n-1$ . We krijgen dan  $s_{XY} = \frac{1}{n-1} \sum x_i y_i = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})$  en deze grootheid wordt ook wel de *covariantie* van  $X$  en  $Y$  genoemd. Covariantie is niet hetzelfde als correlatie. Verderop zullen we zien hoe we correlatie kunnen definiëren met behulp van het begrip covariantie.

### Opgaven

101. Toon aan:  $\sum x_i y_i = \sum X_i Y_i - n \bar{X} \cdot \bar{Y}$
102. Zijn vwo-leerlingen die goed zijn in wiskunde ook goed in natuurkunde? In onderstaande tabel vind je van 9 willekeurige leerlingen

### 7.1. De correlatiecoëfficiënt

het cijfer voor wiskunde B en natuurkunde behaald op het Centraal Eindexamen.

leerling	A	B	C	D	E	F	G	H	I
wiskunde B	7,7	7,2	8,1	5,1	9,5	5,2	6,7	6,8	8,7
natuurkunde	8,2	6,0	7,4	6,0	7,8	6,0	5,7	7,3	7,4

- Bereken in twee decimalen nauwkeurig de covariantie  $s_{XY}$ . (Gebruik het resultaat van opgave 101 en de functie **2VarStats** op de GR.)
- Bereken  $s_X$ ,  $s_Y$  en  $\frac{s_{XY}}{s_X s_Y}$ .
- Reken de cijfers  $X$  en  $Y$  om in percentages (hoogst mogelijke cijfer wordt dan 100 in plaats van 10) en bereken opnieuw  $s_{XY}$ .
- Waarom is de covariantie geen goede correlatiemaat?
- Verklaar hoe je het antwoord van c. ook zonder (GR)-rekenwerk had kunnen vinden.
- Bereken (zonder de GR veel rekenwerk te laten verrichten) met cijfers in percentages:  
 $s_X$ ,  $s_Y$  en  $\frac{s_{XY}}{s_X s_Y}$ . Wat valt je op?

In opgave 102 heb je gezien dat de covariantie  $s_{XY}$  geen goede correlatiemaat is, omdat hij afhangt van de gekozen eenheid. Deel je echter door de standaardafwijkingen  $s_X$  en  $s_Y$ , dan ontstaat een maat die onafhankelijk is van de gebruikte eenheden. Dit leidt tot de definitie van Pearsons product-moment-correlatiecoëfficiënt  $r = \frac{s_{XY}}{s_X s_Y}$ . Uiteraard is deze definitie alleen geldig als  $s_X \neq 0$  en  $s_Y \neq 0$ . In de rest van dit hoofdstuk zullen we steeds aannemen dat aan deze voorwaarden is voldaan.

Samengevat:

Voor een serie waarnemingen  $(X_i, Y_i)$  ( $i = 1, 2, \dots, n$ ) geldt:

$s_{XY} = \frac{1}{n-1} \sum x_i y_i = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})$  is de covariantie

$r = \frac{s_{XY}}{s_X s_Y}$  is de correlatiecoëfficiënt, met

$s_X^2 = \frac{1}{n-1} \sum x_i^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$  en

$s_Y^2 = \frac{1}{n-1} \sum y_i^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$

als steekproefvarianties van  $X$  en  $Y$ .

**Opmerking 7.1.1.** Ga na dat de (steekproef)covariantie symmetrisch is in  $X$  en  $Y$ , d.w.z.  $s_{XY} = s_{YX}$ . De correlatiecoëfficiënt  $r$  is daarom ook symmetrisch in  $X$  en  $Y$ . De regressiecoëfficiënt  $\hat{\beta}$  is niet symmetrisch in  $X$  en  $Y$ .

**Opmerking 7.1.2.** Voor de berekening van de correlatiecoëfficiënt hebben we tot twee keer toe de oorspronkelijke waarnemingen  $(X_i, Y_i)$  getransformeerd: eerst door ons te concentreren op afwijkingen t.o.v. het gemiddelde, vervolgens door deze afwijkingen te delen door hun standaarddeviatie. Dit proces heet het *standaardiseren* van data. De  $\tilde{x}_i = \frac{X_i - \bar{X}}{s_X}$  en  $\tilde{y}_i = \frac{Y_i - \bar{Y}}{s_Y}$  heten de gestandaardiseerde data. Deze data hebben een gemiddelde gelijk aan 0 en een standaardafwijking gelijk aan 1. De correlatiecoëfficiënt van gestandaardiseerde data is gelijk aan die van de oorspronkelijke data.

**Opmerking 7.1.3.** Je kunt de GR zo instellen dat er bij elke regressieberekening de waarde van  $r$  wordt berekend. Je kiest dan **DiagnosticOn**. Je vindt deze opdracht in het **Catalog**-menu.

## 7.2 Een stochastisch model

Net als bij de introductie van lineaire regressie in hoofdstuk 6 hebben we ons bij de introductie van correlatie geconcentreerd op de analyse van waarnemingen zonder ons te bekommeren om de onderliggende werkelijkheid. Als we in opgave 102 de correlatiecoëfficiënt berekenen tussen de cijfers voor wiskunde B en natuurkunde bij negen leerlingen, weten we natuurlijk dat, als we de procedure herhalen bij negen andere leerlingen, de berekende correlatiecoëfficiënt zal afwijken van de eerder gevonden waarde. Welke conclusie kunnen we dan trekken op grond van een enkele serie waarnemingen?

Net als in hoofdstuk 6 zullen we nu onderscheid maken tussen de onderliggende werkelijkheid (het stochastisch model) en de grootheden die we berekenen op grond van een steekproef.

$X$  en  $Y$  zijn kansvariabelen uit een onderliggende populatie. De (populatie)gemiddelden en (populatie)varianties van  $X$  en  $Y$  duiden we aan met respectievelijk:

$$\mu_X = E(X), \mu_Y = E(Y), \sigma_X^2 = Var(X) \text{ en } \sigma_Y^2 = Var(Y)$$

Omdat  $X$  en  $Y$  niet noodzakelijk onafhankelijk zijn, definiëren we hun covariantie als volgt:

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)).$$

Nu kun je zien dat de covariantie  $s_{XY} = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})$  een

## 7.2. Een stochastisch model

schatter is van de covariantie  $Cov(X, Y)$ . Als we duidelijk onderscheid willen maken tussen deze twee soorten covariantie spreken we van steekproefcovariantie respectievelijk populatiecovariantie.

De (populatie)correlatiecoëfficiënt van de kansvariabelen  $X$  en  $Y$  definiëren we als:  $\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$ . Ook hier kun je onmiddellijk zien dat

de (steekproef)correlatiecoëfficiënt  $r = \frac{s_{XY}}{s_X s_Y}$  een schatter is van  $\rho$ . De tabel hieronder geeft een overzicht van een aantal verwante begrippen die betrekking hebben op een populatie of op een steekproef.

	Populatie (of kansvariabele)	Steekproef (of gegevens)
<b>Gemiddelde</b>	$\mu_X$ of $E(X)$	$\bar{X}$
<b>Variantie</b>	$\sigma_X^2$ of $Var(X)$	$s_X^2$
<b>Covariantie</b>	$Cov(X, Y)$	$s_{XY}$
<b>Correlatiecoëfficiënt</b>	$\rho$	$r$

De begrippen uit de rechterkolom (die betrekking hebben op een steekproef) zijn schatters van de overeenkomstige begrippen uit de middelste kolom (die betrekking hebben op de populatie).

In dit hoofdstuk zullen we ons niet bezig houden met de statistische eigenschappen van de schatters  $s_{XY}$  en  $r$  en we zullen dan ook geen betrouwbaarheidsintervallen en toetsen over de werkelijke waarde van  $\rho$  afleiden. In de praktijk volstaan de betrouwbaarheidsintervallen en toetsen over  $\beta$ . In opgave 105 zal dit duidelijk worden.

### Opgaven

103. Om van de (steekproef)covariantie  $s_{XY} = \frac{1}{n-1} \sum x_i y_i$  een correlatiecoëfficiënt te maken, delen we door  $s_X$  en  $s_Y$ . Waarom gebruiken we  $s_X$  en  $s_Y$ , zou je ook  $s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$  en  $s_y = \sqrt{\frac{1}{n-1} \sum (y_i - \bar{y})^2}$  kunnen gebruiken?

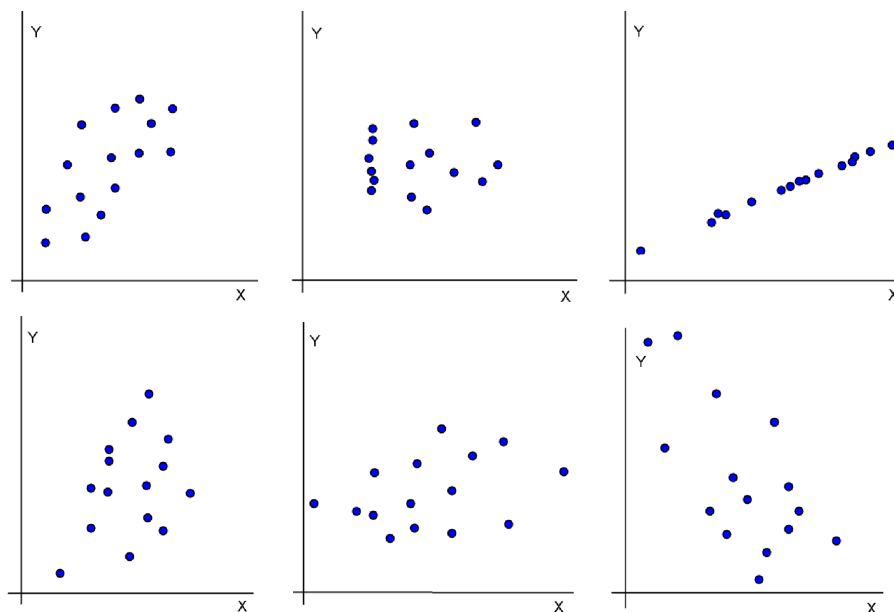
104. Bewijs de volgende formules:

- $\hat{\beta} = r \frac{s_Y}{s_X}$ . (Hierbij is  $\hat{\beta}$  de richtingscoëfficiënt van de OLS-regressielijn van  $Y$  op  $X$ .)
- $\hat{\beta}_X \cdot \hat{\beta}_Y = r^2$ . (Hierbij is  $\hat{\beta}_Y$ , net als  $\hat{\beta}$  bij a., de richtingscoëfficiënt van de OLS-regressielijn van  $Y$  op  $X$  en  $\hat{\beta}_X$  de richtingscoëfficiënt van de OLS-regressielijn van  $X$  op  $Y$ .)

105. Laat zien dat de hypothese  $H_0 : \rho = 0$  equivalent is met  $H_0 : \beta = 0$ .

## Correlatie

106. Zie de spreidingdiagrammen hieronder. Schat bij elk van deze diagrammen de waarde van  $r$ . Kies uit: 1 0,67 0,33 0 -0,33 -0,67 -1.



107. Een klas maakt een proefwerk met 25 vragen. Elk antwoord is óf goed óf fout. De docent zet op elk blaadje het aantal goede antwoorden  $X$  en het aantal foute antwoorden  $Y$ . Hij vindt  $\bar{X} = 18,6$  en  $s_X = 2,9$ . Geef  $\bar{Y}$ ,  $s_Y$  en  $r$ .
108. Op een school doen 22 leerlingen eindexamen in wiskunde B en wiskunde D. In de tabel hieronder zie je wat voor cijfers ze voor deze vakken op hun eindlijst haalden. We willen met de GR de correlatiecoëfficiënt berekenen van het eindcijfer voor de twee vakken.

	wiskunde B					totaal
	6	7	8	9	10	
5		1				1
wis- 6	1	2	4			7
kunde 7			2			2
D 8				6		6
9			1	3	2	6
totaal	1	3	7	9	2	22

- a. Welke gegevens moet je invoeren in L1 en L2 om  $r$  te berekenen? Voer de berekening uit.

Er is ook een eenvoudiger manier om de berekening met de GR uit te voeren. Als je de lijsten L1, L2, L3 vult met respectievelijk het cijfer voor wiskunde D, het cijfer voor wiskunde B en de frequentie van

### 7.3. Interpretatie van $\hat{\beta}$ en $r$

deze cijfercombinatie, geeft `LinReg(a+bx)` L1, L2, L3 de gewenste resultaten.

- b. Bereken  $r$  opnieuw en controleer dat je hetzelfde antwoord krijgt als bij a.
- c. Verschilt de correlatiecoëfficiënt significant van 0? Neem  $\alpha = 0,05$ .

109. Van een puntenwolk  $(X_i, Y_i)$  is gegeven :  $\bar{X} = 5$   $s_X = 2$   $\bar{Y} = 8$   $s_Y = 3$  en  $r = 0,5$ . Stel de regressielijnen op van  $Y$  op  $X$  en van  $X$  op  $Y$ . Gebruik opgave 104.

### 7.3 Interpretatie van $\hat{\beta}$ en $r$

In het voorgaande hebben we gezien dat er een duidelijke samenhang bestaat tussen de begrippen regressie en correlatie en tussen de schatters  $\hat{\beta}$  en  $r$  in het bijzonder. Maar er was ook een verschil in de manier waarop we deze begrippen uiteindelijk introduceerden. We definieerden  $r$  met behulp van de formule  $r = \frac{s_{XY}}{s_X s_Y}$  zonder verdere veronderstellingen te maken behalve de impliciete aanname  $s_X s_Y \neq 0$ . Bij lineaire regressie gingen we in paragraaf 6.4 uit van een stochastisch model waarvoor een aantal veronderstellingen noodzakelijk was. Deze veronderstellingen waren:

$Y_i = \alpha + \beta X_i + e_i$  met de  $e_i$ 's onderling onafhankelijk met  $E(e_i) = 0$  en  $Var(e_i) = \sigma^2$ , waarbij je soms verder aanneemt dat de  $e_i$ 's normaal verdeeld zijn.

Is het nu zo dat je voor regressie wel en voor correlatie geen extra veronderstellingen nodig hebt?

Het is natuurlijk zo dat je  $\hat{\beta}$ , net als  $r$ , uitsluitend via zijn formule  $\hat{\beta} = (\sum x_i Y_i) / \sum x_i^2$  kunt definiëren. Je hebt hiervoor geen aanvullende veronderstellingen nodig, behalve de impliciete voorwaarde  $\sum x_i^2 \neq 0$ . Maar zonder extra aannames heeft de grootte  $\hat{\beta}$  geen betekenis en kun je de uitkomst van de formule niet interpreteren.

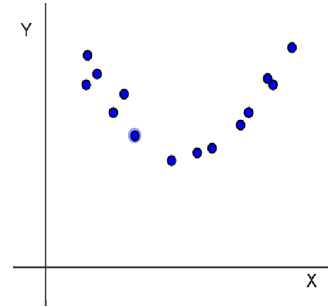
Hetzelfde is waar voor de correlatiecoëfficiënt  $r$ . De correlatiecoëfficiënt is een maat voor de *lineaire* samenhang van  $X$  en  $Y$ . Als er geen sprake is van lineaire samenhang, is de grootte  $r$  zonder betekenis, ook al kun je hem uitrekenen. Voor een zinvolle interpretatie van  $r$  zijn dezelfde veronderstellingen nodig als voor het regressiemodel. Het klakkeloos berekenen van  $\hat{\beta}$  of  $r$ , zonder je af te vragen of aan de onderliggende voorwaarden is voldaan, kan gemakkelijk tot verkeerde conclusies leiden. In de praktijk is het daarom aan te raden om eerst de gegevens te bekijken, zo mogelijk voordat je aan de berekeningen begint. Dikwijls kun je aan een

## Correlatie

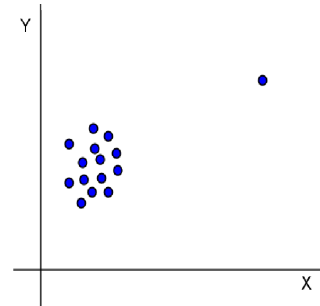
spreidingsdiagram zien of de modelveronderstellingen al dan niet plausibel zijn.

Hieronder zullen we een aantal situaties behandelen waarin de modelveronderstellingen niet zonder meer gemaakt kunnen worden. Dergelijke situaties komen in de praktijk vaak voor.

- A. De figuur hiernaast laat een duidelijk verband zien tussen  $X$  en  $Y$ , maar dit verband is **niet lineair**. Als je klakkeloos  $\hat{\beta}$  en  $r$  berekent, vind je waarden die niet veel van nul verschillen. Als je daaruit de conclusie zou trekken dat er geen verband tussen  $X$  en  $Y$  is, zit je er flink naast. Als je de drie meest rechtse punten buiten beschouwing laat, vind je negatieve waarden van  $\hat{\beta}$  en  $r$ . Laat je de drie meest linkse punten buiten beschouwing, dan worden deze waarden juist positief. In al deze gevallen trek je de verkeerde conclusie.



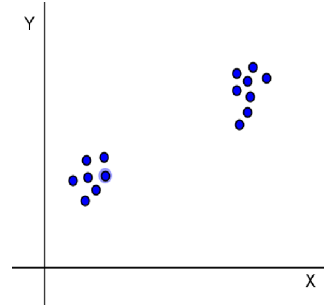
- B. Bij de gegevens uit de figuur hiernaast hoort een correlatiecoëfficiënt van ongeveer 0,7 en een positieve waarde van  $\hat{\beta}$ . Maar je ziet meteen dat deze resultaten worden veroorzaakt door één enkel punt, een zogenaamde **uitbijter** of **outlier**. Zonder dit punt vind je waarden van  $\hat{\beta}$  en  $r$  van ongeveer nul. Je moet je dus afvragen of de uitbijter werkelijk tot de populatie behoort die je bestudeert, voordat je een conclusie kunt trekken.



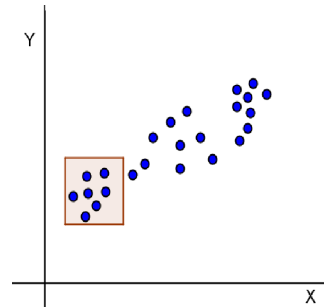


### 7.3. Interpretatie van $\hat{\beta}$ en $r$

- C. Bij de gegevens uit de figuur hiernaast hoort een correlatiecoëfficiënt van boven de 0,9 en een sterk significant positieve waarde van  $\hat{\beta}$ . Maar als je de gegevens opvat als steekproeven uit twee **verschillende populaties** en deze apart analyseert dan vind je waarden van  $\hat{\beta}$  en  $r$  in de buurt van nul. Welke conclusie is de juiste? Zonder additionele informatie geven de gegevens alleen onvoldoende houvast om een lineair verband te veronderstellen, dus is het verstandiger om uit te gaan van twee verschillende populaties. Dat zou bijvoorbeeld betekenen dat je geen uitspraak kunt doen over eventuele toekomstige waarnemingen met een  $X$ -waarde die tussen de twee clusters in ligt.



- D. In de figuur hiernaast moet je je voorstellen dat je alleen de waarnemingen in het vierkant analyseert. Je zou dan tot de conclusie komen dat  $X$  en  $Y$  niet gecorreleerd zijn, terwijl er in feite sprake is van (sterk) positieve correlatie. We hebben hier te maken met een situatie die het tegenovergestelde is van die in C. Bij C spraken we over het risico dat je twee verschillende populaties in één model samenvoegt. Hier bij D wordt gewezen op het gevaar om naar een **deelpopulatie** te kijken in plaats van het geheel. Bijvoorbeeld: je onderzoekt het verband tussen leeftijd ( $X$ ) en lichaamslengte ( $Y$ ) van pubers. Als je alleen gegevens verzamelt over 12-jarigen (het vierkantje) zul je minder duidelijk zien dat  $X$  en  $Y$  gecorreleerd zijn dan wanneer je personen uit je steekproef de leeftijden 12 t/m 18 beslaan.



## Correlatie

Je moet onderscheid maken tussen (lineaire) samenhang en eventueel **oorzakelijk verband** van twee variabelen. Beide zaken kunnen samenhangen, maar dat hoeft niet zo te zijn. En als er sprake is van een oorzakelijk verband, dan kan de correlatiecoëfficiënt niet vertellen welke factor de oorzaak is en welke het gevolg. Een bekend voorbeeld is de sterke positieve correlatie die in de periode 1945-1975 bestond tussen het aantal geboortes per volwassen Nederlandse vrouw en de ooievaarsstand in Nederland, want beide grootheden daalden gestaag in die periode. Zouden ooievaars dan toch iets te maken hebben met geboorte van kinderen?

- E. Ook bestaat er bij mannen een negatieve correlatie tussen de dichtheid van haargroei en de frequentie van hartklachten. Zou haaruitval dan leiden tot hartklachten en zou je hartklachten kunnen bestrijden met een haargroeimiddel? Het antwoord op deze vragen is natuurlijk “nee”. De ooievaar brengt geen baby’s en er bestaat geen oorzakelijk verband tussen haargroei en hartklachten. Er is bij mannen wel een oorzakelijk verband tussen leeftijd en haargroei en tussen leeftijd en hartklachten. Als je de factor leeftijd over het hoofd ziet, kun je tot foutieve conclusies komen. Dit verschijnsel wordt ook wel *confounding* genoemd.

## Opgave

110. Hieronder zie je een lijst met situaties waarin een correlatiecoëfficiënt wordt berekend. Geef in elk van deze situaties aan welk van de bovenstaande problemen er mogelijk speelt. Vermeld zonodig ook de confounding variabele.
- Bij een onderzoek naar de effectiviteit van aandelenhandelaren van een bank wordt de correlatie berekend tussen de jaarlijks geboekte handelswinst en de leeftijd van de handelaar.
  - Ten behoeve van een onderzoek naar de financiële armslag van scholieren wordt op een school een enquête gehouden. De enquêteur komt een lokaal binnen en geeft alle aanwezigen een formulier waarop o.a. de leeftijd en de inhoud van de portemonnee (in euro’s) moet worden ingevuld (anoniem). Hiermee wordt de correlatie berekend tussen leeftijd en de hoeveelheid direct te besteden geld.
  - Uit een uitgebreide enquête onder eerstejaarsstudenten blijkt een positieve correlatie te bestaan tussen de lichaamslengte (in cm) en de alcoholconsumptie (in aantal glazen per week).
  - Voor een studie naar de relatie tussen inkomen en medicijngebruik wordt in de wachtkamer van een huisarts een stapel enquêteformulieren neergelegd. Aan de wachtenden wordt gevraagd een formulier in te vullen en in een daarvoor bestemde bus te deponeren.

#### 7.4. Interpretatie van de correlatiecoëfficiënt

- e. Bij meisjes van een middelbare school bestaat een positieve correlatie tussen het gebruik van orale anticonceptiva en het inkomen (uit zakgeld plus bijbaantjes).
- f. In een specifieke wijk van een stad bestaat een positieve correlatie tussen het inkomen en de consumptie van varkensvlees en een negatieve correlatie tussen het inkomen en de consumptie van lamsvlees. Toch is lamsvlees duurder dan varkensvlees.

### 7.4 Kwantitatieve interpretatie van de correlatiecoëfficiënt bij regressieanalyse

Als we een OLS-regressielijn  $\hat{Y} = \hat{\alpha} + \hat{\beta}X$  geschat hebben voor een puntenwolk  $(X_i, Y_i)$  ( $i = 1, 2, \dots, n$ ), kunnen we schrijven:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Dit is een triviale gelijkheid. Minder triviaal is het feit (en je gaat het bewijzen in opgave 111) dat deze gelijkheid ook geldt als je de individuele termen kwadrateert en de som neemt over de  $i$ 's:

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \quad (*)$$

$\sum (Y_i - \bar{Y})^2 = (n-1)s_Y^2$  is een kwadraatsom die je kunt opvatten als een maat voor spreiding van de  $Y_i$ 's.

Als je bedenkt dat  $\bar{\hat{Y}} = \bar{Y}$ , zul je inzien dat  $\sum (\hat{Y}_i - \bar{Y})^2$  een kwadraatsom is die je kunt opvatten als een maat van de *door het regressiemodel verklaarde spreiding van de  $Y_i$ 's*, want de  $\hat{Y}_i$ 's zijn functies van de  $X_i$ 's.

$\sum (Y_i - \hat{Y}_i)^2 = (n-2)s_e^2$  is een kwadraatsom die je kunt opvatten als een maat van de *niet door het regressiemodel verklaarde spreiding van de  $Y_i$ 's*.

Vergelijking (\*) splitst dus de kwadraatsom van de spreiding van de  $Y_i$ 's in een deel dat wel en een deel dat niet door het regressiemodel kan worden verklaard. In opgave 112 ga je bewijzen dat  $\frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = r^2$ .

Met andere woorden:

$r^2$  is het deel van de kwadraatsom van de spreiding van de  $Y_i$ 's dat verklaard wordt door de  $X_i$ 's (of door het regressiemodel).

#### Opgaven

111. a. Toon aan dat  $\bar{\hat{Y}} = \bar{Y}$ . (Aanwijzing: gebruik het feit dat de OLS regressielijn door het punt  $(\bar{X}, \bar{Y})$  gaat, ofwel  $\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$ ).
- b. Leg uit dat (\*) equivalent is aan  $\sum y_i^2 = \sum (y_i - \hat{y}_i)^2 + \sum \hat{y}_i^2$ .

### Correlatie

c. Leg uit dat  $\hat{y}_i = \hat{\beta}x_i$ .

d. Bewijs nu de gelijkheid uit b. (en daarmee (\*)).

112. In deze opgave ga je bewijzen dat  $\frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = r^2$ . Er geldt:

$$\frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2}. \text{ Maak het bewijs af.}$$

## Hoofdstuk 8

# Meervoudige lineaire regressie

In de regressiemodellen die we tot dusver zijn tegengekomen hadden we te maken met twee variabelen:  $X$  en  $Y$ .  $X$  was de verklarende of onafhankelijke variabele en  $Y$  de verklaarde of afhankelijke. In het eerste voorbeeld uit hoofdstuk 6 was  $Y$  de opbrengst per hectare van een stuk land, terwijl  $X$  de hoeveelheid gebruikte kunstmest per hectare was. Je kunt je voorstellen dat er, naast de hoeveelheid gebruikte kunstmest, nog een andere factor is die de opbrengst beïnvloedt, zoals de hoeveelheid neerslag (in mm). We noemen deze nieuwe variabele  $Z$ . Als we over de gegevens van de hoeveelheid neerslag beschikken, kunnen we misschien een betere verklaring geven van de opbrengst. Ons model wordt nu:

$$Y_i = \alpha + \beta X_i + \gamma Z_i + e_i, \text{ waar } e_i \text{ de storingsterm voorstelt. } (*)$$

In dit geval is  $Y$  nog steeds de verklaarde variabele, maar we hebben nu twee verklarende variabelen:  $X$  en  $Z$ . Bovendien hebben we verondersteld dat de relatie tussen  $Y$  en elk van de verklarende variabelen lineair is.

We kunnen nog verder gaan. Misschien hebben we ook het aantal uren zonneschijn nodig als verklarende variabele, of een kenmerk dat iets zegt over de vruchtbaarheid van de grond. Zo breiden we ons model uit met steeds meer verklarende variabelen en wordt het algemene model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + e_i. \quad (**)$$

We hebben hier van doen met een regressievergelijking met  $k$  variabelen  $X_1, \dots, X_k$ . Een dergelijk model heet een *meervoudig regressiemodel* (*multiple regression model*). De analyse ervan lijkt sterk op die van het enkelvoudige regressiemodel.

Bij enkelvoudige regressie konden we alle formules voor OLS-schatters e.d. uitschrijven, maar het uitwerken van een getalvoorbeeld werd al snel

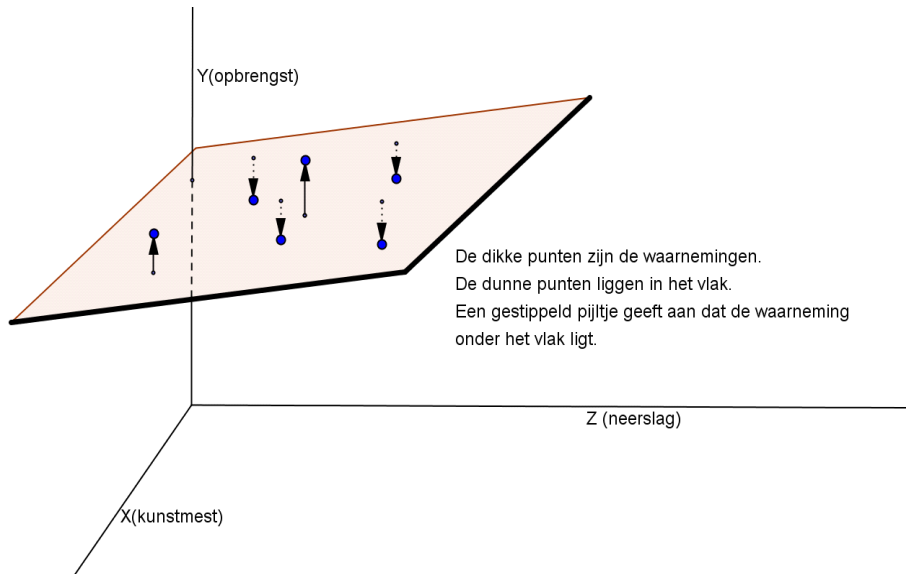
te veel werk zonder GR. Bij meervoudige regressie is ook het uitschrijven van de formules (te) ingewikkeld. We zullen in dit hoofdstuk daarom alleen de principes bespreken met behulp waarvan je de formules zou kunnen uitwerken (als je veel tijd, geduld en niets anders te doen had). De GR kent geen standaardfunctie voor meervoudige regressie, dus voor het uitwerken van getalvoorbeelden hebben we een computer met statistische software nodig.

## 8.1 Een model met twee verklarende variabelen

We keren terug naar model (\*) waarin  $Y$ ,  $X$  en  $Z$  respectievelijk opbrengst, hoeveelheid kunstmest en hoeveelheid neerslag voorstellen. We hadden er ook voor kunnen kiezen om aansluiting te zoeken met het algemene model (\*\*). Dan zou de hoeveelheid kunstmest worden aangeduid met  $X_1$  en de hoeveelheid neerslag met  $X_2$ . Inhoudelijk is er natuurlijk geen enkel verschil. Op dit moment geven we er de voorkeur aan om te werken met vergelijking (\*). Op deze manier vermijden we het werken met dubbele subscripten. Het is echter belangrijk om je te realiseren dat  $X$  en  $Z$  beide verklarende variabelen zijn en  $Y$  de verklaarde.

In hoofdstuk 6 konden we het model met één verklarende variabele mooi grafisch weergeven door middel van een regressielijn door een puntenwolk in een  $OXY$ -assenstelsel. Nu wordt dit een stuk lastiger. We zullen een ruimtelijk  $OXZY$ -assenstelsel moeten gebruiken. De waarnemingen vormen een puntenwolk in dit assenstelsel en de regressievergelijking  $Y = \alpha + \beta X + \gamma Z$  wordt voorgesteld door een vlak dat door de puntenwolk loopt. In de figuur hierna zie je hiervan een plaatje. Je zult direct begrijpen dat, als we een derde verklarende variabele zouden toevoegen, het onmogelijk wordt om de situatie in een plaatje weer te geven, want dan zou je in vier dimensies moeten kunnen tekenen. Maar ook in drie dimensies is het plaatje moeilijker te lezen dan zijn tweedimensionale equivalent. Kun je “op het oog” inschatten of het getekende vlak goed past bij de puntenwolk?

### 8.1. Een model met twee verklarende variabelen



Het is goed om stil te staan bij de betekenis van de parameters  $\alpha, \beta, \gamma$ . (In het algemene model  $(**)$  zouden de parameters  $\beta_0, \beta_1, \beta_2$  heten.)  $\beta$  geeft aan met hoeveel de opbrengst wordt vermeerderd als je één eenheid kunstmest per hectare toevoegt, uitgaande van een gelijkblijvende hoeveelheid neerslag. Het is dus de helling van de lijn die een punt doorloopt als het over het regressievlak beweegt in een richting evenwijdig aan het  $OXY$ -vlak. Op dezelfde manier geeft  $\gamma$  aan met hoeveel de opbrengst wordt vermeerderd als er één eenheid extra neerslag valt, uitgaande van een gelijkblijvende hoeveelheid kunstmest.  $\gamma$  is dus de helling van de lijn die een punt doorloopt als het over het regressievlak beweegt in een richting evenwijdig aan het  $OZY$ -vlak.  $\alpha$  is het snijpunt van het vlak met de  $Y$ -as, want daar zijn  $X$  en  $Z$  beide gelijk aan 0.

In de praktijk zijn de waarden van  $\alpha, \beta, \gamma$  onbekend en willen we ze schatten met behulp van de waarnemingen  $(X_i, Z_i, Y_i)$  ( $i = 1, 2, \dots, n$ ). We gebruiken het principe van de kleinste kwadraten om de schatters te bepalen. Met andere woorden we zoeken waarden van  $a, b, c$  zó dat  $S = \sum (Y_i - a - bX_i - cZ_i)^2$  minimaal is.

Als je  $S$  uitschrijft kun je zien dat, of je  $S$  nu opvat als functie van  $a, b$  of  $c$ ,  $S$  in alle drie de gevallen een dalparabool is. Door de afgeleiden van  $S$  naar elk van de drie parameters gelijk te stellen aan 0, ontstaat het volgende stelsel vergelijkingen:

$$\begin{cases} na + b \sum X_i + c \sum Z_i - \sum Y_i = 0 \\ a \sum X_i + b \sum X_i^2 + c \sum X_i Z_i - \sum X_i Y_i = 0 \\ a \sum Z_i + b \sum X_i Z_i + c \sum Z_i^2 - \sum Z_i Y_i = 0 \end{cases} \quad (*)$$

## Meervoudige lineaire regressie

Dit stelsel ziet er misschien ingewikkeld uit, maar het is eigenlijk heel simpel. Alle grootheden met een  $\sum$ -teken zijn getallen die je kunt uitrekenen met behulp van de waarnemingen. Als je die getallen invult in (\*), resteert een stelsel van drie lineaire vergelijkingen in  $a$ ,  $b$  en  $c$ , dat je eenvoudig kunt oplossen. Deze oplossing duiden we aan met  $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ .  $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$  zijn de OLS-schatters van de parameters  $\alpha, \beta, \gamma$ . Expliciete formules laten we hier achterwege. In de praktijk gebruik je statistische software. Hier gaat het erom dat je inziet dat de berekening van OLS-schatters neerkomt op het oplossen van een stelsel lineaire vergelijkingen.

**Opmerking 8.1.1.** Bij het enkelvoudige regressiemodel hebben we eerst het model herschreven door de  $X_i$ 's te vervangen door de  $x_i$ 's. Dat deden we om het rekenwerk te vereenvoudigen. In bovenstaand geval hebben we daarvan afgezien, omdat we het uiteindelijk rekenwerk toch aan de computer overlaten. Als je alle berekeningen met de hand zou uitvoeren, is het ook hier verstandig om het model eerst te herschrijven in termen van  $x_i$ 's en  $z_i$ 's.

De schatters  $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$  zijn alle een lineaire combinatie van de  $Y_i$ 's. Dit kun je inzien door nog eens goed naar (\*) te kijken. We beschouwen de  $Y_i$ 's als kansvariabelen; de  $X_i$ 's en  $Z_i$ 's zijn vaste getallen. Bij het oplossen van (\*) hoef je nooit te delen door een uitdrukking waar  $Y_i$ 's in voorkomen.

Omdat  $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$  lineaire combinaties zijn van de  $Y_i$ 's, kun je hun verwachtingswaarden en hun varianties berekenen. Voor de verwachtingswaarden geldt  $E(\hat{\alpha}) = \alpha$ ,  $E(\hat{\beta}) = \beta$  en  $E(\hat{\gamma}) = \gamma$ . (We laten de omslachtige algebra achterwege.) We hebben dus te maken met zuivere schatters.

De varianties van  $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$  zijn evenredig met  $\sigma^2$ . Dat wil zeggen dat deze varianties gelijk zijn aan  $\sigma^2$  maal een constante die alleen afhangt van de  $X_i$ 's en  $Z_i$ 's. Ook hier laten we de berekening van deze constanten over aan statistische software.

De schatter van  $\sigma^2$  is:

$$s^2 = \frac{1}{n-3} \sum \hat{e}_i^2 = \frac{1}{n-3} \sum (Y_i - \hat{\alpha} - \hat{\beta}X_i - \hat{\gamma}Z_i)^2.$$

Als de  $e_i$ 's normaal verdeeld zijn, dan heeft  $\frac{(n-3)s^2}{\sigma^2}$  een  $\chi^2$ -verdeling met  $n-3$  vrijheidsgraden.  $s^2$  is dan een zuivere schatter van  $\sigma^2$ . Je ziet dat we drie vrijheidsgraden “kwijtraken” omdat we voor de berekening van  $s^2$  schattingen van drie andere parameters ( $\alpha, \beta, \gamma$ ) nodig hebben.

Met behulp van  $s^2$  zijn de varianties van  $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$  te schatten. We duiden de bijbehorende schattingen van de standaarddeviaties aan met



## 8.2. Het algemene geval met $k$ verklarende variabelen

$s_\alpha, s_\beta, s_\gamma$ . Er geldt nu dat  $\frac{\hat{\alpha} - \alpha}{s_\alpha}$  een t-verdeling heeft met  $n - 3$  vrijheidsgraden. Hetzelfde geldt voor  $\frac{\hat{\beta} - \beta}{s_\beta}$  en  $\frac{\hat{\gamma} - \gamma}{s_\gamma}$ . Deze eigenschap kun je gebruiken om op de gebruikelijk wijze hypothesen te toetsen en betrouwbaarheidsintervallen op te stellen.

Voorts berekenen we de verhouding tussen de totale kwadraatsom en de door het regressiemodel verklaarde kwadraatsom:

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{verklaarde kwadraatsom}}{\text{totale kwadraatsom}}$$

Je kunt aantonen dat  $R$  gelijk is aan de correlatiecoëfficiënt van  $Y$  en  $\hat{Y}$ .

### Opgave

113. Laat zien dat het stelsel (\*) volgt uit  $\frac{dS}{da} = 0$ ,  $\frac{dS}{db} = 0$  en  $\frac{dS}{dc} = 0$ .

## 8.2 Het algemene geval met $k$ verklarende variabelen

De resultaten van paragraaf 8.1 laten zich eenvoudig generaliseren. We vatten samen:

- Het algemene model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + e_i$$

Hierin is  $Y$  de verklaarde of afhankelijke variabele en  $X_1, \dots, X_k$  zijn de verklarende of onafhankelijke variabelen.  $e$  is de storingsterm. De  $e_i$ 's zijn onderling onafhankelijke stochasten, hebben verwachting 0 en een gemeenschappelijke variantie  $\sigma^2$ . Voor het uitvoeren van toetsen of het opstellen van betrouwbaarheidsintervallen wordt daarnaast verondersteld dat ze normaal verdeeld zijn. Als de variabelen  $X_1, \dots, X_k$  geen constanten zijn, maar kansvariabelen, dan zijn de volgende aanvullende veronderstellingen nodig:

(i) de verdeling van  $X_1, \dots, X_k$  wordt niet bepaald door de parameters  $\beta_0, \beta_1, \dots, \beta_k$  en (ii) de  $e_i$ 's zijn onafhankelijk van  $X_1, \dots, X_k$ .

- De OLS-schatters van  $\beta_0, \beta_1, \dots, \beta_k$  worden gevonden door de kwadraatsom

$$S = \sum (Y_i - b_0 - b_1 X_{1i} - \dots - b_k X_{ki})^2$$

## Meervoudige lineaire regressie

te minimaliseren.  $S$  is een dalparabool in elk van de  $b_0, b_1, \dots, b_k$ . Het minimum wordt gevonden door voor  $j = 1, \dots, k$ ,  $\frac{dS}{db_j}$  gelijk te stellen aan 0. Zo ontstaat een stelsel van  $k$  lineaire vergelijkingen in  $k$  onbekenden  $b_0, b_1, \dots, b_k$ . Als regel heeft dit stelsel een unieke oplossing, die wordt aangeduid met  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ .

- De schatters  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  zijn lineaire functies van de  $Y_i$ 's en daarom normaal verdeeld. Het zijn zuivere schatters, want voor  $j = 1, \dots, k$  geldt  $E(\hat{\beta}_j) = \beta_j$ .  $Var(\hat{\beta}_j)$  is evenredig met  $\sigma^2$ .
- De gebruikelijke schatter van  $\sigma^2$  is:

$$s^2 = \frac{1}{n - k - 1} \sum \hat{e}_i^2 = \frac{1}{n - k - 1} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_k X_{ki})^2$$

Als de  $e_i$ 's normaal verdeeld zijn, dan heeft  $\frac{(n - k - 1)s^2}{\sigma^2}$  een  $\chi^2$ -verdeling met  $n - k - 1$  vrijheidsgraden. Het aantal vrijheidsgraden kan als volgt worden gevonden: het aantal waarnemingen moet worden verminderd met het aantal parameters dat moet worden geschat voordat de parameter  $\sigma^2$  kan worden geschat. In dit geval zijn er  $n$  waarnemingen. Voordat  $\sigma^2$  kan worden geschat, moet eerst een schatting worden gemaakt van  $\beta_0$  t/m  $\beta_k$ , d.w.z. van  $k + 1$  andere parameters. Het aantal vrijheidsgraden is dus gelijk aan  $n - (k + 1) = n - k - 1$ .

- Met behulp van  $s^2$  is de schatter van  $Var(\hat{\beta}_j)$  te berekenen. We duiden deze schatter aan met  $s_{\hat{\beta}_j}^2$ . Er geldt nu dat  $\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}}$  een  $t$ -verdeling heeft met  $n - k - 1$  vrijheidsgraden. Deze eigenschap kan worden gebruikt bij het toetsen van hypothesen en het opstellen van betrouwbaarheidsintervallen.
- De verhouding tussen de totale kwadraatsom en de door het model verklaarde kwadraatsom:

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\text{verklaarde kwadraatsom}}{\text{totale kwadraatsom}}$$

geeft een indicatie over de verklarende kracht van het regressiemodel. Je kunt aantonen dat  $R$  gelijk is aan de correlatiecoëfficiënt van  $Y$  en  $\hat{Y}$ .

### 8.3 Een voorbeeld

Een medisch onderzoeker heeft in een ontwikkelingsland uit enkele platte-landsdorpen ruim 1000 mensen willekeurig geselecteerd. Onder meer werden de bloeddruk, het lichaamsgewicht, de leeftijd en de polsfrequentie gemeten. Uit deze groep hebben we aselekt een steekproef getrokken van 31 individuen. De gegevens zijn schematisch weergegeven in de onderstaande tabel:

Y	X1	X2	X3
115	40	24	80
120	47	22	68
148	76	52	92
113	44	65	84
etc	etc	etc	etc

Hierin is :  $Y$  de bloeddruk (mmHg),  $X_1$  het lichaamsgewicht (kg),  $X_2$  de leeftijd (jaren) en  $X_3$  de polsfrequentie (slagen per minuut).

We gebruiken het volgende lineaire model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

Verwerking van deze gegevens met de statistische software R geeft de volgende output:

```

Residuals:
    Min       1Q   Median       3Q      Max
-21.4630   -8.7266    0.5997    5.6449   25.4904

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.2584     20.7832   2.226  0.0346 *
Gewicht       0.4917      0.2069   2.376  0.0248 *
Leeftijd      0.1628      0.1850   0.880  0.3865
Polsfreq      0.5357      0.2489   2.152  0.0405 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.14 on 27 degrees of freedom
Multiple R-squared:  0.3851, Adjusted R-squared:  0.3168
F-statistic: 5.636 on 3 and 27 DF,  p-value: 0.003920
```

Het blokje **Coefficients** bevat de belangrijkste informatie.

- In de kolom “Estimate” staan de OLS-schatters  $\hat{\beta}_0, \dots, \hat{\beta}_k$ . Met “Intercept” wordt de constante term  $\beta_0$  bedoeld. De geschatte OLS-vergelijking is dus:

$$Y = 46,26 + 0,4917X_1 + 0,1628X_2 + 0,5357X_3 + e$$

## Meervoudige lineaire regressie

- De kolom “Std. Error” geeft de bijbehorende schatting van de standaarddeviatie ( $s_{\beta_j}$ ).
- De kolom “t value” is gelijk aan “Estimate” gedeeld door “Std. Error”.  $t = \hat{\beta}_j / s_{\beta_j}$ . Aangezien  $T = (\hat{\beta}_j - \beta_j) / s_{\beta_j}$  een t-verdeling heeft met 27 vrijheidsgraden, kan de waarde van  $t$  worden gebruikt om  $H_0 : \beta_j = 0$  te toetsen. Op een vergelijkbare manier kun je de waarden uit de kolommen “Estimate” en “Std. Error” gebruiken om andere hypothesen te toetsen of om een betrouwbaarheidsinterval voor  $\beta_j$  te construeren. In opgaven 115 en 116 wordt gevraagd dit uit te werken.
- De kolom “Pr(>|t|)” geeft de p-waarde die hoort bij  $H_0 : \beta_j = 0$  en een tweezijdig alternatief. Zo zien we in dit voorbeeld dat  $\hat{\beta}_1$  en  $\hat{\beta}_3$  beide significant van 0 verschillen, maar dat dit voor  $\hat{\beta}_2$  niet het geval is. (Ook de “intercept”  $\hat{\beta}_0$  verschilt significant van 0, maar dat is in dit voorbeeld niet zo relevant.)

Het blokje **Residuals** geeft informatie over de residuen  $\hat{e}_i = Y_i - \hat{Y}_i$ . De mediaan van deze residuen is 0,6 en verschilt niet veel van het gemiddelde dat altijd nul is. Bij een perfect symmetrische verdeling van de residuen zijn de waarden van het 1e en het 3e kwartiel op het teken na aan elkaar gelijk. Hier vinden we -8,7 en 5,6. Deze waarden zijn in absolute waarde niet zodanig verschillend dat we de veronderstelling van een normale (en dus symmetrische) verdeling van de storingstermen  $e_i$  in twijfel hoeven te trekken. Hetzelfde geldt voor de waarden van het grootste en het kleinste residu.

In het **onderste blokje** wordt met “Residual standard error”  $s$ , de schatting van  $\sigma$ , bedoeld. “Multiple R-squared” is  $R^2$ . Dit model slaagt er dus in om 38,5% van de totale kwadraatsom met behulp van de variabelen te verklaren. Je moet hierbij wel bedenken dat we 31 waarnemingen proberen te verklaren met behulp van 3 verklarende variabelen. Er is een vuistregel die zegt dat je voor elke verklarende variabele tenminste 10 waarnemingen moet hebben. Aan deze vuistregel wordt in dit voorbeeld niet voldaan. De “Adjusted R-squared” is een variant op de “Multiple R-squared”. De laatste regel, die begint met “F-statistic” geeft het resultaat van de toets of alle parameters gelijktijdig gelijk zijn aan 0. We gaan op deze laatste twee begrippen nu niet verder in.

## Opgaven

114. Verklaar waarom de “t values” uit de computeroutput hiervoor 27 vrijheidsgraden hebben. Ga met de GR na dat dit aantal vrijheidsgraden in overeenstemming is met de p-waarden in de laatste kolom van de tabel.

#### 8.4. Is het model adequaat?

115. Geef een 95% betrouwbaarheidsinterval voor  $\hat{\beta}_1$  in voorgaand voorbeeld.
116. Zie voorgaand voorbeeld.  
Geef de p-waarde van de toets  $H_0 : \beta_3 = 0, 2$  tegen  $H_1 : \beta_3 > 0, 2$

### 8.4 Is het model adequaat?

Bij enkelvoudige regressie hebben we gezien dat het belangrijk is om na te gaan of de gemaakte veronderstellingen realistisch zijn. We deden dat door naar plaatjes te kijken, waarbij we controleerden of de puntenwolk redelijk rond de regressielijn gespreid lag. Ook controleerden we of de residuen redelijk “willekeurig” leken en geen systematisch patroon vertoonden.

Bij meervoudige regressie is een controle van het onderliggende model eveneens belangrijk. Het maken van plaatjes is ingewikkelder, omdat we daarvoor een meerdimensionale realiteit moeten reduceren tot een tweedimensionaal plaatje. Met de meeste statistische softwarepakketten kunnen dergelijke plaatjes eenvoudig worden gemaakt. Nuttig zijn met name plaatjes waarin de residuen worden weergegeven als functie van elk van de verklarende variabelen.

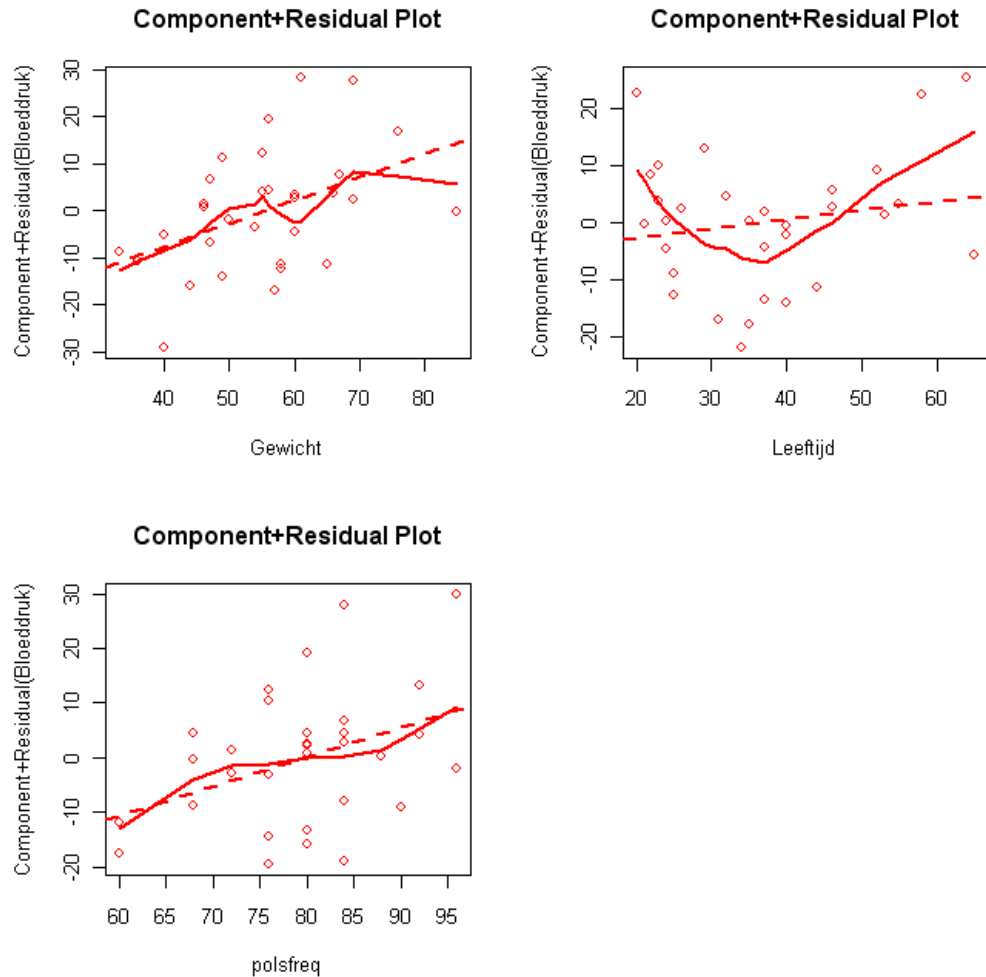
Hierna zie je dergelijke plaatjes voor het voorbeeld van paragraaf 8.3, gemaakt met behulp van R. In deze plaatjes moet je vooral kijken naar de punten zelf. Op de twee lijnen (gestippeld en doorgetrokken) komen we zo terug. Als de gemaakte veronderstellingen realistisch zijn, dan zijn de waargenomen residuen  $\hat{e}_i$  een goede schatting van de werkelijke storingstermen  $e_i$ . Deze zijn volgens het model onderling onafhankelijk, met verwachting 0 en met gelijke variantie  $\sigma^2$ . Deze eigenschappen moet je terugzien in je plot van de residuen. Als je een duidelijk patroon in de residuen ziet, dat niet door louter toeval lijkt te zijn veroorzaakt, is dat een aanwijzing dat het model niet adequaat is.

De twee lijnen (gestippeld en doorgetrokken) zijn hulpmiddelen bij het ontdekken van een eventueel patroon in de residuen en komen tot stand d.m.v. een “regressie-achtige” techniek waar we nu niet verder op ingaan. In de ideale situatie zijn deze beide lijnen recht en horizontaal. Als je de plaatjes beoordeelt, moet je je afvragen of het door de computer “ontdekte” patroon reëel is of dat het net zo goed door toeval zou kunnen zijn veroorzaakt.

Bij het plaatje “Gewicht” suggereren de lijnen dat er een stijgend patroon in de residuen zit. Erg overtuigend is dit niet. Als je je vinger legt op de drie meest linkse punten, blijft er van de stijging weinig over. Min of meer hetzelfde geldt voor het plaatje “Polsfrequentie”. Bij het plaatje “Leeftijd” ligt het anders. Je ziet hier dat de negatieve residuen vooral voorkomen in de leeftijdsgroep 25-45; voor jongere en oudere mensen zijn de residuen vooral positief. Dit geeft aan dat het veronderstelde verband

## Meervoudige lineaire regressie

tussen bloeddruk en leeftijd niet lineair is. In paragraaf 8.5 zullen we bespreken welke consequenties je hieraan kunt verbinden.



## 8.5 Selectie van verklarende variabelen

In het voorbeeld van paragraaf 8.3 gebruikten we een model dat voor de onderzochte populatie de bloeddruk verklaart met behulp van drie verklarende variabelen: gewicht, leeftijd en polsfrequentie. Kennelijk hadden de onderzoekers het idee dat dit een zinvol model is; misschien was er ook een andere reden om waarnemingen te verzamelen over deze variabelen. Als we naar de uitkomsten van de regressieanalyse kijken, zien we dat er een positief verband is gevonden tussen de bloeddruk en het lichaamsgewicht. Geschat wordt dat voor iedere extra kg lichaamsgewicht de bloeddruk met  $\hat{\beta}_1 \approx 0,4917$  mmHg stijgt. Deze geschatte parameter mag dan voorzien

### 8.5. Selectie van verklarende variabelen

zijn van een flinke standaardfout (0,2068), de gevonden waarde verschilt significant van 0, dus je mag concluderen dat de bloeddruk met toenemend lichaamsgewicht ook toeneemt. Toch moet je voorzichtig zijn met een dergelijke conclusie. De gegevens laten alleen zien dat er een significante *samenhang* is tussen bloeddruk en lichaamsgewicht, maar dit zegt niets over een *causaal verband*. Het kan best zo zijn er een andere oorzaak is die de bloeddruk beïnvloedt en dat deze oorzaak vaker of in heviger mate voorkomt bij mensen met een hoog lichaamsgewicht.

We zien ook dat  $\hat{\beta}_2$ , de bij “leeftijd” behorende coëfficiënt, niet significant van 0 verschilt. Je mag hieruit niet concluderen dat leeftijd kennelijk geen invloed uitoefent op de bloeddruk. Misschien dat dit verband wel overtuigend kan worden aangetoond als je meer data tot je beschikking hebt. Verder hebben we bij het bekijken van de plaatjes met residuen in paragraaf 8.4 gezien dat het verband tussen bloeddruk en lichaamsgewicht waarschijnlijk niet lineair is. Daarom kan het verstandig zijn om de factor “leeftijd” te verwijderen uit het model. Je krijgt dan  $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + e$  en je zou m.b.v. dit eenvoudiger model opnieuw de waarden van de parameters kunnen schatten. Je krijgt dan:

```
Residuals:
    Min       1Q   Median       3Q      Max
-21.0878   -7.0521    0.9887    5.4629   24.0807

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.6673     20.6943   2.255  0.03213 *
Gewicht        0.5546      0.1934   2.867  0.00778 **
polsfreq       0.5623      0.2460   2.286  0.03006 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.09 on 28 degrees of freedom
Multiple R-squared:  0.3674, Adjusted R-squared:  0.3223
F-statistic: 8.133 on 2 and 28 DF,  p-value: 0.001642
```

Dit model heeft als voordeel boven de eerste versie dat alle geschatte parameters significant van 0 verschillen, m.a.w: alle verklarende variabelen vertonen een significante samenhang met de verklaarde variabele. In de praktijk wordt regressieanalyse dikwijls gebruikt om een selectie te maken uit verschillende mogelijke verklarende variabelen. Soms stuit je bij dergelijk onderzoek op verrassende resultaten. Zo kan het voorkomen dat een variabele die aanvankelijk significant leek, na toevoeging van (een) andere variabele(n) aan het model zijn significantie kwijtraakt. Dit illustreert nog eens dat je altijd voorzichtig moet zijn met het trekken van conclusies. Een

geconstateerde statistische samenhang hoeft niet te betekenen dat er ook sprake is van een causaal verband.

**Opmerking 8.5.1.** Bij meervoudige regressie moet je proberen te vermijden om verklarende variabelen in het model op te nemen die onderling sterk gecorreleerd zijn. Als je in een model bijvoorbeeld zowel “lichaamslengte” als “schoenmaat” als verklarende variabele opneemt, dan wordt het statistisch lastig om te ontdekken welke van de twee variabelen nu welk effect heeft: lange mensen hebben meestal ook een grote schoenmaat. Als je toch beide variabelen in het model opneemt, zul je zien dat beide parameters een relatief grote standaardafwijking hebben, terwijl die standaardafwijking sterk terugloopt als je één van de twee variabelen uit het model elimineert. Dit verschijnsel wordt in de literatuur *multicollineariteit* (Engels: *multicollinearity*) genoemd.

## 8.6 Categorical variabelen

Veronderstel dat we in het voorbeeld van paragraaf 8.3 ook de variabele “Geslacht” als verklarende variabele in het model willen opnemen en dat we per waarneming ook weten of het een man of een vrouw betreft. We kunnen natuurlijk het bestand splitsen in twee deelbestanden, één voor mannen en de ander voor vrouwen, maar deze handelwijze heeft twee belangrijke nadelen. In de eerste plaats worden de deelbestanden relatief klein en daarmee neemt de betrouwbaarheid van de schattingen af. In de tweede plaats zal elk deelbestand verschillende schattingen opleveren, maar zijn deze verschillen systematisch (veroorzaakt door het verschil in geslacht) of toevallig (andere steekproef)?

Een alternatieve benadering is om een nieuwe variabele te definiëren,  $X_4$ , met  $X_4 = 0$  als we te maken hebben met een man en  $X_4 = 1$  als we te maken hebben met een vrouw. Onze gegevens komen er dan als volgt uit te zien (we laten de variabele “leeftijd” weg):

Y	X1	X3	X4
115	40	80	0
120	47	68	1
148	76	92	0
113	44	84	1
etc	etc	etc	etc



## 8.6. Categorical variabelen

We krijgen de volgende resultaten:

```
Residuals:
    Min       1Q   Median       3Q      Max
-20.279   -6.262   -1.222    6.559   25.123

Coefficients:
              Estimate      Std. Error  t value Pr(>|t|)
(Intercept)  42.4587       20.8177    2.040  0.0513 .
Gewicht      0.5270        0.1932    2.728  0.0111 *
polsfreq     0.6681        0.2592    2.577  0.0157 *
Geslacht    -5.5481        4.5857   -1.210  0.2368
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.99 on 27 degrees of freedom
Multiple R-squared:  0.4, Adjusted R-squared:  0.3333
F-statistic: 5.999 on 3 and 27 DF,  p-value: 0.002862
```

De coëfficiënt voor “Geslacht” is  $\hat{\beta}_4 \approx -5,5481$ . Aangezien  $X_4 = 1$  als we te maken hebben met een vrouw, betekent dit dat de bloeddruk van vrouwen gemiddeld 5,5 lager ligt dan bij mannen, bij gelijke waarden van de overige variabelen. We zien ook dat deze waarde niet significant van 0 verschilt, dus het geconstateerde verschil tussen mannen en vrouwen kan ook door toeval zijn veroorzaakt. Een variabele als  $X_4$  wordt een *0-1-variabele* of een *dummy-variabele* genoemd.

### Opgave

117. Van een groot aantal auto's zijn de volgende gegevens vastgelegd: de brandstofefficiëntie (km per liter), het ledig gewicht (kg), de maximum snelheid en het merk. In het bestand komen drie merken voor: Toyota, Volkswagen en Fiat die elk via verschillende types zijn vertegenwoordigd. Je wilt met name weten welk merk de zuinigste auto's produceert, maar je wilt rekening houden met de invloed van andere kenmerken. Daarom stel je een lineair regressiemodel op met brandstofefficiëntie als verklaarde variabele en als verklarende variabelen het gewicht, de maximumsnelheid en het merk. Op welke manier ga je de variabele “Merk” in het model opnemen?

## 8.7 Niet-lineaire regressie

Vaak is het mogelijk om d.m.v. een nieuwe definitie van de variabelen een niet-lineair model te veranderen in een lineair model dat met de standaard-technieken kan worden geanalyseerd. We geven twee voorbeelden:

- Stel we hebben een puntenwolk  $(X_i, Y_i)$  ( $i = 1, \dots, n$ ) en we vermoeden dat er een parabolisch verband bestaat tussen  $Y$  en  $X$ . We gebruiken het model

$$Y_i = \alpha + \beta X_i + \gamma X_i^2 + e_i, \text{ waarbij } e_i \text{ een storingsterm is.}$$

Definieer nu  $Z_i = X_i^2$  en het model is te herschrijven als:

$$Y_i = \alpha + \beta X_i + \gamma Z_i + e_i.$$

Dit is een meervoudig lineair regressiemodel dat we met de technieken uit dit hoofdstuk kunnen analyseren.

- Stel we hebben een puntenwolk  $(X_i, Y_i)$  ( $i = 1, \dots, n$ ) en we vermoeden dat er een exponentieel verband bestaat tussen  $Y$  en  $X$ . We gebruiken het model

$$Y_i = \alpha \cdot \beta^{X_i} \cdot e_i, \text{ waarbij } e_i \text{ een storingsterm is.}$$

Dit model is te herschrijven als:

$$\log(Y_i) = \log(\alpha) + \log(\beta) \cdot X_i + \log(e_i).$$

Dit heeft de vorm van een enkelvoudig lineair regressiemodel. De technieken uit Hoofdstuk 6 leveren o.a. schattingen op voor  $\log(\alpha)$  en  $\log(\beta)$ .

Op de GR staan onder het menu **STAT - CALC** enkele van dergelijke modellen voorgeprogrammeerd.

## Hoofdstuk 9

### Gemengde opgaven

1. De lichaamslengte van mannen en vrouwen is normaal verdeeld. Voor vrouwen is het gemiddelde  $\mu$  gelijk aan 171 cm en voor mannen is dit gemiddelde gelijk aan 183 cm. De standaarddeviatie  $\sigma$  is voor beide groepen gelijk aan 6 cm.
  - a. Bereken de kans dat een man langer is dan 185 cm en de kans dat een vrouw langer is dan 185 cm.
  - b. Bereken deze kansen nog een keer door alleen gebruik te maken van de standaardnormale verdeling.
  - c. Bereken de kans dat een willekeurige man groter is dan een willekeurige vrouw.
2. De laatste jaren is er veel ophef geweest over het gebruik van EPO in de wielersport. EPO stimuleert de aanmaak van rode bloedlichaampjes en vergroot daardoor de zuurstoftoevoer naar de spieren. Omdat het gebruik van EPO niet direct in het bloed aantoonbaar is, wordt door de internationale wielerveding naar de hematocrietwaarde in het bloed gekeken. Men gaat ervan uit dat bij een waarde van 0,5 of hoger EPO-gebruik vaststaat. Neem bij de volgende opgaven aan dat de hematocrietwaarde bij benadering normaal verdeeld is. Van 6 wielrenners wordt de hematocrietwaarde in het bloed bepaald. Men vindt:

0,39   0,43   0,46   0,52   0,53   0,55.

- a. Bereken het gemiddelde, de mediaan, de range, de variantie en de standaarddeviatie van deze steekproef. Doe dit met en zonder gebruik te maken van de statistische functies op de GR.
- b. Wat is het verschil tussen de mediaan en het gemiddelde? Wat kun je zeggen over de verdeling wanneer de mediaan en het gemiddelde veel van elkaar verschillen?

### Gemengde opgaven

3. In een brief in de Lancet (Marx JJM en Vergouwen PCJ, Lancet 1998;352:451) publiceren twee Utrechtse onderzoekers de hematocrietwaarden van 18 mannelijke en 28 vrouwelijke topatleten, en 134 mannelijke en 144 vrouwelijke controles. Voor zover bekend gebruikte geen van de gemeten personen EPO. De resultaten staan in de volgende tabel.

	Men		Women	
	Mean (SD)	Range	Mean (SD)	Range
<b>Athletes</b>	0.44 (0.03)	0.38-0.52	0.40 (0.02)	0.34-0.47
<b>Controls</b>	0.45 (0.03)	0.37-0.55	0.41 (0.04)	0.30-0.51

- Bereken de kans dat een mannelijke sporter die geen EPO gebruikt toch beschuldigd wordt van dopinggebruik (d.w.z een hematocrietwaarde boven de 0,50 heeft). Gebruik hierbij de normale verdeling met  $\mu=0,44$  en  $\sigma=0,03$ .
  - Een 95%-referentie-interval is een interval waarin 95% van de populatie zich bevindt. Bereken een 95%-referentie-interval voor hematocrietwaarden van mannelijke topatleten. Gebruik hierbij opnieuw de normale verdeling met  $\mu=0,44$  en  $\sigma=0,03$ .
  - Wat vind je van het besluit van de internationale wielervedstrijden?
4. In deze opgave gaan we verder met de analyse van hematocrietwaarden bij atleten en controles (zie opgave 3). De hematocrietwaarden van 18 mannelijke en 28 vrouwelijke topatleten, en 134 mannelijke en 144 vrouwelijke controles staan in de volgende tabel beschreven:

	Men		Women	
	Mean (SD)	Range	Mean (SD)	Range
<b>Athletes</b>	0.44 (0.03)	0.38-0.52	0.40 (0.02)	0.34-0.47
<b>Controls</b>	0.45 (0.03)	0.37-0.55	0.41 (0.04)	0.30-0.51

- Bereken de standaardfout van het gemiddelde (SEM) en het 95% - betrouwbaarheidsinterval voor de hematocrietwaarde van mannelijke topatleten. Doe dit met en zonder gebruik te maken van de statistische functies op de GR.
- Wat stelt dit betrouwbaarheidsinterval precies voor? Wat is het verschil tussen dit interval en het referentie-interval berekend in opgave 3?
- Wat is het verschil tussen de SD en de SEM?

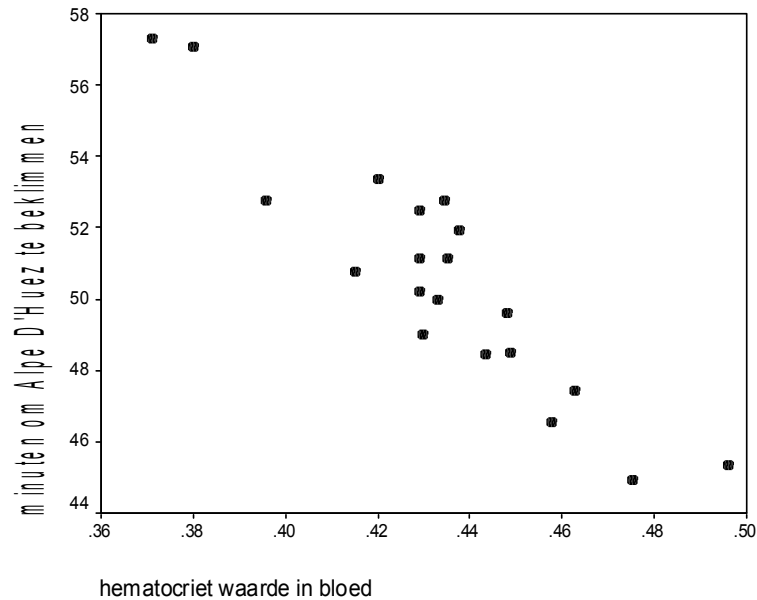
- d. Stel dat het onderzoek nogmaals uitgevoerd wordt, met twee keer zoveel personen. Wordt de SD dan groter, kleiner of blijft ze ongeveer gelijk? En wat gebeurt er met de SEM?
5. Een onderzoeker wil kijken of de hematocrietwaarde verhoogd wordt door het toedienen van EPO. Daarvoor bekijkt hij de hematocrietwaarde in het bloed van 20 renners voor en na het toedienen van EPO. Eén renner blijkt zowel voor als na het toedienen van EPO een hematocrietwaarde boven de 0,50 te hebben. Twaalf renners hebben zowel voor als na het toedienen van EPO een waarde onder de 0,50 en 7 renners hebben voor het gebruik van EPO geen verhoogde waarde maar na het gebruik wel. In een kruistabel zien deze gegevens er als volgt uit:

	hematocriet waarde na EPO		Totaal
	< 50	$\geq 50$	
hematocrietwaarde < 50 voor EPO	12	7	19
$\geq 50$		1	1
Totaal	12	8	20

Kun je op grond van deze gegevens concluderen dat het gebruik van EPO de kans op een hematocrietwaarde van 0,50 of meer significant verhoogt? Neem  $\alpha = 0,05$ .

### Gemengde opgaven

6. Een onderzoeker meet van een groep wielrenners de hematocriet-waarde in het bloed en het aantal minuten dat het de renner kostte om de Alpe d'Huez te beklimmen. Hij maakt van deze waarnemingen het volgende plaatje:



- a. Uit dit plaatje kun je opmaken dat de correlatiecoëfficiënt tussen hematocrietwaarde en de tijd om de Alpe d'Huez te beklimmen:
- a positief is en een waarde heeft dicht in de buurt van 1
  - b positief is en een waarde heeft dicht in de buurt van 0
  - c negatief is en een waarde heeft dicht in de buurt van 0
  - d negatief is en een waarde heeft dicht in de buurt van -1
- b. Kun je een voorbeeld schetsen waarbij de correlatiecoëfficiënt tussen twee continue uitkomsten 0 is terwijl er toch een verband bestaat tussen de twee uitkomsten?

De onderzoeker registreerde ook de hartslag van de renner aan de top van Alpe d'Huez. Hij vond:

Hematocrietwaarde	Hartslag
0,368	145
0,379	135
0,392	115
0,413	123
0,419	148
0,428	125
0,428	119
0,428	121
0,429	135
0,430	128
0,433	110
0,434	125
0,434	132
0,437	141
0,441	127
0,448	112
0,449	130
0,458	125
0,461	125
0,476	128
0,497	135

- c. Wat kun je concluderen over de samenhang tussen de hematocrietwaarde en de hartslag van een wielrenner?

Bij het bepalen van de hematocrietwaarde in het laboratorium is een fout gemaakt, hierdoor zijn alle waarden 0,05 te laag.

- d. Geef aan hoe de correcte figuur voor de samenhang tussen hematocrietwaarde en minuten om de Alpe d'Huez te beklimmen er uit ziet.
- e. Wat kan je zeggen over de correlatiecoëfficiënt tussen de correcte hematocrietwaarde en de hartslag aan de top van de Alpe d'Huez: gelijk aan, kleiner dan of groter dan de bij onderdeel c. gevonden waarde?

### Gemengde opgaven

7. Hieronder staan de resultaten van een gerandomiseerd vergelijkend onderzoek naar het effect van calcium op de bloeddruk van Afro-Amerikaanse mannen. (Naar: Lyle, Roseann. et al. JAMA, 257 (1987)) De behandelgroep van 10 mannen kreeg gedurende 12 weken een calcium-supplement. De controlegroep bestaande uit 11 mannen kreeg gedurende dezelfde periode een placebo. Bij alle mannen werd hun bloeddruk in rust gemeten voor en na de 12-weekse periode. De bovendruk is steeds genoteerd.

Behandeling	Begin	Eind	Behandeling	Begin	Eind
Calcium	107	100	Placebo	123	124
Calcium	110	114	Placebo	109	97
Calcium	123	105	Placebo	112	113
Calcium	129	112	Placebo	102	105
Calcium	112	115	Placebo	98	95
Calcium	111	116	Placebo	114	119
Calcium	107	106	Placebo	119	114
Calcium	112	102	Placebo	112	114
Calcium	136	125	Placebo	110	121
Calcium	102	104	Placebo	117	118
			Placebo	130	133

Wat kun je, bij  $\alpha = 0,05$ , concluderen over het effect van het calcium-supplement?

8. Bij een medische keuring laat een sportarts een aantal gezonde mannelijke vrijwilligers een tijdje onder grote belasting fietsen op een home-trainer. Hij verwacht dat deze mensen dit gemiddeld wel 15 minuten volhouden. Bij 8 proefpersonen meet hij de tijd dat ze het volhouden en vindt (in minuten): 10, 7, 9, 45, 8, 14, 3, 6.  
De sportarts wil nu toetsen of zijn aanname dat men de fietsproef gemiddeld 15 minuten volhoudt wel valide is.
- Bereken het gemiddelde, de mediaan en de range van deze steekproef.
  - Waarom is de t-toets hier niet geschikt? Welke toets zou je hier gebruiken? Er zijn twee mogelijkheden.
  - Formuleer de nulhypothese en de alternatieve hypothese, voer de toets vervolgens uit en trek je conclusie. Neem  $\alpha = 0,05$ . Doe dit voor beide mogelijke toetsen.



9. Dezelfde sportarts heeft bij de 8 proefpersonen uit opgave 7 de bloeddruk gemeten voor en een kwartier na de fietsproef. Hij wil weten of de bloeddruk een kwartier na de fietsproef anders was dan voor de fietsproef. Voor de diastolische bloeddruk vond hij:

persoon-nummer	bloeddruk voor fietsproef (mm Hg)	bloeddruk kwartier na fietsproef (mm Hg)
1	80	83
2	72	77
3	90	95
4	75	72
5	71	80
6	105	95
7	95	100
8	80	82

- a. Formuleer een voor deze onderzoeker geschikte toets. Voer hem uit en trek je conclusie. Gebruik  $\alpha = 0,05$ . Waar nodig mag je veronderstellen dat uitkomsten normaal verdeeld zijn.
  - b. Bereken een 95%-betrouwbaarheidsinterval voor het gemiddelde verschil in bloeddruk voor en na de fietsproef. Leg het verband met de uitkomst van onderdeel a.
  - c. Welke toets is het meest geschikt wanneer je niet kunt aannemen dat de normale verdeling geldt? Formuleer de bijbehorende hypothese, voer de toets uit en trek je conclusie.
10. In een onderzoek onder de Leidse jeugd bleek dat van de 1161 ondervraagde scholieren van het voortgezet onderwijs er 675 in de maand voorafgaand aan het onderzoek alcohol hadden gebruikt.
- a. Bereken de proportie alcoholgebruikers en het 95% betrouwbaarheidsinterval van deze proportie. Doe dit zonder en met gebruikmaking van de statistische functies op de GR.
  - b. Denk je, als je dit interval bekijkt, dat het aannemelijk is dat Leidse scholieren verschillen van de gehele Nederlandse populatie van scholieren, waar het percentage alcoholgebruikers gelijk is aan 54,3%?
11. Men wil kijken of het roken van ouders invloed heeft op de conditie van kinderen met astma. In een pilotonderzoek heeft men een aantal 12-jarige jongens met astma op een lopende band gezet en gemeten hoeveel minuten deze jongens konden rennen zonder uitgeput te raken. Men vond:

### Gemengde opgaven

Ouders roken: 12, 7, 23, 14, 46

Ouders roken niet: 8, 12, 17, 14, 36, 65

- a. Bereken in beide subgroepen het gemiddelde en de mediaan van het aantal minuten dat de jongens lopen. Kijk ook naar het maximum in beide groepen. Wat is je conclusie over de verdeling van deze grootte?
  - b. Is het verschil tussen beide groepen significant? Neem  $\alpha = 0,05$ .
12. In een onderzoek naar eetgewoontes van kinderen met het syndroom van Down (Hopman et al., J Am Diet Assoc 1998;98:790-4) werd gekeken of de voedingsinname adequaat was. In het artikel staan de volgende gegevens van kinderen in de leeftijd 1-4 jaar.

Energy and nutrient intakes per day	Children with Down syndrome (n=33, mean age =26 mo)	Control subjects (n=26, mean age = 28 mo)	RDA (Recommended Daily Allowances)
mean (standard deviation)			
Energy (kcal)	976 (210)	1214 (322)	1326
Energy (kcal/kg)	94 (22)	89 (28)	102

RDA is de dagelijkse aanbevolen hoeveelheid

kcal/kg is de energie-inname gedeeld door het gewicht van het kind.

- a. De RDA is de dagelijkse aanbevolen hoeveelheid voor kinderen van 1-4 jaar. Welke toets zou je gebruiken om na te gaan of de totale energie-inname van kinderen met het syndroom van Down voldoet aan deze aanbeveling? Formuleer ook de nulhypothese en het alternatief.
- b. Bereken de p-waarde van deze toets. Wat is je conclusie?
- c. Bereken ook het 95%-betrouwbaarheidsinterval voor de gemiddelde energie-inname van kinderen met het Down syndroom. Zit de aanbevolen dagelijkse inname in dit interval? Wat concludeer je hieruit?

13. In deze opgave gaan we de gegevens uit opgave 12 over eetgewoontes van kinderen met het syndroom van Down verder analyseren. In het onderzoek naar de voedingsinname van kinderen met het Downsyndroom heeft men ook de voedingsinname van een aantal gezonde kinderen bepaald. Men wil deze controlekinderen graag vergelijken met de kinderen met het Downsyndroom.
- Welke statistische toets is het meest geschikt om de gemiddelde energie-inname tussen controles en kinderen met Downsyndroom te vergelijken? Waar nodig mag je aannemen dat uitkomsten normaal verdeeld zijn.
  - Formuleer de hypothese. Bereken de p-waarde van deze toets en trek je conclusie.
  - Bereken een 95%-betrouwbaarheidsinterval voor het verschil in gemiddelde energie-inname tussen de twee groepen. Maak een berekening zonder en met gebruikmaking van de statistische functies van de GR. Leg het verband tussen het gevonden betrouwbaarheidsinterval en de uitkomst van onderdeel b.
  - Bereken ook een 95%-betrouwbaarheidsinterval voor het verschil in gemiddelde energie-inname per kg lichaamsgewicht. Trek je conclusie waarbij je de eerdere resultaten van deze opgave betreft.
14. In het onderzoek naar eetgewoontes van kinderen met het syndroom van Down werd ook gekeken naar de prevalentie van borstvoeding. Na 8 dagen kregen 19 van de 42 kinderen met het Downsyndroom nog borstvoeding. Men wil weten of dit verschilt van de gehele Nederlandse bevolking, waar in 1993 na 8 dagen nog 54% van de kinderen borstvoeding kreeg.
- Welke toets zou je gebruiken om te toetsen of het percentage kinderen met Downsyndroom dat borstvoeding krijgt gelijk is aan dat van de totale Nederlandse bevolking?
  - Formuleer de bijbehorende hypothese, voer de toets uit en trek je conclusie. Neem  $\alpha = 0,05$ .
  - Bereken het 95%-betrouwbaarheidsinterval voor de proportie kinderen met Downsyndroom dat na 8 dagen nog borstvoeding krijgt. Wat concludeer je op grond van dit interval?

Men heeft ook een controlegroep gezonde kinderen bekeken. Zevenentwintig van de 37 controlegroepkinderen kreeg na 8 dagen nog borstvoeding. Men wil deze kinderen vergelijken met de kinderen met het Downsyndroom.

### *Gemengde opgaven*

- d. Verschilt het percentage kinderen dat na 8 dagen borstvoeding krijgt significant tussen de twee groepen? Neem  $\alpha = 0,05$ . Voer de berekening uit met en zonder gebruik te maken van de statistische functies van de GR.
- e. Bereken (met en zonder gebruik te maken van de statistische functies van de GR) het 95%-betrouwbaarheidsinterval voor het verschil in proportie kinderen dat borstvoeding krijgt. Vergelijk je antwoord met het resultaat bij a.

# Hoofdstuk 10

## R-opgaven

### 10.1 Inleiding

Statistiek wordt in de praktijk vaak gebruikt om grote databestanden te analyseren. Dit kan niet met de hand of met de Grafische rekenmachine, maar moet met de computer gedaan worden. Als de onderzoeksopzet eenmaal is vastgesteld, zijn er twee hoofdtaken:

1. Data verzamelen, documenteren, coderen (bv. aangeven of je met categorische of met numerieke variabelen te maken hebt), controleren en verbeteren,
2. Statistische analyses loslaten op de data.

Een programma dat in de toegepaste statistiek veel gebruikt wordt voor deze 2 fases is SPSS. SPSS heeft een uitgebreide menustructuur en heel veel opties. Het heeft echter ook een paar nadelen: het is duur om aan te schaffen en weinig flexibel als je iets anders wilt dan wat al voorgeprogrammeerd is. Een alternatief voor SPSS is R. Het is freeware, dus voor iedereen gratis te downloaden, en inmiddels zijn er talloze ‘contributed packages’ verschenen voor gespecialiseerde statistische procedures. R is heel flexibel, maar een nadeel daarvan is dat gebruikers zelf ook veel of weinig (afhankelijk van hoe standaard hun wensen zijn) moeten programmeren. Inmiddels is een deel van dit nadeel ondervangen door het package `Rcmdr`, dat een menustructuur toevoegt aan R. Deze structuur lijkt op die van SPSS, maar is minder uitgebreid. R kan werken met databestanden die in andere programma’s samengesteld zijn.

#### 10.1.1 Installeren en gebruiken

Voordat je aan de slag kan met R, moet het natuurlijk eerst op je computer staan. Hieronder staat uitgelegd hoe je R kunt downloaden en hoe je vervolgens het `Rcmdr` package (kort voor: R Commander) opent. Het kan

## *R-opgaven*

gebeuren dat deze aanwijzingen niet meer precies opgaan bij toekomstige versies van R; volg dan de aanwijzingen op je scherm.

### **Thuis R downloaden**

- Ga naar <http://cran.r-project.org/>.
- Klik op ‘Windows’ onder ‘Download and Install R’ (of Linux of Mac, afhankelijk van waarmee je werkt).
- Kies ‘base’.
- Kies ‘Download R [versienummer] for windows’.
- Kies ‘Run’.
- Volg het installatiemenu. Belangrijk: als je het scherm ‘Selecteer Componenten’/‘Select components’ ziet, kies dan voor ‘Volledige installatie’/‘Full installation’.

Je hebt nu de basisversie van R op de computer staan. Nu moet je nog een extra package installeren.

- Open R.
- Klik op ‘Packages’, ‘Install Packages’.
- Klik op ‘Netherlands’ (plaats doet er niet toe).
- Zoek in de lange lijst ‘Rcmdr’.
- De rest gaat vanzelf; er wordt van alles geïnstalleerd.

Als je daarna met het pakket wilt werken:

- Eerst R openen (alleen als R nog niet open was).
- Dan ‘Load Package’.
- ‘Rcmdr’ aanklikken.
- De eerste keer dat je met R Commander werkt, vraagt R je of je nog extra packages wilt installeren die je nodig hebt voor volle functionaliteit van de R Commander. Antwoord ‘ja’ en wacht dan tot R klaar is.
- De bestanden die je nodig hebt, vind je ofwel op de site van cTWO ([www.fi.uu.nl/ctwo/WiskundeD/MateriaalDomeinenWiskundeD/welcome.html](http://www.fi.uu.nl/ctwo/WiskundeD/MateriaalDomeinenWiskundeD/welcome.html)), onder VWO, Domein B), ofwel op een plaats die je docent je opgeeft.

## 10.1. Inleiding

- Als je ze van de cTWO-site afhaalt, moet je eerst het .zip-bestand waar al het materiaal instaat openen, en vervolgens de bestanden kopiëren naar een folder op je eigen computer.

Je bent nu klaar om via R Commander te werken.

### Op school

- Eerst R openen. Je docent vertelt waar je R op de computer kunt vinden.
- Daarna (net als thuis) ‘Load Package’ en ‘Rcmdr’ aanklikken.
- In sommige opgaven wordt verwezen naar bestanden. Je hoort van je docent in welke map ze staan. De meeste van deze bestanden zijn geen R-bestanden, maar SPSS-bestanden (te herkennen aan naam.sav). Deze bestanden kun je inlezen door in het R Commander scherm via ‘Data’ en ‘Import data’ naar ‘from SPSS data set...’ te gaan.
- Als je een commando laat uitvoeren door Rcmdr, krijg je ook de expliciete aanroep van die functie in het ‘Script Window’. Je kunt daarin zelf opties veranderen en kijken wat er gebeurt (typen in Script Window, selecteren, ‘Submit’), de helppagina van de functie oproepen (door ‘?’ gevolgd door de functienaam), of meer informatie opzoeken in boeken of op het web. Voor de opdrachten van dit practicum is dit in principe allemaal niet nodig.
- Negeer de waarschuwing die onderaan het scherm van de Commander verschijnt.

## 10.2 Computeropgaven

De volgende opgaven zijn gebaseerd op opgaven voor studenten geneeskunde aan de Leidse universiteit. Je leert zowel data in te voeren als bestaande databestanden te gebruiken. Je gaat de data zowel beschrijven (hoofdstuk 10.2.1) als analyseren (hoofdstuk 10.2.2).

Ga bij elke opgave na:

- welke medische vraag of hypothese wordt onderzocht?
- welke gegevens heb je? welke mis je?
- waar vind je de uitkomst van de analyse in de computeruitvoer?
- wat is nu het antwoord op de medische vraag?
- had ik dit ook met de hand of met de GR kunnen uitrekenen?

### 10.2.1 Opgaven beschrijvende statistiek

In dit eerste deel van de opgavenset leer je hoe je gegevens uit een dataset kunt samenvatten en hoe je kunt zoeken naar verschillen tussen groepen, zonder dat je daar al statistische conclusies aan verbindt. Voor numerieke variabelen kun je bijvoorbeeld het gemiddelde, de mediaan en de standaarddeviatie uitrekenen, terwijl je bij categorische variabelen overzicht krijgt door aantallen en percentages in de verschillende categorieën uit te rekenen. Ook kun je plaatjes maken (zoals een histogram of een boxplot) om je een beeld te vormen van de spreiding van de data. In de volgende opgaven wordt uitgelegd hoe je dit alles kunt uitvoeren in R.

#### Opgave 1

Stel we willen de volgende gegevens van een zestal zwangerschappen invoeren en analyseren.

Gewicht	Leeftijd	Geslacht
3036	28	1
3005	31	1
3152	32	1
3073	20	2
2882	30	1
2943	30	2

Elke regel geeft gegevens van één bevalling: het geboortegewicht in grammen, de leeftijd van de moeder tijdens de bevalling in jaren en het



## 10.2. Computeropgaven

geslacht van het kind (1=meisje, 2=jongen).

- Voor het invoeren van deze gegevens klik je op ‘Data’ - ‘New Data Set’ - Geef een naam - ‘OK’. Je ziet nu een soort spreadsheet. Klik op het veld var1, vul de naam ‘Gewicht’ in voor deze kolom en vink aan dat het om numerieke gegevens gaat. In de velden daaronder vul je de cijfers in. Je doet hetzelfde met de tweede en derde kolom. Let op: ook bij geslacht voer je numerieke gegevens in (1 en 2), later zul je aangeven wat deze getallen betekenen. Nu sluit je het venster en je gaat terug naar R Commander.
- Nu klik je op ‘Data’ - ‘Manage variables in active data set’ - ‘Convert numeric variables to factors’. Je gaat nu aangeven dat de enen en tweeën in de kolom ‘Geslacht’ geen getallen zijn, maar codes. Je klikt op ‘Geslacht’ en je vinkt aan dat je namen wilt geven en je voert geen nieuwe naam voor ‘geslacht’ in. Achter ‘1’ voer je ‘meisje’ in en achter ‘2’ ‘jongen’. Je komt dan vanzelf terug in R Commander.
- Klik nu op ‘View data set’ en je dat de kolom ‘geslacht’ is aangepast.

Nu kun je het bestand analyseren. Probeer maar wat uit. Doe in ieder geval het volgende:

1. Bereken via ‘Statistics’ - ‘Summaries’ - ‘Numerical Summaries’ allerlei gegevens over de gewichten van de baby’s en de leeftijd van de moeders. Doe hetzelfde uitgesplitst naar jongen/meisje met behulp van de ‘Summarize by groups’-knop.
2. Maak via ‘Graphs’ diverse histogrammen, boxplots en scatter diagrams. Een boxplot geeft de spreiding van de waarnemingen rondom de mediaan aan (50% in het dikke stuk, ‘snorren’ maximaal  $1,5 \times$  de box-afstand, outliers apart).
3. Maak grafieken voor de jongens en de meisjes apart, bijvoorbeeld een histogram voor de gewichten van de meisjes. Hiervoor moet je eerst een nieuw bestand aanmaken, waarbij je alleen de gegevens die op meisjes betrekking hebben selecteert. Dat gaat als volgt. Klik op ‘Data’ - ‘Active Data Set’ - ‘Subset Active Data Set’. Je laat het vinkje achter ‘select all variables’ staan. Bij ‘subset expression’ voer je in: ‘geslacht==“meisje”’ (let op de aanhalingstekens en het dubbele =teken) en je kiest een nieuwe naam voor dit deelbestand. Als je weer terug bent in R Commander, controleer dan via View Data of het bestand is aangepast zoals je wilde. Daarna maak je de grafiek voor de meisjes (via ‘Graphs’ - ‘Histogram’). Om weer de oude dataset terug te krijgen (en dan op dezelfde manier alleen de jongens te selecteren) gebruik je ‘Data’ - ‘Active Data Set’ - ‘Select Active Data Set’.

## *R-opgaven*

Voor de volgende drie opgaven gebruik je een al bestaand bestand dat je op de volgende manier kunt openen: ‘Data’ - ‘Import data’ - ‘from SPSS data set’.

### **Opgave 2**

In de file ‘Down.sav’ staan gegevens over energie-inname van kinderen met het Down-syndroom en kinderen zonder het Down-syndroom. Lees deze file in.

4. Wat voor soort uitkomst is energie-inname?
5. Met een histogram kan de verdeling van energie-inname in kinderen met het Down-syndroom weergegeven worden. Maak deze figuur (hoe dat gaat, kan je teruglezen in de vorige opdracht.) Maak ook zo’n figuur voor kinderen zonder Down-syndroom.
6. Bereken de steekproefgemiddelden en de mediaan van energie-inname voor beide groepen.
7. Wat is je conclusie over de verdelingen? Zijn ze symmetrisch?

### **Opgave 3**

In de file ‘glucose.sav’ staan de glucosewaarden van een aantal controlepersonen.

8. Bereken het steekproefgemiddelde en de mediaan voor deze groep.
9. Maak een histogram. Is de verdeling symmetrisch?
10. Maak ook een boxplot. Zijn er extreme waarden in deze gegevens?
11. Maak tenslotte op twee manieren een 95%-referentie interval:
  - (a) door aan te nemen dat de normale verdeling geldt, en
  - (b) gebaseerd op de percentielen in de gegevens. Gebruik hiervoor het 2.5% en het 97.5% percentiel, die je kunt laten uitrekenen door bij ‘Numerical Summaries’ in het hokje ‘quantiles’ zelf de getallen 0.025 en 0.975 toe te voegen.
12. Welk referentie-interval heeft je voorkeur? Motiveer je antwoord.

### **Opgave 4**

In de file ‘bloeddruk.sav’ staan bloeddrukgegevens van 8 personen voordat ze een fietsproef doen en 15 minuten nadat ze een fietsproef uitgevoerd hebben. Lees deze file in.

13. Wat voor soort uitkomst is bloeddruk?

## 10.2. Computeropgaven

14. Om een idee te krijgen van de samenhang tussen de bloeddruk voor en de bloeddruk na afloop van de fietsproef, kunnen we kijken naar de verschillen tussen deze variabelen: hoe groot zijn deze en wat is hun spreiding? Neemt de bloeddruk misschien sterker toe bij mensen die toch al een hoge bloeddruk hadden (of andersom)?

Om per persoon het verschil in bloeddruk te berekenen, kunnen we in R het volgende doen: ga naar 'Data' - 'manage variables in active data set' - 'compute new variable'. Hier kun je in het hokje 'new variable name' de nieuwe variabele een naam geven, bijvoorbeeld 'verschil'. En in het hokje 'expression to compute' geef je aan wat er precies uitgerekend moet worden. Aangezien het hier om een verschil gaat zetten we hier: 'bloeddruk\_na-bloeddruk\_voor'. Als je nu op 'View data set' klikt, zie je dat de dataset een extra kolom heeft gekregen, waarin per persoon precies de verschillen in bloeddruk staan. Nu kan je van deze nieuwe variabele een histogram maken om te zien hoe groot de verschillen ongeveer zijn en hoe ze verdeeld zijn.

15. Maak dit histogram.

### 10.2.2 Opgaven met toetsen en schatten

In het tweede deel van de opgavenset gaan we er vanuit dat je weet hoe je gemiddeldes e.d. kunt berekenen (ook voor subgroepen), dat je weet hoe je verschillende plaatjes kunt maken, hoe je een bestand opsplijt in bestanden voor verschillende subgroepen en hoe je een nieuwe variabele berekent. In de komende opgaven gaat het niet meer alleen om het beschrijven van de data, maar ook om het testen van verschillende hypothesen en om het zoeken naar verbanden tussen variabelen, met behulp van lineaire regressie en correlatiecoëfficiënten.

Per opgave wordt aangegeven hoe je dit met R kunt doen, tenzij dit al eerder gedaan is. In dat geval moet je zelf even terugzoeken waar het uitgelegd werd.

#### *Toetsen*

##### **Opgave 1**

We bekijken hier gegevens uit een onderzoek naar eetgewoontes van kinderen met het syndroom van Down. In de file 'Down.sav' staat de energie-inname per dag (in kcal) van een aantal kinderen met Down-syndroom en een aantal controles.

16. Lees deze file in via 'Data' - 'Import Data' - 'from SPSS data set', en bekijk de gegevens.

## R-opgaven

17. Bereken voor kinderen met en zonder Down-syndroom apart het gemiddelde, de mediaan en de standaarddeviatie van de energie-inname. Maak ook per groep apart een histogram van de gegevens.
18. Welke file(s) heb je nodig om de gemiddelde energie-intake in beide groepen te vergelijken? Welke toets heb je hiervoor nodig?
19. Formuleer de nulhypothese en het alternatief. Voer de toets uit met  $\alpha = 0.05$ . Klik hiervoor op de commando's 'Statistics' - 'Means' en zoek de goede toets. Vul nu het schermje in. Kies 'yes' bij 'assume equal variances'. Voer als alles ingevuld is het commando uit. Bestudeer de output.
  - Is het aantal vrijheidsgraden correct?
  - Geef het 95% betrouwbaarheidsinterval voor het echte verschil in energie-intake tussen kinderen met en zonder Down-syndroom. Zit 0 in het interval?
  - Geef de p-waarde die bij de niet-gepaarde t-toets hoort.
  - Formuleer je conclusie.

## Opgave 2

Om te kijken of er geen systematische afwijking bestaat tussen verschillende keuringsartsen, werden 32 mensen die in de WAO zaten gekeurd door twee artsen. Beide artsen noteerden voor welk percentage de persoon was afgekeurd. De gegevens staan in de SPSS-file 'WAO.sav'.

20. Haal deze file binnen en bekijk de gegevens. Wat voor uitkomst is percentage 'afgekeurd zijn'? Zijn dit gepaarde of niet-gepaarde gegevens? Zijn ze normaal verdeeld?
21. Bereken voor elke persoon het verschil tussen de meting van arts 1 en arts 2. Maak een histogram van de verschillen en bereken het gemiddelde verschil en de standaarddeviatie van de verschillen. Bereken met de hand de standaardfout van het gemiddelde verschil en het 95% betrouwbaarheidsinterval voor het gemiddelde verschil.
22. Welke toets moet je uitvoeren om te kijken of de twee artsen al dan niet overeenstemmen? Welke veronderstelling over de gegevens maken we hierbij? Denk je dat dit een redelijke veronderstelling is?
23. Zoek de goede toets onder 'Statistics'. Voer het commando uit. Wat is je conclusie?

24. Controleer de waarde van de t-statistic en het aantal vrijheidsgraden. Zit 0 in het 95% betrouwbaarheidsinterval voor het gemiddelde verschil? Geeft dat nog extra informatie?

### Opgave 3

In de file 'borstvoeding.sav' vind je gegevens over borstvoeding uit het onderzoek naar eetgewoontes van kinderen met het syndroom van Down. Lees de file in en bekijk welke variabelen er in deze file zitten, hoe ze heten en wat de value labels zijn. 'NA' is not 'available', deze data zijn onbekend. Ze worden genegeerd in de analyses.

25. Maak een kruistabel van wel of niet starten met borstvoeding tegen de groep via 'Statistics' - 'Contingency Tables' - 'Two-way table'. Bereken het percentage kinderen dat begint met borstvoeding in beide groepen door met het subcommando 'Cells' de relevante percentages uit te rekenen. Wat valt je op?
26. Toets of de percentages kinderen die met borstvoeding beginnen in beide groepen gelijk zijn. Welke toets kun je daarvoor gebruiken? Welke aanname doe je dan? Je kunt hem uitvoeren in hetzelfde venster als je voor de vorige opgave gebruikt hebt. Als de aanname niet opgaat voor je data, kun je Fishers exacte toets gebruiken. Vergelijk de p-waarden van beide toetsen.
27. Bereken ook het percentage kinderen in beide groepen dat na 8 dagen nog borstvoeding krijgt. Toets of deze percentages in beide groepen gelijk zijn. Wat valt je op als je het dit resultaat met dat van de vorige opgave vergelijkt?
28. In dezelfde file staat ook voor die kinderen die met borstvoeding begonnen zijn hoelang zij uitsluitend borstvoeding gekregen hebben. Bereken voor kinderen met en zonder Down-syndroom apart het gemiddelde, de mediaan en de standaarddeviatie van deze borstvoedingsduur. Maak ook per groep apart een histogram van de gegevens. Wat valt je op als je de gemiddelden en medianen in beide groepen vergelijkt?
29. Geef je de voorkeur aan een parametrische toets of een niet-parametrische toets om de borstvoedingsduur in beide groepen te vergelijken? Welke toets is het meest geschikt?
30. Voer deze toets uit. Wat concludeer je uit de uitvoer? Zou je een verklaring kunnen geven?

## R-opgaven

### Opgave 4

In de file ‘enquete.sav’ vind je de gegevens van een vragenlijst. Een aantal Leidse studenten geneeskunde werd gevraagd naar geslacht, leeftijd, lengte, rookgedrag, alcoholgebruik, keuze voor roltrap of gewone trap, en vervoermiddel naar collegezaal.

De codering is als volgt:

Geslacht	geslacht student
Leeftijd	leeftijd student in jaren
Lengte	lengte student in centimeters
Rookgedrag	rookgedrag student
Alcoholgebruik	aantal glazen alcohol dat per week genuttigd wordt
Traplopen	gebruik van trap of roltrap
Vervoer	vervoermiddel naar college

31. Lees de file in en bekijk de gegevens.
32. Beschrijf nu de afzonderlijke variabelen met behulp van kengetallen (gebruik ‘Statistics’ - ‘Summaries’ - ‘Frequency Distributions’ voor de categorische variabelen).
33. Onderzoek of het percentage mannen en vrouwen dat rookt gelijk is. Bereken het percentage mannen en vrouwen dat rookt. Welke toets zou je gebruiken om te toetsen of deze twee percentages significant van elkaar verschillen? Voer deze toets uit en formuleer je conclusie.
34. Men wil weten of mannen en vrouwen gemiddeld even lang zijn. Welke toets gebruik je hiervoor? Voer deze toets uit en formuleer je conclusies.
35. Maak een scatterplot (onder ‘Graphs’) met alcoholgebruik op de verticale as en lengte op de horizontale as. Vink de optie ‘least-squares line’ aan voor een regressielijn.
36. Bereken Pearsons correlatiecoëfficiënt tussen lengte en alcoholgebruik via ‘Statistics’ - ‘Summaries’ - ‘Correlation test’. Is er een significante associatie tussen lengte en alcoholgebruik? Is de correlatie positief of negatief? Klopt dat met wat je ziet op de scatterplot?
37. Is er sprake van ‘confounding’ in de vorige vraag? Hint: bekijk het verband tussen lengte en alcoholgebruik voor 2 goedgekozen groepen apart.

### *Toetsen en regressie*

In de file ‘zwangerschap.sav’ staan enkele gegevens van een aantal veel te vroeg geboren kinderen.

38. Bekijk welke gegevens er in de dataset zitten. Zwangerschapsduur is gemeten in weken (normaal is 37-42 weken); geboortegewicht is in grammen (normaal is ca 2750-4250 gram); sectio betekent keizersnede.
39. Bereken voor de continue variabelen het gemiddelde, de mediaan, de standaarddeviatie en het minimum en maximum. Bereken voor de categorische variabelen de aantallen en percentages in de verschillende categorieën.
40. Maak een kruistabel waarin je eenling/meerling uitzet tegen geboren met of zonder keizersnede. Bereken door de juiste opties te kiezen welk percentage van de eenlingen met een keizersnede geboren is en welk percentage van de meerlingen, en controleer je antwoorden door de percentages met de hand te berekenen.
41. Maak een plaatje om te kijken of er verschillen zijn in geboortegewicht tussen jongens en meisjes. Dit kun je doen met een boxplot met geslacht op de x-as en gewicht op de y-as. Wat valt op?
42. Bereken voor de jongens en meisjes apart het gemiddelde geboortegewicht. Welke toets is het meest geschikt om de geboortegewichten tussen jongens en meisjes te vergelijken? Voer deze toets uit en verwoord je conclusie.
43. Bekijk wat het gemiddelde geboortegewicht is bij een gegeven zwangerschapsduur. Maak daarvoor een plaatje (een Scatterplot) met zwangerschapsduur op de x-as en geboortegewicht op de y-as. Breng een regressielijn in het plaatje aan. Schat met het blote oog wat de constante en de richtingscoëfficiënt van deze lijn zijn.
44. Voer een lineaire regressie uit met gewicht als afhankelijke (response) variabele en zwangerschapsduur als onafhankelijke variabele via de commando’s ‘statistics’ - ‘fit models’ - ‘linear regression’. Noteer de vergelijking van de regressievergelijking en vergelijk deze met de schatting die je gemaakt hebt.
45. Bekijk de regressiecoëfficiënt voor zwangerschapsduur. Hoe interpreteer je deze coëfficiënt? Verschilt deze significant van 0? Wat concludeer je hieruit?
46. Maak nogmaals het plaatje met zwangerschapsduur op de x-as en geboortegewicht op de y-as maar nu met aparte symbolen voor de jongens en de meisjes, en laat in dit plaatje voor de twee groepen apart

### *R-opgaven*

een regressielijn tekenen. Dit kun je doen door ‘plot by groups’ te gebruiken.

47. Voer nu een multiple regressie uit met gewicht als afhankelijke variabele en zowel zwangerschapsduur als geslacht als onafhankelijke variabelen. Let op: dit gaat nu via de commando's ‘statistics’ - ‘fit models’ - ‘linear model’; klik de onafhankelijke variabelen naar het rechtervak en verbind ze door +. Schrijf de regressievergelijking weer op. NB bij factoren, zoals geslacht, neemt R de categorie met de laagste waarde als referentiecategorie, in dit geval is dat ‘meisje’. R maakt zelf zogenaamde dummy-variabelen aan, dat wil zeggen dat de referentiecategorie waarde 0 krijgt, en de andere categorie waarde 1.
48. Welke waarde voor de variabele geslacht hebben meisjes en jongens nu in de vergelijking? Vul deze waarden elk apart in in de regressievergelijking die je in de vorige vraag gevonden hebt en construeer op deze manier de regressielijnen voor meisjes en jongens apart.
49. Hoe interpreteer je nu de regressiecoëfficiënt voor zwangerschapsduur? En hoe interpreteer je de regressiecoëfficiënt voor geslacht?



# Appendices

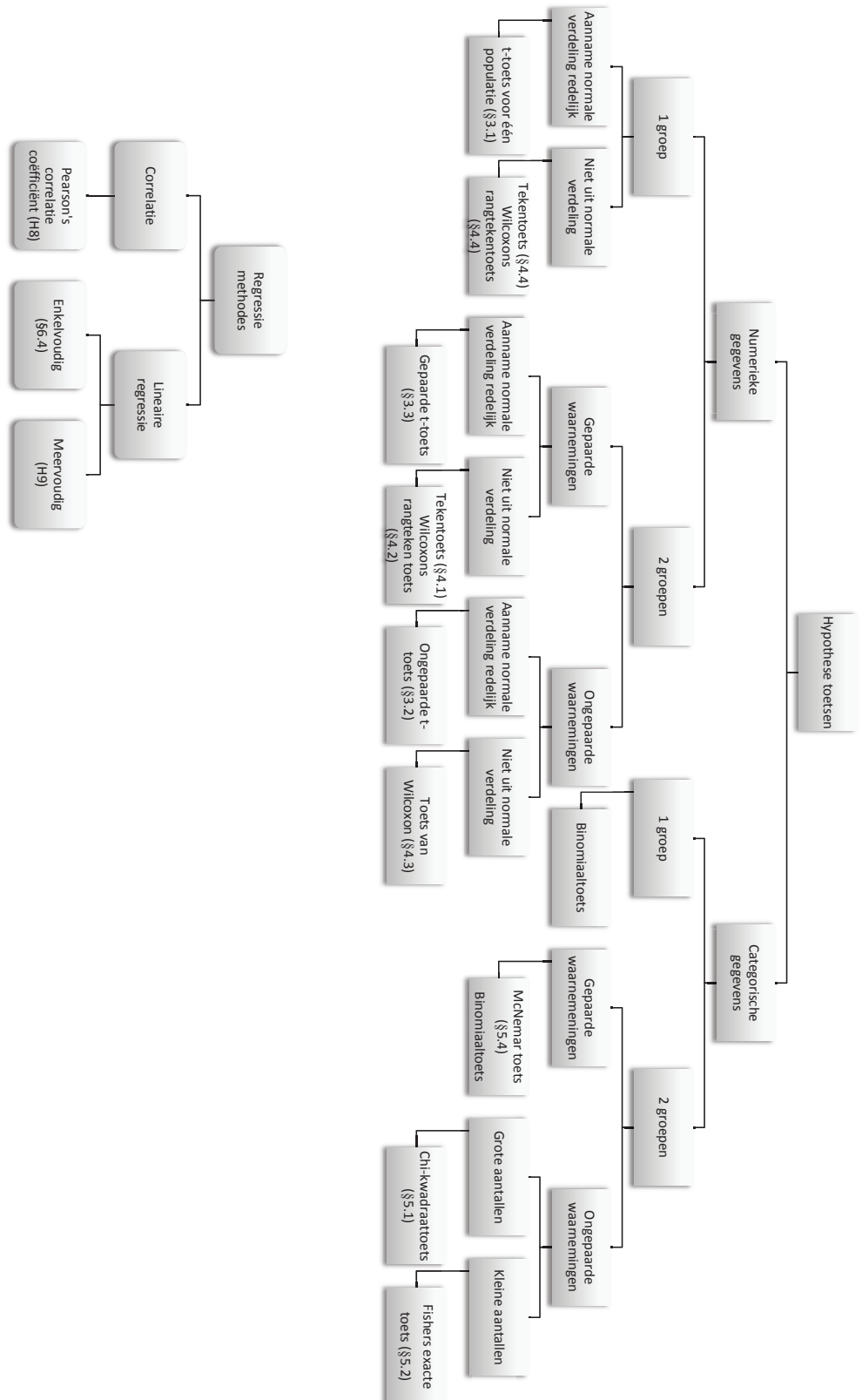
## Appendix 1: Wilcoxon's rangtekentoets

<b>Wilcoxon's rangtekentoets</b> Rechter kritieke waarden voor de som S van de van een teken voorziene rangnummers van verschillen (tweezijdig) De linker kritieke waarde is gelijk aan de rechter voorzien van een minteken.				
n'	$\alpha=0.01$	$\alpha=0.02$	$\alpha=0.05$	$\alpha=0.10$
1	-	-	-	-
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-
5	-	-	-	15
6	-	-	21	17
7	-	28	24	22
8	36	34	30	26
9	43	39	35	29
10	49	45	39	35
11	56	52	46	40
12	64	60	52	44
13	73	67	57	49
14	81	75	63	55
15	90	82	70	60
16	93	90	78	66
17	107	99	85	71
18	117	107	91	77
19	126	116	98	84
20	136	124	106	90
21	147	133	115	97
22	157	143	123	103
23	168	152	130	110
24	178	162	138	118
25	189	173	147	125
26	201	183	155	131
27	212	194	164	140
28	224	204	174	146
29	235	215	183	155
30	247	225	191	163

## Appendix 2: Kritieke waarden bij de toets van Wilcoxon (tweezijdig)

$l$ is het grootste getal waarvoor geldt $\Pr(S_{X_1} \leq l) \leq \frac{1}{2}\alpha$									
$r$ is het kleinste getal waarvoor geldt $\Pr(S_{X_2} \geq r) \leq \frac{1}{2}\alpha$									
		$\alpha = 0,1$		$\alpha = 0,05$		$\alpha = 0,02$		$\alpha = 0,01$	
$n_1$	$n_2$	$l$	$r$	$l$	$r$	$l$	$r$	$l$	$r$
2	5	3	13						
	6	3	15						
	7	3	17						
	8	4	18	3	19				
	9	4	20	3	21				
	10	4	22	3	23				
3	3	6	15						
	4	6	18						
	5	7	20	6	21				
	6	8	22	7	23				
	7	8	25	7	26	6	27		
	8	9	27	8	28	6	30		
	9	10	29	8	31	7	32	6	33
	10	10	32	9	33	7	35	6	36
4	4	11	25	10	26				
	5	12	28	11	29	10	30		
	6	13	31	12	32	11	33	10	34
	7	14	34	13	35	11	37	10	38
	8	15	37	14	38	12	40	11	41
	9	16	40	15	41	13	43	11	45
	10	17	43	15	45	13	47	12	48
5	5	19	36	17	38	16	39	15	40
	6	20	40	18	42	17	43	16	44
	7	21	44	20	45	18	47	17	48
	8	23	47	21	49	19	51	17	53
	9	24	51	22	53	20	55	18	57
	10	26	54	23	57	21	59	19	61
6	6	28	50	26	52	24	54	23	55
	7	29	55	27	57	25	59	24	60
	8	31	59	29	61	27	63	25	65
	9	33	63	31	65	28	68	26	70
	10	35	67	32	70	29	73	27	75
7	7	39	66	36	69	34	71	32	73
	8	41	71	38	74	36	76	34	78
	9	43	76	40	79	37	82	35	84
	10	45	81	42	84	39	87	37	89
8	8	51	85	49	87	46	90	43	93
	9	54	90	51	93	48	96	45	99
	10	56	96	53	99	50	102	47	105
9	9	66	105	63	108	59	112	56	115
	10	69	111	65	115	61	119	58	122
10	10	82	128	78	132	74	136	71	139

### Appendix 3: Schema van statistische technieken



# Verantwoording

In deze publicatie zijn enkele voorbeelden ontleend aan:

J. Devore en R. Peck: *Statistics. The Exploration and Analysis of Data*, Fifth Edition (2005), Brooks/Cole

De gemengde opgaven en de R-opgaven in hoofdstuk 9 en 10 zijn gebaseerd op onderwijsmateriaal voor studenten Geneeskunde en Biomedische Wetenschappen van de Universiteit Leiden, dat ontwikkeld is door Saskia le Cessie (Afdeling Medische Statistiek en Bioinformatica, Leids Universitair Medisch Centrum).