

Uitwerkingen bij Levende Statistiek

Een module voor Wiskunde D VWO

Inhoudsopgave

1	Hoofdstuk 1	3
2	Hoofdstuk 2	9
3	Hoofdstuk 3	12
4	Hoofdstuk 4	18
5	Hoofdstuk 5	23
6	Hoofdstuk 6	28
7	Hoofdstuk 7	35
8	Hoofdstuk 8	39
9	Hoofdstuk 9	41
10	Hoofdstuk 10	51

1 Hoofdstuk 1

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
1. 3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

We stellen een tabel op met alle mogelijke uitkomsten van het gooien met twee dobbelstenen. Elk vakje in de tabel geeft de som van het aantal ogen op beide dobbelstenen weer. De kans op de gebeurtenissen in elk van de hokjes zijn $\frac{1}{36}$. Het getal 4 komt in drie hokjes voor. Dus $\Pr(Y = 4) = \frac{3}{36} = \frac{1}{12}$. Op deze manier stellen we de kansverdeling samen.

y	2	3	4	5	6	7	8	9	10	11	12
$\Pr(Y = y)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

2. $\Pr(X = 0) \approx \text{binompdf}(4, 0.25, 0) = 0,316$ etc.

x	0	1	2	3	4
$\Pr(X = x)$	0,316	0,421	0,211	0,047	0,004

3. $E(Y) = \frac{1}{36} \cdot 2 + \frac{1}{18} \cdot 3 + \dots + \frac{1}{36} \cdot 12 = \frac{252}{36} = 7$. (Je kunt ook zonder te rekenen op grond van symmetrie tot deze uitkomst komen.)
4. $E(X) = \frac{1}{16} \cdot 1 + \frac{1}{8} \cdot 2 + \dots + \frac{7}{16} \cdot 6 = \frac{71}{16} = 4\frac{7}{16}$
5. $E(X) = 0,316 \cdot 0 + 0,421 \cdot 1 + \dots + 0,004 \cdot 4 = 1$ (Je kunt ook gebruik maken van de algemene formule voor de verwachtingswaarde van de binomiale verdeling: $E(X) = n\pi$.)
6. We bereken eerst de mogelijke uitkomsten van $(X - E(X))^2 = (X - 4\frac{7}{16})^2$. Als $X = 1$ krijgen we $(1 - 4\frac{7}{16})^2 \approx 11,816$. Zo verder gaand krijgen we de volgende tabel:

x	11,8164	5,9414	2,0664	0,1914	0,3164	2,4414
$\Pr((X - 4\frac{7}{16})^2 = x)$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{7}{16}$

Er volgt: $Var(X) = E(X - 4\frac{7}{16})^2 = \frac{1}{16} \cdot 11,8164 + \dots + \frac{7}{16} \cdot 2,4414 \approx 2,871$ en $\sigma(X) = \sqrt{Var(X)} \approx 1,694$.

7. Op dezelfde manier als in opgave 6 krijgen we
 $Var(X) = E(X-1)^2 \approx 0,316 \cdot (-1)^2 + 0,421 \cdot 0^2 + \dots + 0,004 \cdot 3^2 = 0,75$.
 Je kunt ook de algemene formule voor de variantie van de binomiale verdeling gebruiken: $Var(X) = n\pi(1-\pi) = 4 \cdot 0,25 \cdot 0,75 = 0,75$.
 Dus $\sigma(X) = \frac{1}{2}\sqrt{3} \approx 0,866$.
8. $E(X+Y) = 0,06(1+1) + 0,06(1+2) + \dots + 0,12(4+3) = 4,8$
 $E(X) = 0,2 \cdot 1 + 0,2 \cdot 2 + 0,3 \cdot 3 + 0,3 \cdot 4 = 2,7$
 $E(Y) = 0,3 \cdot 1 + 0,3 \cdot 2 + 0,4 \cdot 3 = 2,1$
 Wat opvalt is dat $E(X+Y) = E(X) + E(Y)$. Dit was ook zo in het voorbeeld van de *afhankelijke* X en Y .
9. Afhankelijke X en Y : $E(XY) = 0,09 \cdot 1 \cdot 1 + 0,05 \cdot 1 \cdot 2 + \dots + 0,14 \cdot 4 \cdot 3 = 5,84$
 Onafhankelijke X en Y : $E(XY) = 0,06 \cdot 1 \cdot 1 + 0,06 \cdot 1 \cdot 2 + \dots + 0,12 \cdot 4 \cdot 3 = 5,67$
 Wat opvalt is dat, *alleen in het geval van de onafhankelijke* X en Y , geldt
 $E(XY) = E(X) \cdot E(Y)$.
10. Stel $\Pr(X = x_i) = p_i$ ($i = 1, \dots, n$), met $\sum p_i = 1$. Dan geldt:
 $E(X+a) = \sum p_i(x_i+a) = \sum p_i x_i + \sum p_i a = \sum p_i x_i + a \sum p_i = E(X) + a$
 $E(aX) = \sum p_i(ax_i) = a \sum p_i x_i = aE(X)$.
 Om te begrijpen waarom $E(E(X))$ gelijk is aan $E(X)$ is het genoeg om in te zien dat $E(X)$ gewoon een getal is en dus geschreven kan worden als a . Nu krijgen we $E(a) = \sum p_i a = a \sum p_i = a$. Aangezien we $a = E(X)$ hadden, krijgen we nu $E(E(X))$ gelijk is aan $E(X)$.
- 11.
- $$\begin{aligned} Var(X) &= E(X - E(X))^2 = E(X^2 - 2X \cdot E(X) + E^2(X)) \\ &= E(X^2) - 2E(X) \cdot E(X) + E^2(X) \text{ (vanwege opgaven 8 en 10)} \\ &= E(X^2) - E^2(X). \end{aligned}$$
- 12.
- $$\begin{aligned} Var(aX) &= E(a^2 X^2) - E^2(aX) \text{ (vanwege opgave 11)} \\ &= a^2 E(X^2) - a^2 E^2(X) \text{ (vanwege opgave 10)} \\ &= a^2 (E(X^2) - E^2(X)) \\ &= a^2 Var(X) \end{aligned}$$

13.

$$\begin{aligned}
 \text{Var}(X + Y) &= E(X + Y)^2 - E^2(X + Y) \\
 &= E(X^2 + 2XY + Y^2) - (E(X) + E(Y))^2 \\
 &= E(X^2) + 2E(X)E(Y) + E(Y^2) - E^2(X) - 2E(X)E(Y) - E^2(Y) \quad (*) \\
 &= E(X^2) - E^2(X) + E(Y^2) - E^2(Y) \\
 &= \text{Var}(X) + \text{Var}(Y)
 \end{aligned}$$

(*) In deze stap wordt de onafhankelijkheid van X en Y gebruikt, daarom geldt $E(XY) = E(X)E(Y)$

14. a. $E(2X - 3)(Y - 1) = E(2XY - 2X - 3Y + 3)$
 $= 2E(X)E(Y) - 2E(X) - 3E(Y) + 3 = 35.$
- b. Uit $\text{Var}(X) = E(X^2) - E^2(X)$ volgt $E(X^2) = \text{Var}(X) + E^2(X) = 2 + 5^2 = 27.$
Dus $E(X^2 - 5Y) = E(X^2) - 5E(Y) = 27 - 5 \cdot 6 = -3.$
- c. $\text{Var}(2X - 3) = 2^2 \text{Var}(X) = 8$
- d. $\text{Var}(3X - Y) = \text{Var}(3X) + \text{Var}(-Y) = 3^2 \text{Var}(X) + (-1)^2 \text{Var}(Y)$
 $= 3^2 \cdot 2 + 1 \cdot 3 = 21.$
15. a. $\Pr(2 \leq X \leq 3) = \Pr(\frac{2-2,4}{0,7} \leq \frac{X-2,4}{0,7} \leq \frac{3-2,4}{0,7}) = \Pr(-\frac{4}{7} \leq Z \leq \frac{6}{7})$
waarin Z standaardnormaal verdeeld is.
De kans is ongeveer gelijk aan $\text{normalcdf}(-\frac{4}{7}, \frac{6}{7}) = 0,520$
- b. Er geldt $\Pr(X \geq x) = \Pr(Z \geq \frac{x-2,4}{0,7}) = 0,01$ en $\text{invNorm}(0.99) = 2,326.$
Dus $\frac{x-2,4}{0,7} = 2,326$ en $x \approx 2,326 \cdot 0,7 + 2,4 \approx 4,03.$
- c. Er geldt $\Pr(Z \geq \frac{50-\mu}{13}) = 0,07$ met Z standaardnormaal verdeeld en $\text{invNorm}(0.93) = 1,476.$
Dus $\frac{50-\mu}{13} = 1,476$ en $\mu \approx 50 - 13 \cdot 1,476 \approx 30,8.$
- d. Er geldt $\Pr(X \leq 30) = \Pr(Z \leq \frac{30-40}{\sigma}) = 0,25$ met Z standaardnormaal verdeeld en $\text{invNorm}(0,25) = -0,674.$
Dus $-\frac{10}{\sigma} = -0,674$ en $\sigma \approx \frac{-10}{-0,674} \approx 14,8.$
16. \bar{X} is normaal verdeeld met gemiddelde 80 en standaardafwijking $15/\sqrt{20}.$
De gevraagde kans is dus ongeveer $\text{normalcdf}(-10^{\wedge}99, 75, 80, 15/\sqrt{20}) = 0,068.$
17. Noem de score die Jan aan een willekeurige foto geeft $X.$
Er geldt: $E(X) = 0,15 \cdot 1 + 0,55 \cdot 2 + 0,3 \cdot 3 = 2,15$ en
 $\text{Var}(X) = E(X^2) - E^2(X) = 0,15 \cdot 1^2 + 0,55 \cdot 2^2 + 0,3 \cdot 3^2 - 2,15^2 = 0,4275.$
De gemiddelde score van 30 foto's is $\bar{X}.$ Voor deze stochast geldt:
 $E(\bar{X}) = E(X) = 2,15$ en $\text{Var}(\bar{X}) = \frac{1}{n} \text{Var}(X) = 0,4275/30 =$

0,01425,
 dus $\sigma(\bar{X}) = \sqrt{0,01425} \approx 0,119$.

18. Uit de centrale limietstelling volgt dat \bar{X} bij benadering normaal verdeeld is met gemiddelde 2,15 en standaardafwijking 0,119. Zonder toepassing van de continuïteitscorrectie, krijgen we:

$$\Pr(\bar{X} \leq 2) \approx \text{normalcdf}(-10^9, 2, 2.15, 0.119) = 0,104.$$

Met continuïteitscorrectie moet je de kans berekenen via X_{Som} :

$$\Pr(\bar{X} \leq 2) = \Pr(X_{Som} \leq 60) = \Pr(X_{Som} \leq 60,5) = \Pr(\bar{X} \leq 2,0167)$$

De laatste kans is ongeveer $\text{normalcdf}(-10^9, 2.0167, 2.15, 0.119) = 0,131$.

19. De kansverdeling van X_i^2 is:

x	0	1	4
$\Pr(X_i^2 = x)$	0,5	0,2	0,3

De kansverdeling van $Z = X_1^2 + X_2^2$ vind je via de tabel van de samengestelde kansverdeling van X_1^2 en X_2^2 . Je maakt daarbij gebruik van de onafhankelijkheid van X_1^2 en X_2^2 .

	$\Pr(X_1^2 = x_i \wedge X_2^2 = x_j)$		
$\downarrow x_i \backslash x_j \rightarrow$	0	1	4
0	0.25	0.1	0.15
1	0.1	0.04	0.06
4	0.15	0.06	0.09

Hieruit volgt de kansverdeling van $Z = X_1^2 + X_2^2$

z	0	1	2	4	5	8
$\Pr(Z = z)$	0,25	0,2	0,04	0,3	0,12	0,09

De kansverdeling van $Y = X_1^2 + X_2^2 + X_3^2 = Z + X_3^2$ vind je via de tabel van de samengestelde kansverdeling van Z en X_3^2 , waarbij je opnieuw gebruik maakt van onafhankelijkheid.

	$\Pr(X_3^2 = x_i \wedge Z = z_j)$					
$\downarrow x_i \backslash z_j \rightarrow$	0	1	2	4	5	8
0	0,125	0,1	0,02	0,15	0,06	0,045
1	0,05	0,04	0,008	0,06	0,024	0,018
4	0,075	0,06	0,012	0,09	0,036	0,027

Hieruit volgt de kansverdeling van $Y = Z + X_3^2$:

y	0	1	2	3	4	5	6	8	9	12
$\Pr(Y = y)$	0,125	0,15	0,06	0,008	0,225	0,18	0,036	0,135	0,054	0,027

20. Gebruik $Y1 = \chi^2\text{pdf}(X, \dots)$, waar je op de puntjes achtereenvolgens 2, 5, 20 en 50 invult. Als windows kun je bijv nemen:
 $[0, 5] \times [0, 1]; [0, 15] \times [0, 0.25]; [0, 40] \times [0, 0.1]$ en $[20, 80] \times [0, 0.05]$
 Wat opvalt is, dat met het stijgen van het aantal vrijheidsgraden de kromme minder scheef wordt. Dit is een gevolg van de centrale limietstelling, want een chi-kwadraatverdeling met n vrijheidsgraden is de som van n identiek verdeelde, onafhankelijke stochasten. Bij 50 vrijheidsgraden lijkt de kromme sterk op die van een normale verdeling.
21. $\chi^2\text{cdf}(0, 10, 8) = 0,735$.
22. Voer in: $Y1 = \chi^2\text{cdf}(X, 10^99, 15)$ en $Y2 = 0.05$. Window $[15, 30] \times [0, 0.2]$ geeft een geschikte grafiek. Intersect geeft $x \approx 25, 0$.
23. Vanwege $E(Y) = n$ en $Var(Y) = 2n$:
 $\chi^2\text{cdf}(2 - \sqrt{2} \times 2, 2 + \sqrt{2} \times 2, 2) \approx 0, 865$
 $\chi^2\text{cdf}(50 - \sqrt{2} \times 50, 50 + \sqrt{2} \times 50, 50) \approx 0, 686$
 $\text{normalcdf}(-1, 1) \approx 0, 683$
 De kansen bij de standaardnormale verdeling en de chi-kwadraatverdeling met 50 vrijheidsgraden zijn vrijwel gelijk. Dit is een gevolg van de centrale limietstelling.
24. Volgens de definitie van de chi-kwadraatverdeling geldt: $Y = \sum_{i=1}^n X_i^2$, waarbij de X_i 's onderling onafhankelijk en standaardnormaal verdeeld zijn.
 Uit $Var(X_i) = E(X_i^2) - E^2(X_i)$ volgt $E(X_i^2) = Var(X_i) + E^2(X_i) = 1 + 0 = 1$.
 Dus $E(Y) = nE(X_i^2) = n$.
25. Deze kans is ongeveer $\text{tcdf}(-10^99, 2, 7) = 0,957$.
26. Vanwege de symmetrie van de t-verdeling geldt: $\Pr(T \leq x) = 0,975$.
 De gezochte waarde van x is dus ongeveer $\text{invT}(0.975, 4) = 2,776$.
27. 15 keer de tcdf -functie aanroepen levert:

	$\Pr(X \geq 1)$	$\Pr(X \geq 2)$	$\Pr(X \geq 3)$
X heeft t-verdeling met 2 vrijheidsgraden	0,211	0,0918	0,04773
X heeft t-verdeling met 5 vrijheidsgraden	0,182	0,0510	0,01505
X heeft t-verdeling met 10 vrijheidsgraden	0,170	0,0367	0,00667
X heeft t-verdeling met 20 vrijheidsgraden	0,165	0,0296	0,00354
X is standaardnormaal verdeeld	0,159	0,0228	0,00135

Zoals verwacht gaan de kansen bij het stijgen van het aantal vrijheidsgraden steeds meer lijken op die van de standaardnormale verdeling. Maar de relatieve verschillen blijven “ver in de staart” groot. Zo verschilt de waarde van $\Pr(X \geq 1)$ bij een t-verdeling met 20 vrijheidsgraden praktisch weinig van die bij de standaardnormale verdeling, maar de overeenkomstige waarden van $\Pr(X \geq 3)$ schelen een factor 3.

2 Hoofdstuk 2

28. Er geldt $\text{invNorm}(0.975) = 1,96$. Dus als Z standaardnormaal verdeeld is geldt $\Pr(Z \leq 1,96) = 0,975$ en $\Pr(-1,96 \leq Z \leq 1,96) = 0,95$. Verder zijn de volgende ongelijkheden equivalent:

$$\begin{aligned} -1,96 &\leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1,96 \\ \frac{-1,96\sigma}{\sqrt{n}} &\leq \bar{X} - \mu \leq \frac{1,96\sigma}{\sqrt{n}} \\ -\bar{X} - \frac{1,96\sigma}{\sqrt{n}} &\leq -\mu \leq -\bar{X} + \frac{1,96\sigma}{\sqrt{n}} \\ \bar{X} + \frac{1,96\sigma}{\sqrt{n}} &\geq \mu \geq \bar{X} - \frac{1,96\sigma}{\sqrt{n}}. \end{aligned}$$

Voor een 99%-betrouwbaarheidsinterval gebruik je $\text{invNorm}(0.995)=2,576$.

29. In dit geval is σ bekend. Dus we gebruiken dat $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ standaardnormaal is. $\text{invNorm}(0.95) = 1,645$, dus er geldt:

$$\Pr(-1,645 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1,645) = 0,90. \text{ Hieruit volgt}$$

$$\Pr(\bar{X} - \frac{1,645\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{1,645\sigma}{\sqrt{n}}) = 0,90.$$

Het betrouwbaarheidsinterval wordt dus bepaald door $184 \pm \frac{1,645 \cdot 9}{\sqrt{50}}$, en is gelijk aan $[181,9 ; 186,1]$.

30. Voor deze steekproef geldt $\bar{x} = 71$. We maken de volgende tabel:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
70,5	-0,5	0,25
69,5	-1,5	2,25
73,1	2,1	4,41
72,0	1,0	1,0
70,2	-0,8	0,64
69,8	-1,2	1,44
72,7	1,7	2,89
70,2	-0,8	0,64

$$\text{We krijgen } s^2 = \frac{1}{8-1} \sum (x_i - \bar{x})^2 = \frac{1}{7} \cdot 13,52 \approx 1,931,$$

dus $SEM \approx \sqrt{1,931/\sqrt{8}} \approx 0,491$.

$\text{invT}(0.975,7) = 2,365$, dus het betrouwbaarheidsinterval is $71 \pm 0,491 \cdot 2,365$ ofwel $[69,8 ; 72,2]$.

Gebruik makend van STAT - TESTS:

Voer de gegevens in in lijst L1. TInterval met Inpt: Data, List: L1, Freq: 1 en C-Level: 0,95 geeft [69,8 ; 72,2].

31. TInterval met Inpt Stats, \bar{x} 71.5, Sx 3.1, n 30 en C-level 0.95 geeft [70,3 ; 72,7].
32. $\bar{x} = 629 : 48 \approx 13,10$ en $\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = 9517 - 48 \cdot 13,10^2 \approx 1274$ (zie opgave 11). Dus $s \approx \sqrt{1274/47} \approx 5,21$.
Tinterval met Inpt: Stats, \bar{x} : 13.1, Sx: 5.21, n: 48 en C-Level: 0.95 geeft [11,6 ; 14,6].
33. 90%: kleiner; 99%: groter.
34.
 - a. Gegevens invoeren in L1 en TInterval geeft [51,78 ; 55,56]. Of, als je met de hand rekt: $\bar{x} = 644/12 = 53,67$; $s = \sqrt{146,67/11} = 3,65$; $SEM = 3,65/\sqrt{12} = 1,054$
 $53,67 \pm 1,796 \cdot 1,054$ geeft [51,78 ; 55,56].
 - b. Veronderstel dat de vliegtijd X normaal verdeeld is met gemiddelde $\mu = 53,67$ en standaardafwijking $\sigma = 3,65$. x wordt dan gegeven door:
 $\text{InvNorm}(0.95, 53.67, 3.65) = 59,7$.
De te publiceren aankomsttijd is dus 11.00 uur.
 - c. $X - \bar{X}$ is normaal verdeeld met verwachting 0 en variantie $\sigma^2 + \sigma^2/12 = 13\sigma^2/12$, want X en \bar{X} zijn onafhankelijk.
Dus $\frac{X - \bar{X}}{\sigma\sqrt{13/12}}$ is standaardnormaal verdeeld.
We weten: $11s^2/\sigma^2$ heeft een $\chi^2_{[11]}$ -verdeling en is onafhankelijk van $X - \bar{X}$.
Dus (zie 1.5.2) $T = \frac{X - \bar{X}}{\sigma\sqrt{13/12}} / \sqrt{\frac{1}{11} \cdot \frac{11s^2}{\sigma^2}} = \frac{X - \bar{X}}{s\sqrt{13/12}}$ heeft een t-verdeling met 11 vrijheidsgraden.
 - d. $\text{invT}(0.95, 11) = 1,796$, dus $\Pr(\frac{X - \bar{X}}{s\sqrt{13/12}} \leq 1,796) = 0,95$.
Met $\bar{X} = 53,67$ en $s = 3,65$ geeft dit $\Pr(X \leq 53,67 + 1,796 \cdot 3,65\sqrt{13/12} = 60,49) = 0,95$.
We vinden dus $x = 60,49$ en de te publiceren aankomsttijd is 11.01 uur.
35. Voer in Y1 = binomcdf (50, X, 6) en Y2 = 0.025. Intersect geeft (X =) $\pi = 0,243$.
36. Voer in Y1 = 1-binomcdf (50, X, 5) en Y2 = 0,025. Intersect geeft (X =) $\pi = 0,045$.

37. $z_{0,05} \approx \text{invNorm}(0.975) = 1,96$ en $SEM = \sqrt{p(1-p)/n} = \sqrt{0,12 \cdot 0,88/50} \approx 0,046$.

Dus $\pi = 0,12 \pm 1,96 \cdot 0,046$ geeft $[0,030 ; 0,210]$. Dit interval is een benadering van de exacte variant $[0,045 ; 0,243]$.

38. *Exact:*

Voer in $Y1 = \text{binomcdf}(64, X, 27)$ en $Y2 = 0.05$. **Intersect** geeft $(X =) \pi = 0,532$.

Voer in $Y1 = 1 - \text{binomcdf}(64, X, 26)$ en $Y2 = 0.05$. **Intersect** geeft $(X =) \pi = 0,317$.

Het betrouwbaarheidsinterval is dus $[0,317 ; 0,532]$

Benadering met normale verdeling:

$\text{invNorm}(0.95) = 1,645$ en $SEM = \sqrt{p(1-p)/n} = \sqrt{\frac{27}{64} \cdot \frac{37}{64}/64} \approx 0,0617$

Dus $\pi = \frac{27}{64} \pm 1,645 \cdot 0,0617$ geeft het interval $[0,320 ; 0,523]$.

39. a. Voer in $Y1 = \text{binomcdf}(25, X, 1)$ en $Y2 = 0.025$. **Intersect** geeft $(X =) \pi = 0,204$.

Voer in $Y1 = 1 - \text{binomcdf}(25, X, 0)$ en $Y2 = 0.025$. **Intersect** geeft $(X =) \pi = 0,001$.

Het betrouwbaarheidsinterval is dus $[0,001 ; 0,204]$.

Je mag dit interval niet benaderen m.b.v. de normale verdeling, want $np = 1 \leq 5$. Doe je dit toch dan vind je als ondergrens een negatieve kans.

$$\left(\frac{1}{25} - 1,96 \cdot \sqrt{\frac{1}{25} \cdot \frac{24}{25}}/25 \approx -0,04\right)$$

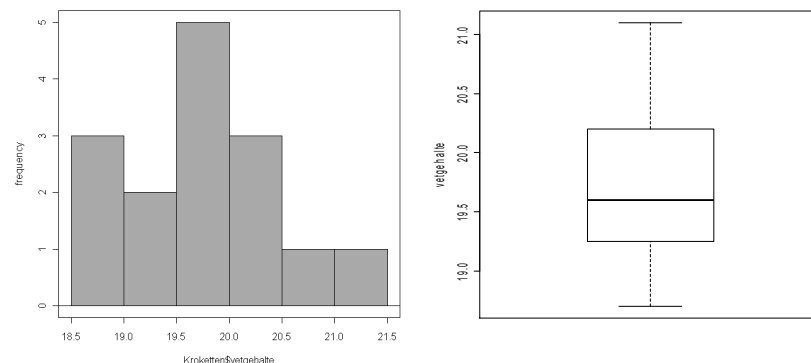
b. a is de oplossing van de vergelijking $\Pr(X \leq 1) = 0,05$, waarin X binomiaal verdeeld is met parameters $n = 25$ en onbekende π .

Voer in $Y1 = \text{binomcdf}(25, X, 1)$ en $Y2 = 0.05$. **Intersect** geeft $(X =) a = 0,113$.

Het betrouwbaarheidsinterval is dus $[0 ; 0,113]$

3 Hoofdstuk 3

40. De kritische grens van een eenzijdige toets ($H_1 : \mu > \mu_0$) met significantieniveau α is gelijk aan de kritische rechtergrens van een tweezijdige toets met significantieniveau 2α . We weten dat we bij een tweezijdige toets de nulhypothese kunnen verwerpen als μ_0 buiten het bijbehorende betrouwbaarheidsinterval voor μ ligt. Dus bij een eenzijdige toets ($H_1 : \mu > \mu_0$) kunnen we verwerpen als de μ_0 kleiner is dan de ondergrens van het $100(1 - 2\alpha)\%$ betrouwbaarheidsinterval voor μ .
41. a. Sla de gegevens op in L1. Om een histogram te maken ga je naar **Stat Plot - Plot1** en je kiest bij **Type** het figuurtje van het histogram. Vervolgens zet je de schakelaar op **On**. Als **Window** kies je eerst **Xmin=18**, **Xmax=22** en **Xscl=0,5**. Zo maak je de klassenindeling 18,0-18,5 18,5-19,0 t/m 21,5-22,0. Vervolgens kies je **Ymin=0** en **Ymax=5** (geen enkele klasse heeft meer dan 5 waarnemingen). Zorg ervoor dat er geen functies aan staan in je functie-invoerscherm. **Graph** geeft het hier afgebeelde histogram.



Je kunt ook een boxplot maken door (bij verder dezelfde instellingen) onder **Plot1** voor het desbetreffende icoontje te kiezen. De hierboven afgebeelde boxplot is een kwartslag gedraaid ten opzichte van de boxplot die je op je GR te zien krijgt. Beide figuren zijn niet perfect symmetrisch, terwijl de normale verdeling dat wel is. Je kunt eenvoudig inzien dat dit wordt veroorzaakt door het geringe aantal waarnemingen. Verder zien we een zekere concentratie van waarnemingen “in het midden” en treden er ook geen vreemde uitschieters op. We kunnen in dit geval concluderen dat de gegevens uit de steekproef zich niet verzetten tegen de veronderstelling dat het vetgehalte van kroketten normaal verdeeld is. Hoewel dit is iets anders dan een bewijs van normaliteit, zullen we deze aanname verder hanteren.

- b. We noemen het vetgehalte van een willekeurige kroket X en veronderstellen dat X normaal verdeeld is met gemiddelde μ en variantie σ^2 . We toetsen: $H_0 : \mu = 20$ tegen het alternatief $H_1 : \mu < 20$
De functie **T-Test** met $\mu_0=20$, **List** L1, **Freq** 1 en $< \mu_0$ levert een p-waarde op van 0,079. We kunnen de hypothese dus niet verwerpen. De fabrikant kan de bewering van Lisa niet bestrijden bij een significantieniveau van 5%.
42. a. $SEM = 12,4/\sqrt{12} \approx 3,58$; $t_{[11],0,05} \approx \text{invT}(0.975, 11) = 2,20$;
 $244,3 \pm 3,58 \cdot 2,20$ geeft $[236,4 ; 252,2]$.
Via de GR: **Tinterval** met **Stats** en \bar{x} : 244.3 , **Sx**: 12.4 , **n**: 12, **C-level**: 0.95 geeft hetzelfde resultaat.
- b. μ is de verwachtingswaarde van het (normaal verdeelde) aantal kcal in een willekeurig exemplaar van het diepvriesproduct. We toetsen
 $H_0 : \mu = 240$ tegen $H_1 : \mu \neq 240$.
Bij 11 vrijheidsgraden is de overschrijdingskans
 $\Pr(T \geq (244,3 - 240)/3,58) \approx \text{tcdf}(1.2011, 10^99) = 0,127 > \frac{1}{2}\alpha$, dus niet verwerpen.
Via de GR: **T-Test** met dezelfde invoer als bij a. plus $\mu_0 = 240$ en $\mu \neq \mu_0$ geeft een p-waarde van 0,255, hetgeen (zoals te verwachten bij een tweezijdige toets) het dubbele is van de eerder berekende overschrijdingskans.
- c. Benadering met normale verdeling: veronderstel X is n.v. met parameters 244,3 en 12,4.
Dan geldt $\Pr(X > 250) \approx \text{normalcdf}(250, 10^99, 244.3, 12.4) = 0,323$. (Dit is niet correct, omdat je werkt met geschatte waarden van μ en σ .)
Via de t -verdeling: $(X - \bar{X})/(\sigma\sqrt{1 + \frac{1}{12}})$ is standaardnormaal verdeeld.
Uit 2.1 volgt dat $11s^2/\sigma^2$ een χ^2 -verdeling met 11 vrijheidsgraden heeft en onafhankelijk is van $(X - \bar{X})/(\sigma\sqrt{1 + \frac{1}{12}})$. Hieruit volgt (zie 1.5.2) dat $T = (X - \bar{X})/(s\sqrt{1 + \frac{1}{12}})$ een t -verdeling heeft met 11 vrijheidsgraden.
Dus $\Pr(T > (250 - 244,3)/(12,4\sqrt{1 + \frac{1}{12}})) \approx \text{tcdf}(0.4416, 10^99, 11) = 0,334$.
43. Onder H_0 zijn \bar{X}_1 en \bar{X}_2 onafhankelijk en normaal verdeeld met gemiddelde $\mu_1 = \mu_2$ en standaardafwijking $\sigma_1/\sqrt{n_1}$ respectievelijk $\sigma_2/\sqrt{n_2}$. Daarom is $\bar{X}_1 - \bar{X}_2$ ook normaal verdeeld met gemiddelde $\mu_1 - \mu_2 = 0$ en standaardafwijking $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Hieruit volgt het gestelde.

44. a. Er geldt: $\frac{(n_1 - 1)s_1^2}{\sigma^2}$ heeft een χ^2 -verdeling heeft met $n_1 - 1$ vrijheidsgraden en $\frac{(n_2 - 1)s_2^2}{\sigma^2}$ heeft een χ^2 -verdeling heeft met $n_2 - 1$ vrijheidsgraden.
- b. s_1^2 en s_2^2 zijn onafhankelijk. Daarom volgt uit a. en paragraaf 1.5.2 dat $\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{\sigma^2} = \frac{(n_1 + n_2 - 2)s^2}{\sigma^2}$ een χ^2 -verdeling heeft met $n_1 + n_2 - 2$ vrijheidsgraden.
- c. T is te schrijven als een breuk waarvan de teller gelijk is aan $\frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{1/n_1 + 1/n_2}}$ en de noemer $\frac{s}{\sigma} = \sqrt{\frac{s^2(n_1 + n_2 - 2)}{\sigma^2(n_1 + n_2 - 2)}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{\sigma^2(n_1 + n_2 - 2)}}$

$$= \sqrt{\frac{1}{n_1 + n_2 - 2} \left(\frac{(n_1 - 1)s_1^2}{\sigma^2} + \frac{(n_2 - 1)s_2^2}{\sigma^2} \right)}.$$
Onder H_0 is de teller standaardnormaal verdeeld.
In de uitdrukking voor de noemer is $\frac{(n_1 - 1)s_1^2}{\sigma^2} + \frac{(n_2 - 1)s_2^2}{\sigma^2}$ de som van twee onafhankelijke χ^2 -variabelen met $n_1 - 1$ respectievelijk $n_2 - 1$ vrijheidsgraden. Deze som heeft dus een χ^2 -verdeling met $n_1 + n_2 - 2$ vrijheidsgraden (zie 1.5.1). De noemer is dus de wortel van een χ^2 variabele gedeeld door zijn aantal vrijheidsgraden. Bovendien zijn teller en noemer onafhankelijk (zie paragraaf 2.1). Uit paragraaf 1.5.2 volgt daarom het gestelde.
45. Het betrouwbaarheidsinterval voor $\mu_1 - \mu_2$ wordt gegeven door $\bar{X}_1 - \bar{X}_2 \pm t \cdot s \sqrt{1/n_1 + 1/n_2}$. Hierin is t het getal waarvoor geldt $\Pr(-t \leq T \leq t) = \alpha$, waarbij T een t -verdeling heeft met $n_1 + n_2 - 2$ vrijheidsgraden.
46. a. Sla de waarnemingen “Nooit” op in L1 en “Ooit” in L2
Via **STAT - CALC - 1-Var Stats L1** vind je voor groep 1 “Nooit”:
 $\bar{x} = 1,080$; $s = 0,1597$ en een mediaan van 1,11
Voor groep 2 “Ooit” vind je via **1-Var Stats L2** :
 $\bar{x} = 1,004$; $s = 0,1391$ en een mediaan van 0,975.
De boxplots zien er redelijk symmetrisch uit zonder uitschieters, de medianen liggen in de buurt van het gemiddelde en de standaardafwijkingen verschillen niet te veel. We kunnen dus de standaardveronderstellingen van paragraaf 3.2 hanteren.
- b. We toetsen $H_0 : \mu_1 = \mu_2$ tegen $H_1 : \mu_1 > \mu_2$.
Handmatig: T heeft een t -verdeling met 18 vrijheidsgraden en we vinden

$s = \sqrt{(9 \cdot 0,1597^2 + 9 \cdot 0,1391^2)/18} \approx 0,1497$ en

$T = 0,076/(0,1497\sqrt{10^{-1} + 10^{-1}}) \approx 1,1348$.

De overschrijdingkans $\Pr(T \geq 1,1348) \approx \text{tcdf}(1.1348, 10^9, 18) = 0,136 > \alpha$.

Dus we kunnen de hypothese niet verwerpen.

Met de GR: `2-SampleTTest` met `Data L1` en `L2`, `Freq1=1` , `Freq2=1` , $\mu_1 > \mu_2$ en `Pooled: Yes` geeft als p-waarde dezelfde overschrijdingskans.

- c. Kies je bij `Pooled: No`, dan voert de GR de in opmerking 3.2.2 aangeduide toets uit die rekening houdt met ongelijke varianties in de populaties. Deze toets geeft eveneens een p-waarde van 0,136.
 - d. `2-SampleTInt` met `Data (etc)`, `C-level 0.9` en `Pooled: Yes` geeft $[-0,04 ; 0,19]$.
 Bij b. hebben we gezien dat de eenzijdige hypothese met $\alpha = 0,05$ niet kon worden verworpen. Dat betekent dat de tweezijdige hypothese met $\alpha = 0,1$ ook niet kan worden verworpen en dat het getal 0 in het 90%-betrouwbaarheidsinterval ligt.
47. a. De aantallen in beide groepen zijn voldoende groot om het gebruik van de t-toets te rechtvaardigen, zeker omdat uitschieters zich niet voor kunnen doen (zie opmerking 3.2.3).
- b. Noem de score van een willekeurige “knappe pasfoto” K en de score van een willekeurige lelijke pasfoto L . Noem verder $E(K) = \mu_K$ en $E(L) = \mu_L$. We veronderstellen verder $Var(K) = Var(L) = \sigma^2$. Deze veronderstelling is niet onredelijk in het licht van de waarnemingen.
 We toetsen $H_0 : \mu_K = \mu_L$ tegen $H_1 : \mu_K \neq \mu_L$ en gebruiken de ongepaarde t-toets.
 Voer in `L1: 1, 2, 3, 4; L2: 0, 4, 15, 6` en `L3: 2, 5, 14, 1`.
`2-SampleTTest` met `Data, List1: L1, List2: L1, Freq1: L2, Freq2:L3, $\mu_1 \neq \mu_2$, Pooled: Yes` geeft een p-waarde van 0,031. We kunnen de hypothese verwerpen. De gemiddelde scores (3,08 en 2,64) verschillen significant.
`2-SampTInt` met dezelfde invoer en `C-level: 0.95` geeft als 95%-betrouwbaarheids-interval voor $\mu_M - \mu_L$: $[0,042 ; 0,845]$.
48. a. Gepaard
 b. Ongepaard
 c. Ongepaard
 d. Gepaard

49. a. Het gaat hier om gepaarde gegevens. Eén proefpersoon bekijkt 3 knappe en 3 lelijke foto's. Als je op deze gegevens de 2-Sample t-test loslaat, handel je op twee punten in strijd met de aannames. Noem de scores van de "knappe" en de "lelijke" foto's X_i resp. Y_i ($i = 1, \dots, 36$). De X_i 's zijn niet onafhankelijk omdat per drielik de scores door dezelfde proefpersoon worden vastgesteld. Idem voor de Y_i 's. Om dezelfde reden zijn de X_i 's en de Y_i 's niet onderling onafhankelijk.
- b. De waarnemingen zijn realisaties van een kansvariabele die we X noemen (andere X dan bij onderdeel a.). De X_i 's zijn onderling onafhankelijk. Zijn ze normaal verdeeld? Iedere X_i is de som van zes onafhankelijke kansvariabelen. Dat aantal is te klein om daaruit te concluderen dat X normaal verdeeld is. Het aantal van 12 waarnemingen is ook te klein om daaruit te rechtvaardigen dat we met niet normaal verdeelde waarnemingen toch een t-test uitvoeren. Het histogram van de x_i 's ziet er echter mooi "normaal" uit. Daarom (en vooral omdat we op dit ogenblik geen beter alternatief voorhanden hebben) voeren we een t-toets uit. Met $E(X) = \mu$ toetsen we $H_0 : \mu = 0$ tegen $H_1 : \mu > 0$. Het eenzijdige alternatief vindt zijn rechtvaardiging in het feit dat we verwachten dat knappe mensen sympathieker worden gevonden dan lelijke mensen en omdat deze verwachting is bevestigd door het experiment met één proefpersoon in opgave 47. Een keuze voor $H_1 : \mu \neq 0$ zou conservatief zijn, maar ook te verdedigen. Voer de gegevens in in L1. T-Test geeft $\bar{x} = 1,17$ $s = 1,85$ $t = 2,18$ bij 11 vrijheidsgraden en een p-waarde van 0,026. We kunnen de nulhypothese verwerpen. Knappe mensen worden op basis van hun pasfoto door 16-18-jarige gymnasiasten als sympathieker beoordeeld dan lelijke mensen.
50. Deze opgave bevat deels gepaarde en deels ongepaarde gegevens. De gegevens voor mannen zijn gepaard, die voor vrouwen ook. Als we mannen en vrouwen willen vergelijken hebben we te maken met ongepaarde gegevens.
- a. Definieer X als het verschil (na min voor) bij mannen. Y is hetzelfde bij vrouwen. Twee maal TInt gebruiken levert: $\bar{x} = 13,625$; $s_X = 8,280$; $\bar{y} = 8,0$; $s_Y = 6,403$. Voor mannen is het 90% betrouwbaarheidsinterval voor μ_X $[8,1; 19,2]$, Voor vrouwen is het 90% betrouwbaarheidsinterval voor μ_Y $[3,3; 12,7]$.

- b. De betrouwbaarheidsintervallen suggereren dat er een verschil is. Voor een formele toets moeten we een vergelijking maken van de (ongepaarde) variabelen X en Y . We toetsen $H_0 : \mu_X = \mu_Y$ tegen $H_1 : \mu_X \neq \mu_Y$. De gevonden steekproefstandaardafwijkingen liggen voldoende dicht bij elkaar om te veronderstellen dat de onderliggende parameters aan elkaar gelijk zijn. Gebruik van 2 - `SampleTTest` levert: $s = 7,472$; $t = 1,454$ (13 vrijheidsgraden) en de (tweezijdige)overschrijdingskans $p = 0,17$. We kunnen de hypothese van gelijke verschillen bij mannen en vrouwen dus niet verwerpen.

4 Hoofdstuk 4

51. Uitgaande van $G_1 - G_0$ zijn er 12 “+”, 4 “-” en 4 “0”. Noem het aantal “+” X .
 X is binomiaal verdeeld met $n' = 16$ en onbekende π . We toetsen $H_0 : \pi = \frac{1}{2}$ tegen $H_1 : \pi > \frac{1}{2}$. De overschrijdingskans is $\Pr(X \geq 12) = 1 - \Pr(X \leq 11) = 1 - \text{binomcdf}(16, 0.5, 11) \approx 0,038 < \alpha$. We kunnen de hypothese verwerpen. We concluderen dat de rekenmethode een positief effect heeft op de prestaties van de kinderen.
52. De gebeurtenis $S = 192$ treedt op als de volgende rangnummers een negatief teken hebben: 9 of (1,8) of (2,7) of (3,6) of (4,5) of (1,2,6) of (1,3,5) of (2,3,4). Dus $\Pr(S = 190) = 8 \cdot (\frac{1}{2})^{20}$.
53. Bij een eenzijdige toets met $\alpha = 0,05$ komt de kritieke waarde van S overeen met de kritieke waarde van een tweezijdige toets met $\alpha = 0,1$, in dit geval dus met 90.
54. Als $n' = 7$ is de kans op de maximale en minimale waarde van S gelijk aan $2 \cdot (\frac{1}{2})^7 \approx 0,015 > \alpha = 0,01$. Voor $n' = 7$ en $\alpha = 0,01$ bestaat dus geen kritieke waarde. Voor $n' = 8$ en $\alpha = 0,01$ bestaat wel een kritieke waarde, want $2 \cdot (\frac{1}{2})^8 \approx 0,0078 < 0,01$. Idem voor de andere waarden van α .
55. a. Onder de nulhypothese hebben T_+ en T_- dezelfde verdeling, dus je verwacht dat de helft van de som van de rangnummers terecht komt bij T_+ en de andere helft bij T_- . Met andere woorden $E(T_+) = E(T_-) = \frac{1}{4}n'(n' + 1)$.
 b. $T_- = \frac{1}{2}n'(n' + 1) - T_+$
 c. $S = T_+ - T_- = 2T_+ - \frac{1}{2}n'(n' + 1)$.
 Dus $\text{Var}(S) = \text{Var}(2T_+ - \frac{1}{2}n'(n' + 1)) = \text{Var}(2T_+) = 4\text{Var}(T_+)$.
 Dus $\text{Var}(T_+) = \frac{1}{4}\text{Var}(S) = \frac{1}{24}n'(n' + 1)(2n' + 1)$
 Uiteraard geldt $\text{Var}(T_-) = \text{Var}(T_+)$.
56. Er geldt $n' = 37 - 2 = 35$ en S is bij benadering normaal verdeeld met onder de nulhypothese $E(S) = 0$ en $\text{Var}(S) = \frac{1}{6} \cdot 35 \cdot 36 \cdot 71 = 14910$. Met continuïteitscorrectie:
 $p = 2 \cdot \Pr(S \geq 316) = 2 \cdot \Pr(S \geq 315) \approx 2 \cdot \text{normalcdf}(315, 10^{\wedge}99, 0, \sqrt{14910}) = 0,0099$
57. a. Op basis van Voor min Na tellen we 10 “plussen” 1 “nul” en 4 “minnen”. Uitgaande van 14 verschillen $\neq 0$ noemen we het aantal “minnen” X .
 X is binomiaal verdeeld met $n' = 14$ en π . We toetsen $H_0 : \pi = \frac{1}{2}$ tegen $H_1 : \pi < \frac{1}{2}$.

$\Pr(X \leq 4) \approx \text{binomcdf}(14, 0.5, 4) = 0,090 > \alpha$. Het aantal storingen is niet significant afgenomen. De p-waarde is 0,090.

- b. We kunnen de verschillen Voor min Na als volgt van rangnummers voorzien:

Verschil	3	4	3	0	-1	3	-1	4
Rang met teken	10,5	13,5	10,5	*	-2,5	10,5	-2,5	13,5

Verschil	2	-1	1	2	2	-2	3
Rang met teken	6,5	-2,5	2,5	6,5	6,5	-6,5	10,5

We vinden $S = 77$ bij $n' = 14$. We toetsen eenzijdig, dus we kijken in de tabel bij $\alpha = 0,1$. $77 > 55$ dus we kunnen de nulhypothese verwerpen. Uit de tabel kunnen we aflezen dat 77 groter is dan de grenswaarde uit de kolom " $\alpha = 0.02$ " en kleiner dan de grenswaarde uit de kolom " $\alpha = 0.01$ ".

Omdat we eenzijdig toetsen, concluderen we dat de p-waarde van deze toets tussen 0,005 en 0,01 ligt.

- c. Er zijn veel knopen, dus de kritieke grenzen uit de tabel zijn niet exact, maar conservatief. De werkelijke p-waarde is vermoedelijk kleiner.

58. We toetsen de hypothese H_0 : Hartslag "Voor" en "Na" hebben dezelfde verdeling tegen

H_1 : Hartslag "Na" is t.o.v. hartslag "Voor" verschoven naar rechts.

We kijken naar de verschillen "Na"- "Voor". Als we deze rangschikken van kleine absolute waarde naar grote absolute waarde en deze verschillen vervolgens voorzien van een rangnummer met teken, krijgen we:

Verschil	-1	-2	-2	2	4	4	-5	-7	7	7	8	14
Rangnummer met teken	-1	-3	-3	3	5,5	5,5	-7	-9	9	9	11	12

En $S = -1 - 3 - 3 + \dots + 12 = 32$. We toetsen eenzijdig met $\alpha = 0,05$. In de tabel van Appendix 1 kijken we bij $n' = 12$ en $\alpha = 0,1$ en vinden we de kritieke waarde 44. We kunnen de hypothese dus niet verwerpen. (Opmerking: vanwege het grote aantal knopen is de toets niet exact, maar eerder conservatief.)

59. We kunnen de verschillen als volgt van rangnummers voorzien:

Verschil	2	5	-1	0	1	-2	3	0	1	1	2	2
Rang met teken	6,5	10	-2,5	*	2,5	-6,5	9	*	2,5	2,5	6,5	6,5

We vinden $S = 37$ en $n' = 10$. We toetsen eenzijdig, dus we kijken in de tabel bij $\alpha = 0,1$. $37 > 35$, dus we kunnen de nulhypothese

verwerpen. Dit resultaat komt overeen met dat van de t-toets.

60. Op de volgende 5 manieren kan een selectie van 8 rangnummers uit 1 t/m 17 een som van 40 hebben:

$$1+2+\dots+4+6+7+8+9$$

$$1+2+\dots+5+7+8+10$$

$$1+2+\dots+6+7+12$$

$$1+2+\dots+6+8+11$$

$$1+2+\dots+6+9+10$$

$$\text{Dus } \Pr(S_Y = 40) = 5 / \binom{17}{8} = 5/24310 \approx 2 \cdot 10^{-4}$$

61. De rangnummers zijn $1, 2, \dots, n_1 + n_2$. De gemiddelde waarde van een rangnummer is dus $\frac{1}{2}(1 + n_1 + n_2)$. Onder de nulhypothese heeft ieder rangnummer dezelfde kans om aan één van de n_1 waarnemingen van X_1 te worden toegewezen. De verwachte waarde van de som van deze rangnummers is dus $n_1 \cdot \frac{1}{2}(1 + n_1 + n_2)$.

62. De kleinst mogelijke waarde van S_{X_1} is $1 + \dots + n_1 = 1 + 2 = 3$.

$$\Pr(S_{X_1} = 3) = 1 / \binom{9}{2} = 1/36 > \frac{1}{2}\alpha, \text{ dus zelfs } 3 \text{ is te groot als kritieke waarde.}$$

63. S_B is bij benadering normaal verdeeld met $E(S_B) = \frac{1}{2} \cdot 7 \cdot (1 + 7 + 12) = 70$ en $Var(S_B) = \frac{1}{12} \cdot 7 \cdot 12 \cdot (1 + 7 + 12) = 140$.

Dus de p-waarde van de tweezijdige toets is (met continuïteitscorrectie):

$$2 \cdot \Pr(S_B \leq 49) = 2 \cdot \Pr(S_B \leq 49,5) \approx 2 \cdot \text{normalcdf}(-10^{99}, 49.5, 70, \sqrt{140}) = 0,0831.$$

64. We toetsen H_0 : “De waarnemingen van K en N komen uit dezelfde verdeling” tegen het eenzijdige alternatief H_1 : “De verdeling is een naar rechts verschoven versie van de verdeling van N .” Hieronder zie je de waarnemingen voorzien van rangnummers. We vinden $S_N = 36,5$. In de tabel vinden we bij $\alpha = 0,1$ (eenzijdige toets): $l = 41$. We kunnen de nulhypothese verwerpen. Ook in de kolom $\alpha = 0,05$ vinden we een linkergrens die groter is dan 36,5, maar in de kolom $\alpha = 0,02$ is dit net niet meer het geval. De p-waarde van de toets ligt dus tussen 0,025 en 0,01.

K	rang	N	rang
10,3	9	7,2	2
11,8	13	9,5	5
9,8	6,5	11,2	11
12,6	15	8,0	3
11,4	12	6,9	1
8,3	4	10,1	8
12,0	14	9,8	6,5
10,9	10		

65. a. $\bar{P} = 11,4$ en $\bar{N} = 12,4$
b.

P	rang	N	rang
12,3	8	12,7	10
10,8	2	13,0	11
11,5	6	10,9	3
11,0	4	11,2	5
10,7	1	13,2	12
12,0	7	13,4	13
		12,5	9

We toetsen H_0 : “ P en N hebben dezelfde verdeling” tegen het tweezijdig alternatief H_1 : “ P en N hebben niet dezelfde verdeling”. Hierboven zie je de waarnemingen voorzien van rangnummers. We vinden $S_P = 28$. In de tabel vinden we $l = 27$, we kunnen de nulhypothese niet verwerpen. We kunnen niet concluderen dat het placebo de prestatie beïnvloedt.

- c. De spreiding van de gegevens zorgt ervoor dat een verschil in gemiddelde tijd van 1,1 seconde niet significant is. De spreiding wordt veroorzaakt door het feit dat het ene kind sneller kan lopen dan het andere, los van een eventueel effect van het placebo. Deze spreiding kan worden verminderd door gepaarde data te verzamelen: je laat ieder kind twee keer lopen, één keer met en één keer zonder placebo. Je moet er dan wel voor zorgdragen dat de omstandigheden per race zoveel mogelijk gelijk zijn door bijvoorbeeld de helft van de kinderen eerst met en daarna zonder placebo te laten lopen en de andere helft andersom.
66. a. $\overline{BmD} = 8,0$ $\overline{D} = 7,5$ $\overline{BzD} = 6,9$
b. We hebben te maken met gepaarde waarnemingen. Van de 12 verschillen $Bmd - D$ zijn er 2 nul, 1 negatief en 9 positief. De tekentoets levert al een significant resultaat: we toetsen H_0 : “De kans op een negatief verschil is 0,5” tegen het tweezijdige alternatief H_1 : “De kans op een negatief verschil verschilt van

0,5”.

Noem het aantal negatieve verschillen X . We vinden de p-waarde: $2 \cdot \Pr(X \leq 2) \approx 2 \cdot \text{binomcdf}(10, 0.5, 1) = 0,021$. We kunnen de nulhypothese verwerpen.

- c. We hebben te maken met ongepaarde gegevens. We gebruiken de toets van Wilcoxon om de nulhypothese H_0 : “ BmD en BzD hebben dezelfde verdeling” tegen het tweezijdige alternatief H_1 : “De verdeling van BmD verschilt van de verdeling van BzD ”.

In de tabel hieronder zijn de waarnemingen van een rangnummer voorzien.

BmD	rang	BzD	rang
6,0	3	4,9	1
6,1	4	5,8	2
6,8	8	6,2	5
7,0	10	6,3	6
8,0	16	6,5	7
8,1	17	6,9	9
8,4	18,5	7,2	11
8,4	18,5	7,3	12,5
8,5	20	7,3	12,5
9,0	21	7,4	14
9,6	23	7,5	15
9,7	24	9,2	22

We vinden $S_{BzD} = 117$. Onder de nulhypothese is S_{BzD} bij benadering normaal verdeeld met $E(S_{BzD}) = \frac{1}{2} \cdot 12 \cdot 25 = 150$ en $Var(S_{BzD}) = \frac{1}{12} \cdot 12 \cdot 12 \cdot 25 = 300$.

Met continuïteitscorrectie vinden we als p-waarde: $2 \cdot \Pr(S_{BzD} \leq 117) = 2 \cdot \Pr(S_{BzD} \leq 117,5) \approx 2 \cdot \text{normalcdf}(-10^{\wedge}99, 117.5, 150, \sqrt{300}) = 0,061$.

We kunnen de nulhypothese niet verwerpen.

- d. De verschillen tussen BmD en D zijn (ondanks het gebruik van een “zwakke” toets) significant bij een verschil in gemiddeld cijfer van 0,5. De verschillen tussen BmD en BzD zijn (ondanks het gebruik van een “sterke” toets) niet significant bij een groter verschil in gemiddeld cijfer: 1,1. Dit opmerkelijke verschil wordt veroorzaakt door het verschil tussen gepaarde en ongepaarde waarnemingen. Bij gepaarde waarnemingen wordt de spreiding afkomstig van het niveauverschil tussen de ene leerling en de andere geëlimineerd.

5 Hoofdstuk 5

67. a. Onder de nulhypothese geldt: X_i is binomiaal verdeeld met parameters n_i en π (i is 1 of 2). Er geldt dus:
 $E(X_i) = n_i\pi$ en $Var(X_i) = n_i\pi(1 - \pi)$. Hieruit volgt:
 $E(p_i) = E(X_i/n_i) = n_i\pi/n_i = \pi$ en $Var(p_i) = Var(X_i/n_i) = n_i\pi(1 - \pi)/n_i^2 = \pi(1 - \pi)/n_i$.
- b. X_1 en X_2 zijn onafhankelijk (resultaten van een steekproef uit aselekt samengestelde populaties), dus p_1 en p_2 zijn onafhankelijk. Er geldt:
 $E(p_1 - p_2) = E(p_1) - E(p_2) = \pi - \pi = 0$ (onafhankelijkheid niet vereist) en
 $Var(p_1 - p_2) = Var(p_1) + Var(p_2) = \pi(1 - \pi)(n_1^{-1} + n_2^{-1})$ (onafhankelijkheid wél vereist).
- c. Bij “grote” waarden van n_i zijn X_i en dus p_i bij benadering normaal verdeeld. Dit geldt dus ook voor $p_1 - p_2$. Deling door de standaardafwijking (wortel van de variantie) en vervanging van π door zijn schatting p levert het gestelde.
68. Groep 1 meldt zich aan via de website, groep 2 telefonisch. Met $\pi_{(i)}$ geven we de kans op tevredenheid in de betreffende (sub)groep aan. We toetsen $H_0 : \pi_1 = \pi_2$ tegen $H_1 : \pi_1 < \pi_2$. De aantallen zijn groot genoeg om de benadering via de normale verdeling te gebruiken. We vinden:
 $p_1 = 59/80 \approx 0,7375$ $p_2 = 48/60 = 0,8$
 $p = (59 + 48)/(80 + 60) \approx 0,7643$
 $Z \approx (0,7375 - 0,8)/\sqrt{0,7643 \cdot 0,2357(80^{-1} + 60^{-1})} \approx -0,862$.
 Hierbij hoort een p -waarde van $\text{normalcdf}(-10^99, -0.862) \approx 0,194$ (eenzijdig). We kunnen de hypothese niet verwerpen.
 Met de GR gebruiken we **2-PropZTest**. We voeren in: `x1:59 n1:80 x2:48 n2:60 p1<p2`. **Calculate** geeft de hiervoor vermelde waarden.
69. Noem de kans dat een onderbouw- respectievelijk bovenbouwleerling lid is van de leerlingenvereniging π_1 respectievelijk π_2 . We toetsen $H_0 : \pi_1 = \pi_2$ tegen $H_1 : \pi_1 \neq \pi_2$. Verder geldt $n_1 = 185 + 86 = 271$ en $n_2 = 152 + 90 = 242$. De aantallen zijn groot genoeg om de benadering via de normale verdeling te gebruiken. We gebruiken **2-PropZTest** en voeren in: `x1:185 n1:271 x2:152 n2:242 p1≠p2`. **Calculate** geeft: $Z = 1,299$ en de bijbehorende p -waarde van 0,194. We kunnen de nulhypothese niet verwerpen. Er is geen significant verschil.
70. Noem de kans dat een allochtone leerling minstens één vulling heeft π_1 . Bij een autochtone leerling is deze kans π_2 . We toetsen $H_0 : \pi_1 = \pi_2$

tegen $H_1 : \pi_1 \neq \pi_2$. De waargenomen aantallen zijn te klein om de benadering via de normale verdeling toe te passen ($n_1 p = \frac{24 \cdot 26}{126} \approx 4,95 < 5$), dus we gebruiken Fishers exacte toets. De “vaas” bestaat uit $24 + 102 = 126$ leerlingen waarvan 26 met en 100 zonder vulling(en). Onder de nulhypothese zijn de allochtone leerlingen op te vatten als een “greep zonder terugleggen” van 24 willekeurige leerlingen uit de vaas. We berekenen de kans dat hier 2 of minder leerlingen met vulling(en) bij zitten. Deze kans is gelijk aan:

$$\frac{\binom{26}{0} \binom{100}{24} + \binom{26}{1} \binom{100}{23} + \binom{26}{2} \binom{100}{22}}{\binom{126}{24}} \approx 0,0782.$$

De p -waarde van de toets is dus $2 \cdot 0,0782 \approx 0,156 > \alpha$. We kunnen de nulhypothese niet verwerpen. Er is geen significant verschil in de gebitsstoestand tussen beide groepen leerlingen.

71. Noem de kans op 1 of meer hersenschuddingen π_1 bij voetballers en π_2 bij andere sporters.

We toetsen $H_0 : \pi_1 = \pi_2$ tegen $H_1 : \pi_1 > \pi_2$. Deze aantallen zijn groot genoeg om de benadering via de normale verdeling te kunnen gebruiken. We gebruiken `2-PropZTest` en voeren in: `: x1:45 n1:91 x2:28 n2:96 p1>p2. Calculate` geeft $Z = 2,842$ en een p -waarde van $0,0022 < \alpha$. We verwerpen de nulhypothese.

Voetballers hebben significant vaker een hersenschudding dan andere sporters.

72. a. Voor mensen boven deze inkomensgrens geldt dat ze zich (financieel) een auto kunnen veroorloven. Zonder een dergelijke grens zou in eventueel verschil in de mate van autobezit kunnen worden veroorzaakt door een eventueel verschil in gemiddeld inkomen tussen de twee partijen.

- b. Noem de kans op het niet bezitten van een auto π_1 bij leden van de PvdA met een inkomen van 75 000 euro of meer en π_2 bij soortgelijke leden van GroenLinks. We toetsen $H_0 : \pi_1 = \pi_2$ tegen $H_1 : \pi_1 \neq \pi_2$.

Er geldt $n_2 p = 37 \cdot 11/96 \approx 4,2 < 5$, dus we gebruiken Fishers exacte toets. De “vaas” bestaat uit 96 personen: 85 hebben een auto en 11 niet. Onder de nulhypothese zijn de leden van de PvdA op te vatten als een “greep zonder terugleggen” van 59 willekeurige personen uit de vaas. We berekenen de kans dat hier 4 of minder personen zonder auto bij zitten. Deze kans is gelijk aan:

$$\frac{\binom{11}{0}\binom{85}{59} + \binom{11}{1}\binom{85}{58} + \binom{11}{2}\binom{85}{57} + \binom{11}{3}\binom{85}{56} + \binom{11}{4}\binom{85}{55}}{\binom{96}{59}} \approx 0,070.$$

De p -waarde van deze toets is dus $2 \cdot 0,070 \approx 0,14 > \alpha$. We kunnen de nulhypothese niet verwerpen. Het verschil tussen de desbetreffende leden van de PvdA en GroenLinks is niet significant.

- c. De hoofdvraag ging over het “persoonlijk in praktijk brengen van milieubewustzijn”; het onderzoek spitst zich toe op autobezit, en dan nog wel op het meer gestelde deel van de partijleden. Dat is veel minder ruim.
73. a. $E(p_1) = \pi_1 \quad E(p_2) = \pi_2 \quad Var(p_1) = \pi_1(1 - \pi_1)/n_1$
 $Var(p_2) = \pi_2(1 - \pi_2)/n_2$
- b. p_1 en p_2 zijn onafhankelijk, dus $p_1 - p_2$ is bij benadering normaal verdeeld met verwachting $\pi_1 - \pi_2$ en variantie $\pi_1(1 - \pi_1)/n_1 + \pi_2(1 - \pi_2)/n_2$.
74. Het betrouwbaarheidsinterval wordt gegeven door
 $p_1 - p_2 \pm z_{0,05} \cdot \sigma(p_1 - p_2)$. Met $p_1 - p_2 = 60/100 - 88/104 \approx -0,24615$,
 $z_{0,05} \approx \text{invNorm}(0,975) = 1,96$ en $\sigma(p_1 - p_2) = \sqrt{\frac{0,6 \cdot 0,4}{100} + \frac{88/104 \cdot 16/104}{104}} \approx 0,0604$ wordt het interval $[-0,36 ; -0,13]$.
 Je kunt ook de functie **2-PropZInt** gebruiken. Voer in: **x1:60 n1:100 x2:88 n2:104 C-level:0.95**. **Calculate** geeft hetzelfde resultaat.
75. a. Het betrouwbaarheidsinterval wordt gegeven door
 $p_1 - p_2 \pm z_{0,1} \cdot \sigma(p_1 - p_2)$. Met $p_1 - p_2 = 106/143 - 67/119 \approx 0,1782$,
 $z_{0,1} \approx \text{invNorm}(0,95) = 1,645$ en
 $\sigma(p_1 - p_2) = \sqrt{\frac{106/143 \cdot 37/143}{143} + \frac{67/119 \cdot 52/119}{119}} \approx 0,05838$ wordt het interval $[0,08 ; 0,27]$.
 Je kunt ook de functie **2-PropZInt** gebruiken. Voer in: **x1:106 n1:143 x2:67 n2:119 C-level:0.9**. **Calculate** geeft hetzelfde resultaat.
- b. We toetsen $H_0 : \pi_1 = \pi_2$ tegen $H_1 : \pi_1 \neq \pi_2$
 Voor de toets gebaseerd op Z vinden we $p = \frac{106+67}{143+119} \approx 0,660$ en
 $Z \approx \frac{0,1782}{\sqrt{0,660 \cdot 0,340(143^{-1} + 119^{-1})}} \approx 3,03$. De bijbehorende p -waarde (tweezijdig) is $2 \cdot \text{normalcdf}(3,03, 10^99) = 0,0024 < 0,1$. We kunnen de hypothese dus verwerpen.
 Voor de toets gebaseerd op Z' kunnen we direct concluderen dat we de hypothese kunnen verwerpen, omdat 0 niet in het bij a. gevonden betrouwbaarheidsinterval ligt. Er geldt

$$\sigma(p_1 - p_2) = \sqrt{\frac{106 \cdot 37}{143^3} + \frac{67 \cdot 52}{119^3}} \approx 0,05838 \text{ en } Z' \approx \frac{0,1782}{0,0583} \approx 3,05.$$

De p -waarde (tweezijdig) is $2 \cdot \text{normalcdf}(3.06, 10^{-99}) = 0,0022$.

- c. We hebben nu $H_1 : \pi_1 > \pi_2$. De waarden van Z en Z' veranderen uiteraard niet. De p -waarden kunnen nu worden gehalveerd.
 - d. De steekproef is genomen in slechts twee gemeenten. Geconstateerde verschillen in de proportie van kinderen met cariës worden dus niet noodzakelijkerwijs veroorzaakt door de aan- of afwezigheid van fluor in het drinkwater. Een andere oorzaak zou kunnen zijn: het verschil in eetgewoonten tussen beide gemeenten, veroorzaakt door verschillen sociale achtergrond, lokaal voedselaanbod e.d.
76. We gebruiken de notatie van de uitwerking van opgave 71. Het 95%-betrouwbaarheidsinterval voor $\pi_1 - \pi_2$ kan worden gevonden m.b.v. de functie **2-PropZInt**. Voer in: **x1:45 n1:91 x2:28 n2:96 C-level:0.95**. **Calculate** geeft $[0,07 ; 0,34]$.
77. Onder H_0 is N_{+-} binomiaal verdeeld met parameters $n' = 11$ en $\pi = \frac{1}{2}$. De gevraagde toets komt dus overeen met $H_0 : \pi = \frac{1}{2}$ tegen $H_1 : \pi \neq \frac{1}{2}$. De overschrijdingskans is: $\Pr(N_{+-} \leq 1) = \text{binomcdf}(11, 0.5, 1) = 0,006$. Dit geeft een p -waarde van $0,012 < \alpha$. Dus we verwerpen de nulhypothese. Middel B heeft een significant hogere kans om te werken dan middel A .
78. We kijken alleen naar de 12 personen die maar van één soort wijn houden. Noem het aantal personen dat wel van rood, maar niet van wit houdt X . X is binomiaal verdeeld met parameters $n' = 12$ en onbekende π . We toetsen $H_0 : \pi = \frac{1}{2}$ tegen $H_1 : \pi < \frac{1}{2}$. De overschrijdingskans (die bij deze eenzijdige toets gelijk is aan de p -waarde) is gelijk aan: $\Pr(X \leq 3) = \text{binomcdf}(12, 0.5, 3) = 0,073 > \alpha$. We kunnen Liesbeth op grond van deze waarnemingen geen gelijk geven.
79. a. We hebben te maken met gepaarde waarnemingen. We beperken ons tot de 76 leerlingen die òf Frans òf Duits volgen. Noem het aantal leerlingen dat Frans volgt X . X is binomiaal verdeeld met parameters $n' = 76$ en π . We toetsen $H_0 : \pi = \frac{1}{2}$ tegen $H_1 : \pi > \frac{1}{2}$. De p -waarde van deze toets is gelijk aan de overschrijdingskans:
 $\Pr(X \geq 45) = 1 - \Pr(X \leq 44) \approx 1 - \text{binomcdf}(76, 0.5, 44) = 0,068 > \alpha$.
 We kunnen de nulhypothese niet verwerpen. Meneer Stolts krijgt gelijk.
- b. We hebben hier te maken met ongepaarde waarnemingen. Noem de kans dat een bovenbouwleerling Frans volgt π_1 bij het Pearson

College en π_2 bij het Fisher Gymnasium. We toetsen $H_0 : \pi_1 = \pi_2$ tegen $H_1 : \pi_1 > \pi_2$. De aantallen zijn groot genoeg om de populatiepercentages te vergelijken m.b.v. de benadering via de normale verdeling.

We gebruiken **2-PropZTest** en voeren in:

x1:45+12=57 n1:12+45+31+51=139 x2:59 n2:190 p1:>p2.

Calculate geeft $Z = 1,87$ en een p -waarde van 0,03. We kunnen de nulhypothese verwerpen. Meneer Fier krijgt gelijk.

- c. We hebben opnieuw te maken met ongepaarde waarnemingen. Noem de kans dat een bovenbouwleerling Frans volgt π_1 bij het Pearson College en π_2 bij de jaarlaag van meneer Fier. We toetsen $H_0 : \pi_1 = \pi_2$ tegen $H_1 : \pi_1 < \pi_2$.

Er geldt $p = \frac{57+19}{139+23} \approx 0,469$ en $n_2 p \approx 23 \cdot 0,469 \approx 10,8 > 5$.

Het is dus niet noodzakelijk om Fishers exacte toets te gebruiken.

- d. De “vaas” bestaat uit 139+23=162 leerlingen: 57+19=76 volgden Frans en 82+4=86 niet. Onder de nulhypothese zijn de leerlingen van de jaarlaag van meneer Fier op te vatten als een “greep zonder terugleggen” van 23 willekeurige leerlingen uit de vaas. We berekenen de kans dat hier 4 of minder leerlingen bij zitten die geen Frans volgen. Deze kans is gelijk aan:

$$\frac{\binom{86}{0} \binom{76}{23} + \binom{86}{1} \binom{76}{22} + \binom{86}{2} \binom{76}{21} + \binom{86}{3} \binom{76}{20} + \binom{86}{4} \binom{76}{19}}{\binom{162}{23}}$$

$\approx 1,9 \cdot 10^{-4}$.

Dit is tevens de p -waarde van deze toets.

Voor de berekening van de p -waarde met de benadering van de normale verdeling gebruiken we **2 - PropZTest** met als invoer
x1:57 n1:139 x2:19 n2:23 p1:<p2.

Calculate geeft een p -waarde van $1,1 \cdot 10^{-4}$.

6 Hoofdstuk 6

80. a. Algemeen geldt:

$$(a + b + c)^2 = (a + b + c)(a + b + c) = a(a + b + c) + b(a + b + c) + c(a + b + c) = a^2 + ab + ac + ab + b^2 + bc + ac + bc + c^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc$$
 Toepassing hiervan (rekening houdend met tekens) levert het gevraagde.
- b. Algemeen geldt: $\sum (A_i + B_i + \dots + Z_i) = \sum A_i + \sum B_i + \dots + \sum Z_i$, omdat je in een optelsom de volgorde van de termen kunt veranderen.
 Verder geldt: $\sum \alpha' = n\alpha'$ (als je n keer α' optelt, krijg je $n\alpha'$)
 en $\sum \beta^2 x_i^2 = \beta^2 x_1^2 + \beta^2 x_2^2 + \dots + \beta^2 x_n^2 = \beta^2 (x_1^2 + x_2^2 + \dots + x_n^2) = \beta^2 \sum x_i^2$
 Herhaalde toepassing van dit principe levert het gevraagde.
- c. Per definitie geldt: $\bar{Y} = \frac{1}{n} \sum Y_i$, dus $\sum Y_i = n\bar{Y}$.
 Verder: $\sum x_i = \sum (X_i - \bar{X}) = \sum X_i - \sum \bar{X} = n\bar{X} - n\bar{X} = 0$
- d. **
81. De nieuwe Y -as is ontstaan door een translatie van de oorspronkelijke Y -as met \bar{X} .
 Het snijpunt van de lijn met de nieuwe Y -as is \bar{Y} . Als we langs de regressielijn (met rico $\hat{\beta}$) horizontaal $-\bar{X}$ opschuiven, dan schuiven we verticaal $-\hat{\beta}\bar{X}$. Het snijpunt met de oorspronkelijke Y -as is dus $\bar{Y} - \hat{\beta}\bar{X}$.
82. $\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$. De teller van deze breuk kun je als volgt uitwerken:

$$\sum x_i Y_i = \sum (X_i - \bar{X}) Y_i = \sum X_i Y_i - \sum \bar{X} Y_i = \sum X_i Y_i - \bar{X} \sum Y_i = \sum X_i Y_i - n\bar{X} \cdot \bar{Y}$$
 De noemer is als volgt te schrijven:

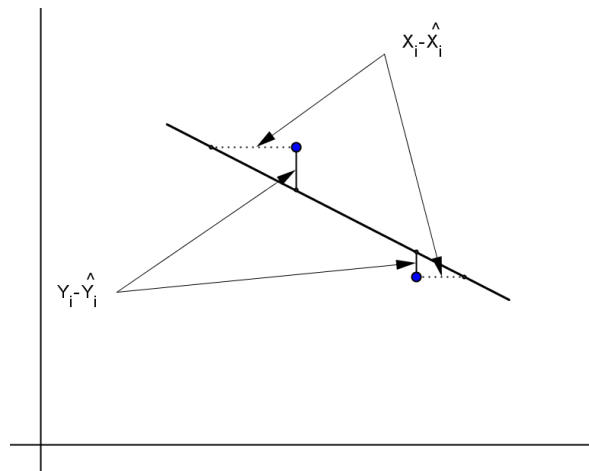
$$\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sum (X_i^2 - 2\bar{X}X_i + \bar{X}^2) = \sum X_i^2 - 2\bar{X} \sum X_i + n\bar{X}^2 = \sum X_i^2 - 2\bar{X}n\bar{X} + n\bar{X}^2 = \sum X_i^2 - n\bar{X}^2$$
 Hieruit volgt het gevraagde.
83. a. Voer met **STAT - EDIT** de Y 's in in L1 en de S 'en in L2.
LinReg(a+bx) L1, L2, Y1 geeft de regressievergelijking:
 $S = -1172 + 0,141Y$
- b. De regressievergelijking is nu opgeslagen als functie Y1. Vul nu via **STAT - EDIT** de lijst L3 met **RESID** (via **LIST - NAMES**) en je ziet dat het grootste residu optreedt bij de waarneming (18000, 2100). De grootte van het residu is 726.
- c. We vinden (via **2-VarStats**):
 $\bar{X} = 24400$, $\bar{Y} = 2280$, $\sum X_i^2 = 3146 \cdot 10^6$, $\sum X_i Y_i = 3021 \cdot 10^5$.

$$\text{Dus } \hat{\beta} = \frac{\sum X_i Y_i - n \bar{X} \cdot \bar{Y}}{\sum X_i^2 - n \bar{X}^2} = \frac{3021 \cdot 10^5 - 5 \cdot 24400 \cdot 2280}{3146 \cdot 10^6 - 5 \cdot 24400^2} = \frac{23940000}{169200000} \approx 0,141$$

$$\text{en } \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} = 2280 - 0,141 \cdot 24400 \approx -1172$$

- d. $\hat{S} = -1172 + 0,141 \cdot 30000 \approx 3100$. Met de GR: Y1(30000) geeft ongeveer 3100.

84. a. Voer met **STAT - EDIT** de X 'en in in **L1** en de Y 'en in **L2**. Op de plot (via **Plot1**) lijken de punten redelijk gespreid rond een (denkbeeldige) rechte lijn
- b. **LinReg(a+bx)** **L1, L2, Y1** geeft de regressievergelijking: $Y = 88,5 - 2,304X$.
Op de plot lijken de punten redelijk gespreid rond deze regressielijn.
- c. $\hat{Y} = 88,5 - 2,304 \cdot 10,4 \approx 64,6$ of (eenvoudiger) **Y1(10.4)** geeft 64,6. De geschatte tijd van deze atleet op de 20 km langlaufen is dus 64,6 minuten.
- d. Uit de regressievergelijking volgt $2,304X = 88,5 - Y$, beide zijden van de vergelijking delen door 2,304 levert $X = 38,4 - 0,4340Y$.
- e. **LinReg L2,L1,Y2** geeft $X = 27,8 - 0,2731Y$. (We slaan de nieuwe formule op in **L2**.)
- f. In deze opgave constateer je dat regressie van Y op X een andere lijn oplevert dan regressie van X op Y . Bij regressie van Y op X zoek je een lijn zó dat $\sum (Y_i - \hat{Y}_i)^2$ minimaal wordt. Bij regressie van X op Y zoek je een lijn zó dat $\sum (X_i - \hat{X}_i)^2$ minimaal wordt. (Zie het plaatje hieronder.) Deze twee criteria zijn verschillend en zullen over het algemeen niet dezelfde regressielijn produceren.



- g. Je wilt een voorspelling \hat{X} die zo dicht mogelijk bij X ligt. De regressielijn van X op Y heeft als criterium dat $\sum (X_i - \hat{X}_i)^2$ minimaal wordt. Dus kun je het best de lijn bij e. gebruiken.
De voorspelling wordt: $\hat{X} = 27,8 - 0,273 \cdot 70 \approx 8,6$ of (eenvoudiger) $Y2(70)$ geeft 8,6. We schatten dat deze atleet het 8,6 minuten op de loopband volhoudt.
- h. Je kunt natuurlijk klakkeloos $Y1(1)$ uitrekenen en concluderen dat je leraar tenminste 86,2 minuten nodig heeft om 20 km te langlaufen, maar dit is geen verstandige berekening. Het lineaire verband is geschat voor atleten met tijden die (op de loopband en voor het langlaufen) enigszins bij elkaar in de buurt liggen. Voor je wiskundeleraar geldt dit niet. Als je zo ver extrapoleert als in dit voorbeeld is de kans groot dat je er flink naast zit.
85. $S = \sum (X_i - \alpha)^2 = \sum (X_i^2 - 2\alpha X_i + \alpha^2) = \sum X_i^2 - 2\alpha \sum X_i + n\alpha^2$.
Hieraan kun je zien dat S een kwadratische functie van α is. Deze dalparabool heeft een minimum als $\frac{dS}{d\alpha} = -2 \sum X_i + 2n\alpha = 0$, d.w.z. als $\alpha = \frac{1}{n} \sum X_i = \bar{X}$.
Als je een aantal getallen moet vervangen door één getal, dan is de “natuurlijke” keuze om het gemiddelde van die getallen te nemen. In deze opgave heb je laten zien dat deze “natuurlijke” keuze berust op het principe van kleinste kwadraten.
86. Er geldt $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$, dus $\hat{\alpha}$ is een lineaire combinatie van \bar{Y} en $\hat{\beta}$, die beide lineaire combinaties zijn van de Y_i 's. Daarom is $\hat{\alpha}$ ook een lineaire combinatie van de Y_i 's.
Om dit argument meer expliciet te maken, kun je, gebruik makend van $\hat{\beta} = \sum w_i Y_i$, schrijven:
$$\hat{\alpha} = \frac{1}{n} \sum Y_i - \bar{X} \sum w_i Y_i = \sum \frac{1}{n} Y_i - \sum \bar{X} w_i Y_i = \sum (\frac{1}{n} - \bar{X} w_i) Y_i.$$
87. $E(\hat{\alpha}) = E(\bar{Y} - \hat{\beta}\bar{X}) = E(\bar{Y}) - \bar{X}E(\hat{\beta})$.
Uit $E(\bar{Y}) = E(\frac{1}{n} \sum Y_i) = \sum \frac{1}{n} E(Y_i) = \sum \frac{1}{n} (\alpha + \beta X_i) = \alpha + \beta \bar{X}$
en $E(\hat{\beta}) = \beta$ volgt $E(\hat{\alpha}) = \alpha + \beta \bar{X} - \beta \bar{X} = \alpha$.
88. In dit geval geldt: $\hat{\alpha} = \bar{Y}$. Uit de onderlinge onafhankelijkheid van de Y_i 's volgt:
$$Var(\bar{Y}) = Var(\frac{1}{n} \sum Y_i) = \frac{1}{n^2} \sum Var(Y_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$
89. $E(\hat{\alpha}) = \alpha$ en $Var(\hat{\alpha}) = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2})$,
dus $T = \frac{\hat{\alpha} - \alpha}{s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}}}$ heeft een t -verdeling met $n-2$ vrijheidsgraden.

Het betrouwbaarheidsinterval wordt dus gegeven door:

$$\alpha = \hat{\alpha} \pm t_{[n-2], 0.05} \cdot s \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}}.$$

90. a. Voer de Y's in bij L1 en de S'en bij L2. **LinRegTTest** met **Xlist:L1**, **Ylist:L2**, **Freq:1**, $\beta > 0$ geeft een p-waarde van 0,034. We verwerpen de hypothese, want $0,034 < 0,05$.
- b. Bij tweezijdig toetsen verdubbelt de p-waarde. $2 \cdot 0,034 > 0,05$, dus dan kunnen we niet verwerpen.
91. a. Voer de x'en in bij L1 en de y's bij L2. **LinRegInt** met **Xlist:L1**, **Ylist:L2**, **Freq:1**, **C-level 0.95**, **RegEQ:Y1** en **Calculate** geeft als regressievergelijking:
 $Y = 4,50 - 5,65X$ en als betrouwbaarheidsinterval voor β
 $[-6,61; -4,69]$.

- b. Voor deze toets gebruiken we de toetsingsgrootheid

$$T = \frac{\hat{\beta} - \beta}{\sqrt{s^2 / \sum x_i^2}} \text{ met (onder } H_0) \beta = 4.$$

We weten (uit de GR-uitvoer van onderdeel a. dat $\hat{\beta} \approx -5,65$.

De waarde van $\sqrt{s^2 / \sum x_i^2}$ kunnen we eveneens afleiden uit de resultaten van a.

De bovengrens van het bij a. berekende betrouwbaarheidsinterval volgt uit de formule:

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

Er geldt $t_{[n-2], 0.05} \approx \text{invT}(0.975, 13) = 2,160$.

Dus: $-4,69 = -5,65 + 2,16 \cdot \frac{s}{\sqrt{\sum x_i^2}}$. Hieruit volgt

$$\frac{s}{\sqrt{\sum x_i^2}} \approx 0,442.$$

Voor de gevraagde toets is de toetsingsgrootheid

$$T = \frac{\hat{\beta} - \beta}{\sqrt{s^2 / \sum x_i^2}} \approx \frac{-5,65 - -4}{0,442} \approx -3,73.$$

Hierbij hoort $p \approx \text{tcdf}(-10^99, -3.72, 13) = 0,0013$.

- c. Als je de residuen plot, zie je dat de eerste twee en de laatste 4 positief zijn. Daartussenin zijn de residuen negatief. Dit duidt erop dat het werkelijke verband tussen X en Y niet lineair is. De gevonden regressievergelijking geeft wel een goede benadering van de puntenwolk, maar toetsen en betrouwbaarheidsintervallen hebben geen praktische betekenis.

92. Als $X_0 = 0$, geldt $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta} \cdot 0 = \hat{\alpha}$.
 Met $x_0 = X_0 - \bar{X} = -\bar{X}$, volgt nu

$$Var(\hat{\alpha}) = Var(\hat{\alpha} + \hat{\beta}X_0) = \sigma^2\left(\frac{1}{n} + \frac{x_0^2}{\sum x_i^2}\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right)$$
93. a. $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta}X_0$, dus $\hat{\mu}_0$ is een lineaire combinatie van $\hat{\alpha}$ en $\hat{\beta}$, die op hun beurt weer lineaire combinaties zijn van de Y_i 's. Hieruit volgt het gestelde.
- b. Er geldt $Z = (\hat{\mu}_0 - \mu_0)/\sqrt{Var(\hat{\mu}_0)}$ en $Var(\hat{\mu}_0) = \sigma^2\left(\frac{1}{n} + \frac{x_0^2}{\sum x_i^2}\right)$.
 Vervang σ^2 door s^2 en je krijgt dat $\frac{\hat{\mu}_0 - \mu_0}{s\sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2}}}$ een t-verdeling heeft met $n-2$ vrijheidsgraden. Het 95%-betrouwbaarheidsinterval wordt dus gegeven door:

$$\mu_0 = \hat{\mu}_0 \pm t_{[n-2],0.05} \cdot s\sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2}}$$
94. a. Na invoer van de gegevens in L1 en L2, ga je naar **LinRegTInt** en voer je **in:Xlist:L1 Ylist:L2 en RegEq:Y1**
Y1(24000) geeft $Y_0 = 2223$ als $X_0 = 24000$. De uitvoer geeft $s = 658,4$
2-VarStats geeft $\bar{X} = 24400$ en $\sum X_i^2 = 3146 \cdot 10^6$. Dus $x_0 = 24000 - 24400 = -400$ en $\sum x_i^2 = \sum X_i^2 - n\bar{X}^2 = 3146 \cdot 10^6 - 5 \cdot 24400^2 = 1692 \cdot 10^5$.
 $t_{[3],0.1} = \text{invT}(0.95, 3) = 2,353$
 Dus $\mu_0 = 2223 \pm 2,353 \cdot 658,5\sqrt{\frac{1}{5} + \frac{(-400)^2}{1692 \cdot 10^5}}$ geeft $[1528 ; 2918]$.
 Voor $X_0 = 30000$ voeren we dezelfde berekening uit. De verschillen zijn:
 $Y_0 = 3072$, $x_0 = 30000 - 24400 = 5600$.
 Dus $Y_0 = 3072 \pm 2,353 \cdot 658,5\sqrt{\frac{1}{5} + \frac{5600^2}{1692 \cdot 10^5}}$ geeft $[2110 ; 4034]$.
- b. Vanwege de grotere waarde van x_0 is de lengte van het tweede interval groter. Dat zit hem in de tweede term van de uitdrukking onder het wortelteken. Hoe verder X_0 verwijderd is van \bar{X} , hoe meer de onzekerheid over β meespeelt in de marge rond $\hat{\mu}_0$.
95. a. $\hat{\mu}_0$ is een functie van Y_i ($i = 1, \dots, n$) en is dus onafhankelijk van Y_0 .
 Dus $Var(Y_0 - \hat{\mu}_0) = Var(Y_0) + Var(\hat{\mu}_0) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{x_0^2}{\sum x_i^2}\right) =$

$$\sigma^2\left(\frac{1}{n} + \frac{x_0^2}{\sum x_i^2} + 1\right)$$

b. $Z = \frac{Y_0 - \hat{\mu}_0}{\sigma \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2} + 1}}$ is dus standaardnormaal verdeeld.

Vervang σ^2 door s^2 en je krijgt dat $\frac{Y_0 - \hat{\mu}_0}{s \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2} + 1}}$ een t-verdeling

heeft met $n-2$ vrijheidsgraden.

Het 95%-betrouwbaarheidsinterval wordt dus gegeven door:

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

96. Verwijzend naar de uitwerking van opgave 94 krijgen we nu:

$$Y_0 = 3072 \pm 2,353 \cdot 658,5 \sqrt{\frac{1}{5} + \frac{5600^2}{1692 \cdot 10^5} + 1} \text{ geeft } [1248 ; 4896].$$

97. a. Voer in $X \rightarrow L1$ en $Y \rightarrow L2$. **LinReg Y1** geeft $Y = 55,9 + 0,312X$. De plot van de lijn door de punten laat het volgende beeld zien:

(i) de punten lijken redelijk op een rechte lijn te liggen en er lijkt geen patroon te zijn van positieve en negatieve residuen;

(ii) er lijkt geen patroon te zijn in de grootte van de residuen. Dus de veronderstelling van een lineair model en van gelijke variantie voor de storingstermen lijkt gerechtvaardigd.

b. **LinRegTInt** met **C-level** 0.95 geeft $[0,25 ; 0,37]$.

c. **Y1(800)** geeft $Y_0 = 305,4$ als $X_0 = 800$. De uitvoer van b. geeft $s=11,176$.

2-VarStats geeft $\bar{X} = 519,2$ en $\sum X_i^2 = 3134543$. Dus $x_0 = 800 - 519,2 = 280,8$ en $\sum x_i^2 = \sum X_i^2 - n\bar{X}^2 = 3134543 - 11 \cdot 519,2^2 = 169288$.

$$t_{[9],0.05} = \text{invT}(0.975, 9) = 2,262$$

$$\text{Dus } Y_0 = 305,4 \pm 2,262 \cdot 11,176 \sqrt{\frac{1}{11} + \frac{280,8^2}{169288} + 1} \text{ geeft}$$

$[274; 337]$, d.w.z. tussen de 274.000 en de 337.000 verkeersongelukken.

d. We weten dat $\frac{Y_0 - \hat{\mu}_0}{s \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2} + 1}}$ een t -verdeling heeft met 9 vrij-

heidsgraden. Dus:

$$\Pr(Y_0 > 325) = \Pr\left(\frac{Y_0 - \hat{\mu}_0}{s \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2} + 1}} > \frac{325 - 305,4}{11,176 \sqrt{\frac{1}{11} + \frac{280,8^2}{169288} + 1}}\right)$$

$$= 1,406)$$

Deze kans is gelijk aan $\text{tcdf}(1.406, 10^{99}, 9) = 0,10$.

e. $Y_1(1500)$ geeft 523. Het model voorspelt dus 523.000 verkeersongelukken. De waarnemingen die ten grondslag liggen aan het model betreffen een aantal auto's dat toeneemt van 3,5 tot 7,4 miljoen. Het is maar de vraag of het model ook geldig is voor 15 miljoen auto's. Tegen de tijd dat het wagenpark zover is gegroeid, gelden wellicht andere wetmatigheden. Extrapoleren is altijd gevaarlijk.

98. De formule is $X_0 = \bar{X} \pm t_{[n-1],0,05} s \sqrt{\frac{1}{n} + 1}$.
Deze formule geeft een 95% betrouwbaarheidsinterval voor een nieuwe waarneming uit een normaal verdeelde populatie waaruit reeds n waarnemingen bekend zijn.

7 Hoofdstuk 7

99. a. positief
 b. positief
 c. negatief
 d. positief
 e. negatief
 f. positief
 g. positief
 h. niet
 i. positief
 j. positief
 k. negatief
 l. negatief (als het zomer is in Amsterdam, is het winter in Sydney en omgekeerd)
 m. positief
 n. niet
100. a. goed
 b. fout
 c. goed
 d. goed
 e. fout

101. Uitschrijven geeft:

$$\begin{aligned}\sum x_i y_i &= \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - \bar{X} \sum Y_i - \bar{Y} \sum X_i + n \bar{X} \cdot \bar{Y} = \\ &= \sum X_i Y_i - n \bar{X} \cdot \bar{Y} - n \bar{X} \cdot \bar{Y} + n \bar{X} \cdot \bar{Y} = \sum X_i Y_i - n \bar{X} \cdot \bar{Y}\end{aligned}$$

102. a. Na invoer X in L1 en Y in L2 geeft **2-VarStats L1, L2**:
 $\sum X_i Y_i = 454,39$ $\bar{X} = 7,2222$ $\bar{Y} = 6,8667$ en $n = 9$
 Dus $s_{XY} = \frac{1}{8}(454,39 - 9 \cdot 7,2222 \cdot 6,8667) \approx 1,007$
- b. **2-VarStats** geeft eveneens $s_X = 1,4754$ $s_Y = 0,93675$.
 Dus $\frac{s_{XY}}{s_X \cdot s_Y} = \frac{1,007}{1,4754 \cdot 0,93675} \approx 0,73$
- c. Dezelfde berekening levert $s_{XY} = 100,7$
- d. De covariantie is afhankelijk van de gekozen eenheid.
- e. Aan de formule voor de covariantie kun je zien dat, als X en Y 10 keer zo groot worden, de covariantie 100 keer zo groot wordt.

- f. De nieuwe waarden zijn $s_X = 14,754$ en $s_Y = 9,3675$.
 $\frac{s_{XY}}{s_X \cdot s_Y}$ blijft ongewijzigd.

103. Vanwege $\bar{x} = 0$ geldt:

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

Op dezelfde manier geldt $s_y^2 = s_Y^2$. Het maakt dus geen verschil.

104. a. $\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} = \frac{\sum x_i (y_i + \bar{Y})}{\sum x_i^2} = \frac{\sum x_i y_i}{\sum x_i^2}$, want $\bar{Y} \sum x_i = 0$.
 $r \frac{s_Y}{s_X} = \frac{s_{XY}}{s_X s_Y} \cdot \frac{s_Y}{s_X} = \frac{(n-1)s_{XY}}{(n-1)s_X^2} = \frac{\sum x_i y_i}{\sum x_i^2}$.
 b. $\hat{\beta}_X \cdot \hat{\beta}_Y = \frac{\sum x_i y_i}{\sum y_i^2} \cdot \frac{\sum x_i y_i}{\sum x_i^2} = \frac{(n-1)^2 s_{XY}^2}{(n-1)^2 s_X^2 s_Y^2} = r^2$

105. Dit volgt rechtstreeks uit de formule van opgave 104a, omdat s_X en s_Y beide positief zijn.

106. $\begin{matrix} 0,67 & 0 & -1 \\ 0,33 & 0,33 & -0,67 \end{matrix}$

107. Omdat elk antwoord goed of fout is, geldt voor iedere i : $X_i + Y_i = 25$.
 Dus $\bar{Y} = \overline{(25 - X)} = 25 - \bar{X} = 25 - 18,6 = 6,4$ en $x_i = -y_i$.
 Hieruit volgt $s_Y^2 = \frac{1}{n-1} \sum y_i^2 = \frac{1}{n-1} \sum x_i^2 = s_X^2$. Dus $s_Y = 2,9$
 $r = -1$, want als je X kent, ken je Y . Je kunt dit ook algebraïsch inzien:
 $s_{XY} = \frac{1}{n-1} \sum x_i y_i = \frac{1}{n-1} \sum (-x_i^2) = -s_X^2 = -s_X s_Y$. Hieruit volgt $r = -1$.

108. a. Je moet elk cijferpaar net zo vaak invoeren als het voorkomt. Dus als je het cijfer voor wiskunde D invoert in L1 en het cijfer voor wiskunde B in L2 (het mag natuurlijk ook andersom), krijg je :
 L1= 5 6 6 6 6 6 6 6 ...
 L2= 7 6 7 7 8 8 8 8 ...
 LinReg (a+bx) L1, L2 (met Diagnostics On) geeft $r = 0,81$.
 b. Je voert als bij a. elk cijferpaar in in L1 en L2, maar je doet het maar één keer. De frequentie vul je in in L3, dus
 L1= 5 6 6 6 ...
 L2= 7 6 7 8 ...
 L3= 1 1 2 4 ...
 LinReg (a+bx) L1, L2 (met Diagnostics On) geeft $r = 0,81$.

- c. We toetsen $H_0 : \rho = 0$ tegen $H_1 : \rho \neq 0$. Dit is gelijkwaardig met $H_0 : \beta = 0$ tegen $H_1 : \beta \neq 0$. (Zie opgave 105.)
 Gebruik **LinRegTTest** en voer in: **Xlist:** L1 **Ylist:** L2
 (of andersom dat maakt niet uit) **Freq:** L3 $\beta \neq 0$.
Calculate geeft $t = 6,18$ met 20 vrijheidsgraden en $p = 4,9 \cdot 10^{-6}$. We verwerpen de nulhypothese. De correlatiecoëfficiënt verschilt significant van 0.

109. *Regressie van Y op X* : $\hat{\beta} = \hat{\beta}_Y = r \frac{s_Y}{s_X} = 0,5 \frac{3}{2} = 0,75$. De OLS-regressie lijn gaat door het punt (\bar{X}, \bar{Y}) , dus $8 = \hat{\alpha} + 0,75 \cdot 5$, dus $\hat{\alpha} = 4,25$. Dus $Y = 4,25 + 0,75X$.
Regressie van X op Y: Op een soortgelijke manier: $Y = 2\frac{1}{3} + \frac{1}{3}X$.

110. a. A. Vermoedelijk neemt de effectiviteit aanvankelijk toe als de handelaar ouder wordt (ervaring) en daarna af. (In de aandelenhandel moet je heel snel kunnen reageren.)
 b. B. Als de docent ook een formulier invult...
 c. E. Mannelijke studenten drinken doorgaans meer dan vrouwelijke studenten en ze zijn gemiddeld langer. Geslacht is een confounder.
 d. D. Je hebt te maken met de deelpopulatie van mensen met een medische klacht. Anders zaten ze niet in de wachtkamer van de huisarts.
 e. E. Oudere meisjes zullen vaker orale anticonceptiva gebruiken dan jongere meisjes. Oudere meisjes hebben doorgaans ook meer inkomen dan jongere. Leeftijd is een confounder.
 f. C. Het kan zijn dat in de desbetreffende wijk veel moslims wonen met gemiddeld een lager inkomen dan de niet-moslims. Moslims eten geen varkensvlees, maar wel lamsvlees. Je hebt te maken met verschillende deelpopulaties.

111. a. $\bar{\hat{Y}} = \frac{1}{n} \sum \hat{Y}_i = \frac{1}{n} \sum (\hat{\alpha} + \hat{\beta} X_i) = \frac{1}{n} \sum \hat{\alpha} + \hat{\beta} \cdot \frac{1}{n} \sum X_i = \hat{\alpha} + \hat{\beta} \bar{X} = \bar{Y}$
 b. Per definitie geldt $Y_i - \bar{Y} = y_i$ en $\hat{Y}_i - \bar{Y} = \hat{y}_i$. Dus $Y_i - \hat{Y}_i = (y_i + \bar{Y}) - (\hat{y}_i + \bar{Y}) = y_i - \hat{y}_i$. Het gevraagde volgt.
 c. Als we van de variabelen (X_i, Y_i) overstappen naar (x_i, y_i) , verplaatsen we de oorsprong van het oorspronkelijke assenstelsel naar (\bar{X}, \bar{Y}) . De richtingscoëfficiënt van de OLS-regressielijn verandert hierdoor niet, maar in het nieuwe assenstelsel gaat deze lijn door de oorsprong. Zijn formule is dus $y = \hat{\beta}x$. Er geldt dus $\hat{y}_i = \hat{\beta}x_i$.
 d. Er geldt:

$$\begin{aligned} \sum (y_i - \hat{y}_i)^2 + \sum \hat{y}_i^2 &= \sum (y_i - \hat{\beta}x_i)^2 + \sum \hat{\beta}^2 x_i^2 \\ &= \sum y_i^2 - 2\hat{\beta} \sum x_i y_i + \hat{\beta}^2 \sum x_i^2 + \hat{\beta}^2 \sum x_i^2. \end{aligned}$$

Uit $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ volgt $2\hat{\beta} \sum x_i y_i = 2\hat{\beta} \cdot \hat{\beta} \cdot \sum x_i^2 = 2\hat{\beta}^2 \sum x_i^2$.

Als deze gelijkheid wordt gesubstitueerd in de daaraan voorafgaande formule, volgt het gevraagde.

112.

$$\frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} = \frac{(\sum x_i y_i)^2 \sum x_i^2}{(\sum x_i^2)^2 \sum y_i^2} = \frac{(\sum x_i y_i)^2}{(\sum x_i^2)(\sum y_i^2)} = r^2.$$

8 Hoofdstuk 8

113. Uitschrijven van S geeft:

$$\begin{aligned} S &= \sum (Y_i^2 + a^2 + b^2 X_i^2 + c^2 Z_i^2 - 2aY_i - 2bX_iY_i - 2cZ_iY_i + 2abX_i + 2acZ_i + 2bcX_iZ_i) \\ &= \sum Y_i^2 + na^2 + b^2 \sum X_i^2 + c^2 \sum Z_i^2 - 2a \sum Y_i - 2b \sum X_iY_i - 2c \sum Z_iY_i + 2ab \sum X_i + 2ac \sum Z_i + 2bc \sum X_iZ_i \end{aligned}$$

$$\frac{dS}{da} = 2na - 2 \sum Y_i + 2b \sum X_i + 2c \sum Z_i = 0. \text{ Delen door 2 levert de eerste vergelijking. De andere vergelijkingen volgen analoog uit } \frac{dS}{db} = 0 \text{ en } \frac{dS}{dc} = 0.$$

114. Het aantal vrijheidsgraden is gelijk aan het aantal waarnemingen (31) verminderd met het aantal parameters dat moet worden geschat voordat we σ^2 kunnen schatten. In dit geval zijn het 4 parameters (β_0 t/m β_3). Het aantal vrijheidsgraden is dus $31 - 4 = 27$. De p-waarde voor de “Intercept” kan als volgt worden berekend met de GR:
 $2 * \text{tcdf} (2.226, 10^{99}, 27) = 0,0246$.
 De andere p-waarden volgen op dezelfde manier.

115. $t_{0,05[27]} \approx \text{invT}(0.075, 27) = 2,0518$

Het 95%-betrouwbaarheidsinterval wordt gegeven door:

$$\hat{\beta}_1 \pm t_{0,05[27]} \cdot s_{\beta_1} = 0,4917 \pm 2,0518 \cdot 0,2069. \text{ Dus } [0,067 ; 0,916]$$

116. Er geldt: $T = \frac{\hat{\beta}_3 - \beta_3}{s_{\beta_3}}$ heeft een t-verdeling met 27 vrijheidsgraden.

In dit voorbeeld vinden we $T = \frac{0,5357 - 0,2}{0,2289} \approx 1,349$. De overschrijdskans (p-waarde) van deze toets is dus $\Pr(T \geq 1,349) \approx \text{tcdf}(1.349, 10^{99}, 27) = 0,094$.

117. Definieer :

Y : brandstofefficiëntie

X_1 : gewicht

X_2 : maximumsnelheid

X_3 : Volkswagen ja/nee, d.w.z

$X_3 = 1$ als de auto een Volkswagen is en

$X_3 = 0$ als de auto geen Volkswagen is.

X_4 : Fiat ja/nee, d.w.z.

$X_4 = 1$ als de auto een Fiat is en

$X_4 = 0$ als de auto geen Fiat is.

Het model wordt nu $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i$. Hierin stelt de parameter β_3 de extra brandstofefficiëntie van een Volkswagen voor in vergelijking met de brandstofefficiëntie van een Toyota. β_4 is de extra brandstofefficiëntie van een Fiat in vergelijking met een Toyota.

Je ziet dat je, om een categorische variabele met drie niveaus in je model op te nemen (hier: Toyota, Volkswagen of Fiat), twee dummyvariabelen nodig hebt. We gebruikten hier Toyota als “referentiemerk”. Die keuze is natuurlijk willekeurig.

9 Hoofdstuk 9

1. a. Noem de lengte van een vrouw V en die van een man M .
 Voor vrouwen: $\Pr(V > 185) = \text{normalcdf}(185, 10^99, 171, 6) = 0,01$
 Voor mannen: $\Pr(M > 185) = \text{normalcdf}(185, 10^99, 183, 6) = 0,37$
 - b. Voor vrouwen: $Z = \frac{V - \mu}{\sigma} = \frac{V - 171}{6}$ is standaardnormaal verdeeld.
 $\Pr(V > 185) = \Pr(Z > \frac{185 - 171}{6}) = \Pr(Z > 2,333) \approx \text{normalcdf}(2,333, 10^99) = 0,01$.
 Voor mannen: $Z = \frac{M - \mu}{\sigma} = \frac{M - 183}{6}$ is standaardnormaal verdeeld.
 $\Pr(M > 185) = \Pr(Z > \frac{185 - 183}{6}) = \Pr(Z > 0,3333) \approx \text{normalcdf}(0,3333, 10^99) = 0,37$.
 - c. $\Pr(M > V) = \Pr(M - V > 0)$. $M - V$ is normaal verdeeld met gemiddelde $183 - 171 = 12$ en variantie $6^2 + 6^2 = 72$. Dus $\Pr(M - V > 0) \approx \text{normalcdf}(0, 10^99, 12, \sqrt{72}) = 0,92$.
2. a. Het gemiddelde is 0,48. De mediaan is $\frac{0,46 + 0,52}{2} = 0,49$. De range is $0,55 - 0,39 = 0,16$. De variantie is $s^2 = \frac{(0,39 - 0,48)^2 + (0,43 - 0,48)^2 + \dots + (0,55 - 0,48)^2}{6 - 1} \approx 0,004$.
 De standaarddeviatie is $s \approx \sqrt{0,004} \approx 0,063$.
 Eenvoudiger is het om de gegevens in te voeren in L1. **1-Var Stats L1** geeft alle gevraagde waarden.
 - b. De mediaan is de middelste van de observaties en minder gevoelig voor extreme uitkomsten dan het gemiddelde. Als de verdeling symmetrisch is, zijn het gemiddelde en de mediaan ongeveer gelijk. Als de mediaan kleiner dan het gemiddelde is, is de verdeling scheef naar rechts en als de mediaan groter dan het gemiddelde is de verdeling scheef naar links. De normale verdeling is symmetrisch. Een groot verschil tussen gemiddelde en mediaan van een steekproef duidt er op dat de verdeling van de waarnemingen niet normaal is.
3. a. Stel X is de hematocrietwaarde van een willekeurige mannelijke sporter die geen EPO gebruikt. We veronderstellen dat X normaal verdeeld is met $\mu = 0,44$ en $\sigma = 0,03$.
 $\Pr(X > 0,50) \approx \text{normalcdf}(0,5, 10^99, 0,44, 0,03) = 0,023$.
 - b. Vanwege $\text{invNorm}(0,975) = 1,96$, geldt dat ongeveer 95% van de hematocrietwaarden ligt binnen het interval:

$$(\mu - 1,96\sigma, \mu + 1,96\sigma) = (0,44 - 1,96 \times 0,03 ; 0,44 + 1,96 \times 0,03) = (0,381 ; 0,499).$$

(Let op: een referentie-interval is niet hetzelfde als een betrouwbaarheidsinterval voor μ .)

Deze berekening van een referentie-interval is niet helemaal correct omdat we geen rekening houden met het feit dat 0,44 en 0,03 *schattingen* zijn van μ respectievelijk σ . De theoretisch juiste berekening (met de t-verdeling) zou echter vrijwel dezelfde resultaten opleveren.

- c. Als de data van deze Utrechtse sporters representatief is voor alle mannelijke topsporters die geen EPO gebruiken, zou 2,3% van de niet-gebruikers toch positief testen.
4. a. De standaarddeviatie van de steekproef is $s = 0,03$. Er waren 18 mannelijke atleten. De standaardfout van het gemiddelde is $SEM = s/\sqrt{n} = 0,03/\sqrt{18} \approx 0,0071$. Het 95%-betrouwbaarheidsinterval voor het onbekende populatiegemiddelde μ is

$$\bar{X} \pm t \cdot SEM.$$

We hebben te maken met een t-verdeling met $18 - 1 = 17$ vrijheidsgraden. Voor een 95%-betrouwbaarheidsinterval gebruiken we $t = \text{invT} (0.975, 17) = 2,11$

Het 95%-betrouwbaarheidsinterval wordt: $0,44 \pm 2,11 \cdot 0,0071$, dus $[0,425 ; 0,455]$.

Met de GR: gebruik **Tinterval** met **Inpt: Data; \bar{x} : 0.44**
Sx: 0.03 n: 18 C-level: 0.95.

Calculate geeft het gevraagde interval.

- b. Dit interval geeft aan hoe nauwkeurig het gemiddelde geschat is. Bij 95 van de 100 studies ligt het onbekende populatiegemiddelde in het 95%-betrouwbaarheidsinterval. In opgave 3 is het referentieinterval berekend, dat aangeeft waarbinnen 95% van de individuele waarden vallen.
- c. De standaarddeviatie $SD = s$ geeft de spreiding aan van de individuele waarnemingen. De standaardfout van het gemiddelde $SEM = s/\sqrt{n}$ geeft aan hoe nauwkeurig het gemiddelde geschat is.
- d. De standaarddeviatie blijft bij stijgende steekproefgrootte ongeveer gelijk. Als er meer waarnemingen zijn kunnen we het gemiddelde nauwkeuriger schatten en wordt de SEM dus kleiner. Bij 2 keer zo veel waarnemingen wordt SEM $\sqrt{2}$ keer kleiner.

5. We hebben te maken met gepaarde gegevens.

We toetsen H_0 : “De kansen op een hematocrietwaarde van 0,50 of meer zijn voor en na het gebruik van EPO gelijk.” tegen H_1 : “De kans op een hematocrietwaarde van 0,50 of meer is na gebruik van EPO groter dan ervoor.”

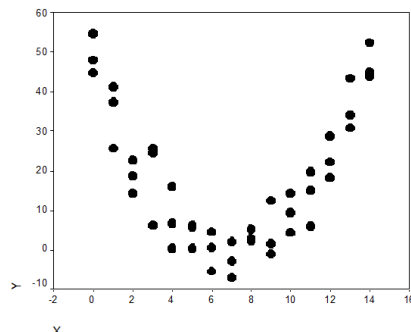
We kijken naar de 7 waarnemingen waarbij de hematocrietwaarde vóór het gebruik van EPO aan de andere kant van 0,50 lag dan erna. Van deze 7 waarnemingen is er een aantal waarbij de hematocrietwaarde voor het gebruik van EPO $\geq 0,50$ en na het gebruik van EPO $< 0,50$ was. Dit laatste aantal noemen we X . X is binomiaal verdeeld met parameters $n' = 7$ en π . De nulhypothese en het alternatief kunnen nu als volgt worden geformuleerd:

$H_0 : \pi = \frac{1}{2}$ en $H_1 : \pi < \frac{1}{2}$.

De steekproef levert $X = 0$, dus de p-waarde van deze toets is $\Pr(X \leq 0) = (\frac{1}{2})^7 \approx 0,008$ (of via `binomcdf (7, 0.5, 0)`). We kunnen de nulhypothese verwerpen. De kans op een hematocrietwaarde van 0,5 of meer is na gebruik van EPO significant hoger dan ervoor.

6. a. De correlatie is negatief (want als x groter wordt, wordt y kleiner).
De punten liggen redelijk op een rechte lijn, de correlatie ligt dus in de buurt van de -1.

b. Bijvoorbeeld:



- c. We voeren een lineaire regressie uit van de hartslag op de hematocrietwaarde.

In lijst L1 voeren we de hematocrietgegevens in en in lijst L2 de hartslaggegevens.

Daarna gebruiken we de functie `LinRegTTest` met invoer: `Xlist:L1`
`Ylist:L2` `Freq:1` `$\beta \& \rho \neq 0$` `RegEQ: Y1`.

`Calculate` geeft:

- De geschatte regressievergelijking: $\text{harts} = 151 - 53,9 \cdot \text{he}$
- De correlatiecoëfficiënt $r = -0,16$

- De p-waarde 0,49 van de toets $H_0 : \beta = 0$ tegen het tweezijdig alternatief. Deze toets is gelijkwaardig met de toets $H_0 : \rho = 0$ tegen het tweezijdig alternatief.
- Plots van de gegevens met de geschatte regressielijn en van de residuen brengen geen opmerkelijke patronen aan het licht.

We concluderen dat de geschatte correlatie tussen hematocriet-waarde en hartslag licht negatief is, maar niet significant afwijkt van 0. We hebben dus geen verband tussen deze twee grootheden kunnen aantonen.

- De punten in de grafiek schuiven iets op naar rechts.
 - Door een horizontale (of verticale) verschuiving verandert er niets in de samenhang. Dus $r = -0,16$.
7. We zijn geïnteresseerd in de invloed van het gebruik van calcium in vergelijking met de placebo, dus we berekenen eerst de afname van de bloeddruk per individu. We vinden:

Afname bloeddruk calcium	7	-4	18	17	-3	-5	1	10	11	-2	
Afname bloeddruk placebo	-1	12	-1	-3	3	-5	5	2	-11	-1	-3

We hebben te maken met ongepaarde waarnemingen uit twee onafhankelijke steekproeven, waarvan we de gemiddelden willen vergelijken. Als we kunnen aannemen dat de waarnemingen normaal verdeeld zijn, kunnen we de twee-steekproeven-t-toets gebruiken, anders gebruiken we de toets van Wilcoxon.

We voeren de verschillen calcium in in **L1** en de verschillen placebo in **L2**. We bekijken achtereenvolgens de resultaten van **1-Var Stats L1** en **1-Var Stats L2**. We constateren dat de beide steekproeven redelijk symmetrisch zijn (geen groot verschil tussen gemiddelde en mediaan) en dat zich ook geen vreemde uitschieters voordoen. Bovendien lijkt het a priori niet vreemd dat een variabele als “afname bloeddruk” normaal verdeeld is. Verder verschillen de gevonden varianties niet veel van elkaar. We besluiten daarom de genoemde t-toets uit te voeren.

We toetsen $H_0 : \mu_{\text{calcium}} = \mu_{\text{placebo}}$ tegen $H_1 : \mu_{\text{calcium}} \neq \mu_{\text{placebo}}$. (Als je een aanwijzing hebt dat het gebruik van calcium de bloeddruk zal verlagen, is de keuze van een eenzijdig alternatief gerechtvaardigd; hier hebben we “conservatief” voor een tweezijdig alternatief gekozen.)

We gebruiken **2 - SampTTest** met als invoer **Inpt:Data List1:L1 List2:L2 Freq1:1 Freq2:1 $\mu_1 \neq \mu_2$ Pooled:Yes**.

Calculate geeft de resultaten. De calciumgebruikers toonden een gemiddelde afname van de bloeddruk met 5, bij de placebo gebruikers was dit -0,3 (een kleine toename dus). De p-waarde van de toets is 0,12 (t van 1,63 bij 19 vrijheidsgraden), dus de gevonden verschillen

zijn niet significant.

8.
 - a. Het gemiddelde is 12,75 minuten, de mediaan 8,5 minuten, het minimum 3 en het maximum 45.
 - b. De t-toets veronderstelt normaliteit van de uitkomst. Dat lijkt hier niet het geval, het verschil tussen gemiddelde en mediaan is relatief groot en er is een extreme uitschieter (45 minuten). Bovendien is het aantal waarnemingen klein. In dit geval kun je de tekentoets gebruiken of Wilcoxon's rangtekentoets.
 - c. Beide toetsen gaan over de mediaan m (niet over het gemiddelde). We toetsen $H_0 : m = 15$ tegen $H_0 : m \neq 15$.
Bij de *tekentoets* gebruiken we dat onder de nulhypothese de kans op een waarneming kleiner dan 15 gelijk is aan 0,5. Het aantal waarnemingen X dat kleiner is dan 15 is dus binomiaal verdeeld met $n = 8$ en $\pi = 0,5$. 7 waarnemingen zijn kleiner dan 15. De p-waarde van de (tweezijdige) toets is dus:
 $2 \cdot \Pr(X \geq 7) = 2 \cdot (1 - \Pr(X \leq 6)) \approx 2 \cdot (1 - \text{binomcdf}(8, 0.5, 6)) = 0,07 > \alpha$.
We kunnen de nulhypothese niet verwerpen.
Bij *Wilcoxon's rangtekentoets* noteren we de "waarnemingen minus 15" en krijgen:
-5, -8, -6, 30, -7, -1, -12, -9.
We voorzien de absolute waarde van deze getallen van een rangnummer en voegen er het oorspronkelijke teken aan toe. We krijgen:
-2, -5, -3, 8, -4, -1, -7, -6. Opgeteld: $S = -20$. In de tabel lezen we bij $n' = 8$ en $\alpha = 0,05$ dat de grenswaarden -30 en 30 zijn. -20 ligt daartussen. We kunnen de nulhypothese niet verwerpen. Overigens zien we dat we ook bij $\alpha = 0,1$ de nulhypothese niet kunnen verwerpen. De p-waarde van deze toets is dus groter dan 0,1.
9.
 - a. We gebruiken de gepaarde t-toets.
 H_0 :de gemiddelde bloeddruk voor is gelijk aan de gemiddelde bloeddruk na de inspanning.
 H_1 :deze gemiddelden zijn ongelijk.
We veronderstellen dat de verschillen (bloeddruk na - voor) normaal verdeeld zijn. De waargenomen verschillen zijn:

persoonnummer	bloeddruk voor	bloeddruk na	verschil (na-voor)
1	80	83	3
2	72	77	5
3	90	95	5
4	75	72	-3
5	71	80	9
6	105	95	-10
7	95	100	5
8	80	82	2

Voer deze verschillen in in L1 en gebruik T-Test. Voer in Inpt:

Data $\mu_0 : 0$ List:L1 Freq:1 $\mu \neq \mu_0$.

Calculate geeft $t = 0,954$ bij 7 vrijheidsgraden en een bijbehorende p-waarde van 0,37. We kunnen de nulhypothese niet verwerpen. Er is geen significant verschil tussen bloeddruk voor en bloeddruk na de fietsproef.

- b. Voor het 95%-betrouwbaarheidsinterval gebruiken we TInterval met dezelfde invoer als bij onderdeel a en C-level: 0.95.

Calculate geeft het interval $[-3,0 ; 7,0]$.

Omdat 0 in dit interval zit, kunnen we de hypothese dat de bloeddruk voor de fietsproef gelijk is aan erna niet verwerpen bij $\alpha = 0,05$.

- c. De meest geschikte toets is Wilcoxon's rangtekentoets.

H_0 : bloeddruk voor en na de fietsproef hebben dezelfde verdeling

H_1 : bloeddruk voor en na de fietsproef hebben niet dezelfde verdeling.

We voorzien de absolute waarde van deze getallen van een rangnummer en voegen er het oorspronkelijke teken aan toe. We krijgen:

2,5 5 5 -2,5 7 -8 5 1. Opgeteld: $S = 15$. In de tabel lezen we bij $n' = 8$ en $\alpha = 0,05$ dat de grenswaarden -30 en 30 zijn. 15 ligt daartussen. We kunnen de nulhypothese niet verwerpen. Overigens zien we dat we ook bij $\alpha = 0,1$ de nulhypothese niet kunnen verwerpen. De p-waarde van deze toets is dus groter dan 0,1.

10. a. Onder Leidse scholieren is de proportie alcoholgebruikers gelijk aan $p=675/1161=0,58$.

Aangezien zowel $np \geq 5$ als $n(1-p) \geq 5$ mogen we de benadering via de normale verdeling gebruiken. Het 95%-betrouwbaarheidsinterval voor een de werkelijke proportie π is:

$$\pi = p \pm z \cdot \sqrt{\frac{p(1-p)}{n}} \text{ met } z = \text{invNorm}(0.975) = 1,96, \\ \text{dus } (0,55 ; 0,61).$$

Met de GR gebruik je 1-PropZInt met x: 675 n:1161

C-level: 0.95.

Calculate geeft het gevraagde interval.

- b. We weten met 95% zekerheid dat de proportie Leidse scholieren die alcohol gebruiken tussen de 0,55 en 0,61 ligt. Van de gehele Nederlandse scholierenbevolking is de proportie die alcohol gebruikt 0,543. Dit getal zit niet in het 95%-betrouwbaarheidsinterval van de Leidse scholieren. Conclusie: De gemiddelde Leidse scholier gebruikt significant vaker alcohol dan de gemiddelde Nederlandse scholier.
11. a. Ouders roken: gemiddelde 20,4, mediaan 14, maximum 46 minuten
Ouders roken niet : gemiddelde 25,3, mediaan 15,5, maximum 65 minuten.
De verdelingen zijn scheef. We hebben niet te maken met een normale verdeling.
- b. We hebben te maken met ongepaarde, niet normaal-verdeelde gegevens uit twee onafhankelijke steekproeven. We gebruiken daarom de toets van Wilcoxon.
We toetsen H_0 : De gegevens uit beide steekproeven hebben dezelfde verdeling
tegen H_1 : Beide verdelingen verschillen.
De rangnummers zijn:
Ouders roken: 3,5 1 8 5,5 10
Ouders roken niet: 2 3,5 7 5,5 9 11
De som van de rangnummers van de kleinste groep is $S = 28$.
In de tabel kijken we bij $n_1 = 5$, $n_2 = 6$ en $\alpha = 0,05$. We constateren dat S tussen de beide grenzen (18 en 42) ligt. We kunnen de nulhypothese niet verwerpen. Het verschil tussen beide groepen is niet significant. Overigens ligt S ook tussen de grenzen die horen bij $\alpha = 0,1$ (20 en 40). De p-waarde van deze toets is dus groter dan 0,1.
12. a. We vergelijken het gemiddelde van één populatie (energie-inname van kinderen met Down syndroom) met een vooraf vaststaande waarde (RDA), dus de (één-steekproef) t-toets lijkt het meest geschikt. Voor deze toets moeten we aannemen dat de energie-inname normaal verdeeld is. Dit lijkt een redelijke veronderstelling, maar met de beschikbare gegevens kunnen we deze niet controleren. Zelfs als de verdeling niet normaal zou zijn, is het aantal waarnemingen groot genoeg om de t-toets acceptabel te doen zijn. Noem de gemiddelde intake in de populatie kinderen met Down syndroom μ .

Dan toetsen we $H_0 : \mu = 1326$ tegen $H_1 : \mu \neq 1326$.

- b. Gebruik op de GR de functie **T-Test**. Voer in: **Inpt: Stats**
 $\mu_0 : 1326$ $\bar{x} : 976$ **Sx:** 210 **n:**33 $\mu \neq \mu_0$.
Calculate geeft $t = -9,57$ met 32 vrijheidsgraden en een p-waarde van $6,5 \cdot 10^{-11}$. Dus we verwerpen de nulhypothese bij elk redelijk significantieniveau.

We concluderen dat de kinderen met het Downsyndroom een significant lagere energie-inname hebben dan de aanbevolen hoeveelheid.

- c. We gebruiken **TInterval** met grotendeels dezelfde invoer als bij onderdeel b.

C-level: 0.95 en **Calculate** geven: (901, 1051 kcal).

De aanbevolen hoeveelheid RDA=1326 kcal zit niet in dit interval. Dus net als bij onderdeel b concluderen we dat de kinderen met Downsyndroom significant minder energie binnen krijgen dan aanbevolen wordt.

13. a. We hebben een numerieke, continue uitkomst die bij benadering normaal verdeeld is. Twee ongerelateerde groepen, dus we gebruiken de niet-gepaarde t-toets.

- b. Noem de gemiddelde energie-inname per groep μ_1 (Down) respectievelijk μ_2 (controle).

We toetsen $H_0 : \mu_1 = \mu_2$ tegen $H_1 : \mu_1 \neq \mu_2$.

We gebruiken **2-SampTTest**. Voer in: **Inpt:Stats** $\bar{x}_1 : 976$
Sx1: 210 **n1:**33 $\bar{x}_2 : 1214$ **Sx2:**322 **n2:**26 $\mu_1 \neq \mu_2$ **Pooled:** Yes.

Calculate geeft $t = -3,42$ met 57 vrijheidsgraden en een p-waarde van 0,0011. We verwerpen de nulhypothese. Er is een significant verschil in gemiddelde energie-inname tussen beide groepen.

- c. *Zonder GR:*

Het 95% betrouwbaarheidsinterval voor het echte verschil $\mu_1 - \mu_2$ is $\bar{X}_1 - \bar{X}_2 \pm t \cdot SEM$.

Hierin is $SEM = s \sqrt{n_1^{-1} + n_2^{-1}}$ met

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{32 \cdot 210^2 + 25 \cdot 322^2}{57} \approx 70233.$$

Dus $SEM \approx \sqrt{70233(33^{-1} + 26^{-1})} \approx 69,5$.

We hebben te maken met 57 vrijheidsgraden, dus $t = \text{invT}(0.975, 57) = 2,00$.

Via $\bar{X}_1 - \bar{X}_2 = 976 - 1214 = -238$ vinden we het interval $[-377, -99]$.

Met GR :

We gebruiken de functie **2 - SampTInt** met dezelfde invoer als bij onderdeel b. en **C-level:**0.95. We krijgen hetzelfde interval

als hierboven.

Omdat 0 niet in dit interval zit, concluderen we, net als bij onderdeel b, dat de gemiddelde energie-inname van kinderen met het Downsyndroom significant verschilt van die van de kinderen uit de controlegroep.

- d. We gebruiken opnieuw 2 - `SampTInt`, ditmaal met invoer:
`Inpt:Stats` $\bar{x}_1 : 94$ `Sx1:` 22 `n1:`33 $\bar{x}_2 : 89$ `Sx2:`28
`n2:`26 `C-level:`0.95 `Pooled:` Yes.
`Calculate` geeft het interval $[-8, 18]$. Omdat 0 in dit interval zit concluderen we dat de gemiddelde energie-inname per kg lichaamsgewicht van kinderen met het Downsyndroom niet significant verschilt van die van de kinderen uit de controlegroep. Kinderen met het Downsyndroom zijn gemiddeld lichter dan de kinderen uit de controlegroep. Als je hun energie-inname corrigeert voor dit lagere gewicht, blijkt deze niet significant af te wijken van de energie-inname van de kinderen uit de controlegroep.
14. a. We vergelijken een populatiepercentage met een vooraf gegeven waarde (54%). We kunnen de binomiaaltoets toepassen. (Omdat zowel $19 \geq 5$ als $42 - 19 \geq 5$, zouden we de binomiale verdeling mogen benaderen door de normale. Nodig is dit niet.)
- b. Noem de proportie in de populatie kinderen met Downsyndroom dat na 8 dagen borstvoeding krijgt π .
We toetsen $H_0 : \pi = 0,54$ tegen $H_1 : \pi \neq 0,54$.
Noem het aantal kinderen uit de steekproef dat na 8 dagen nog borstvoeding kreeg X . Onder de nulhypothese is X binomiaal verdeeld met $n = 42$ en $\pi = 0,54$. De p-waarde van de (tweezijdige) toets is:
 $2 \cdot \Pr(X \leq 19) \approx 2 \cdot \text{binomcdf}(42, 0.54, 19) = 0,32 > \alpha$.
We kunnen de nulhypothese niet verwerpen. De geobserveerde proportie borstvoeding onder kinderen met het syndroom van Down wijkt niet significant af van die van de Nederlandse bevolking.
- c. Voor het betrouwbaarheidsinterval gebruiken we de benadering via de normale verdeling. Gebruik `1-PropZInt`. Voer in: `x:`19 `n:`42 `C-level:`0.95. `Calculate` geeft $[0,30; 0,60]$.
0,54 ligt in dit interval dus (net als bij onderdeel c) concluderen we dat kinderen met Downsyndroom qua borstvoeding niet significant van de gemiddelde Nederlandse bevolking verschillen. In theorie is het mogelijk dat de uitkomsten van c en d strijdig zijn, omdat we alleen bij d gebruik maken van een benadering.

- d. De groep kinderen met het syndroom van Down duiden we aan met de index 1, de controlegroep met 2.

We toetsen $H_0 : \pi_1 = \pi_2$ tegen $H_1 : \pi_1 \neq \pi_2$.

Zonder GR:

We vinden $p_1 = 19/42 \approx 0,4524$ en $p_2 = 27/37 \approx 0,7297$.

Onder H_0 schatten we de gemeenschappelijke waarde van π_1 en π_2 met

$$p = (19 + 27)/(42 + 37) = 46/79 \approx 0,5823.$$

Onder H_0 heeft

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)(n_1^{-1} + n_2^{-1})}} \text{ bij benadering een standaardnor-}$$

male verdeling.

De benadering is acceptabel, want $42 \cdot 0,5823$, $37 \cdot 0,5823$, $42 \cdot (1 - 0,5823)$ en $37 \cdot (1 - 0,5823)$ zijn alle ≥ 5 .

$$\text{We vinden } Z \approx \frac{0,4524 - 0,7297}{\sqrt{0,5823(1 - 0,5823)(42^{-1} + 37^{-1})}} \approx -2,49.$$

De p-waarde van de toets is $\Pr(|Z| \geq 2,49) \approx 2 \cdot \text{normalcdf}(2.49, 10^99) = 0,013$.

We kunnen de nulhypothese verwerpen. Het verschil tussen beide groepen is significant.

Met GR:

We gebruiken `2 - PropZTest` en voeren in: `x1:19 n1: 42`
`x2: 27 n2: 37 p1:≠p2`.

`Calculate` geeft dezelfde resultaten.

- e. *Zonder GR:*

Het betrouwbaarheidsinterval voor $\pi_1 - \pi_2$ wordt (bij benadering) gegeven door

$$p_1 - p_2 \pm z \cdot \sigma(p_1 - p_2) \text{ met } \sigma(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

De benadering is acceptabel omdat $19, 27, 42-19$ en $37-27$ alle ≥ 5 zijn.

Met $z = \text{invNorm}(0.975) = 1,96$ vinden we $[-0,49 ; -0,07]$.

Met GR:

We gebruiken `2 - PropZInt` met dezelfde invoer als bij onderdeel a en `C-level:0.95`. `Calculate` geeft het gevraagde betrouwbaarheidsinterval.

Aangezien 0 niet tot dit interval behoort, kunnen we (zoals verwacht) de nulhypothese uit onderdeel a verwerpen. Toch zijn beide methodes niet equivalent omdat we bij het toetsen van de hypothese een (iets) andere toetsingsgrootte gebruiken dan voor de bepaling van het betrouwbaarheidsinterval. In de praktijk zullen beide methodes zelden tot verschillende conclusies leiden.

10 Hoofdstuk 10

Opgave 1

1. Variable: Gewicht

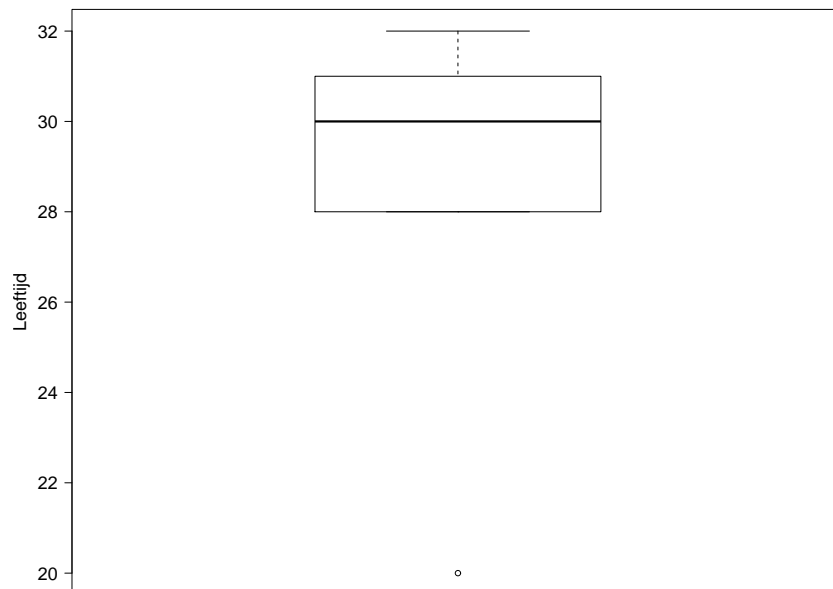
	mean	sd	0%	25%	50%	75%	100%	n
jongen	3008.00	91.92388	2943	2975.50	3008.0	3040.5	3073	2
meisje	3018.75	110.96959	2882	2974.25	3020.5	3065.0	3152	4

Variable: Leeftijd

	mean	sd	0%	25%	50%	75%	100%	n
jongen	25.00	7.071068	20	22.5	25.0	27.50	30	2
meisje	30.25	1.707825	28	29.5	30.5	31.25	32	4

2. Een boxplot van de leeftijd van de moeders staat in figuur 1:

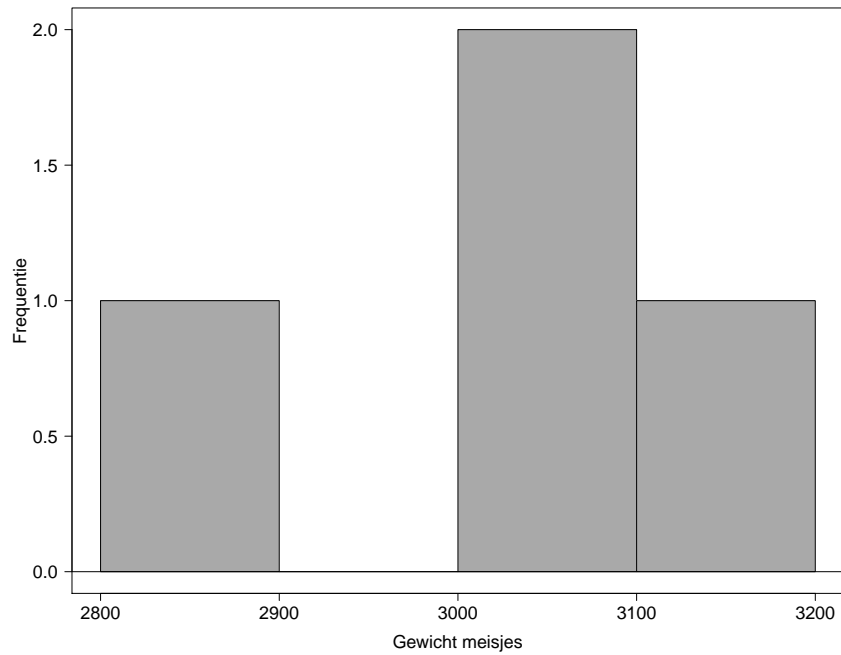
```
> boxplot(opgave1$Leeftijd, cex.axis = 1.5, cex.lab = 1.5, ylab = "Leeftijd",  
+         las = 1)
```



Figuur 1:

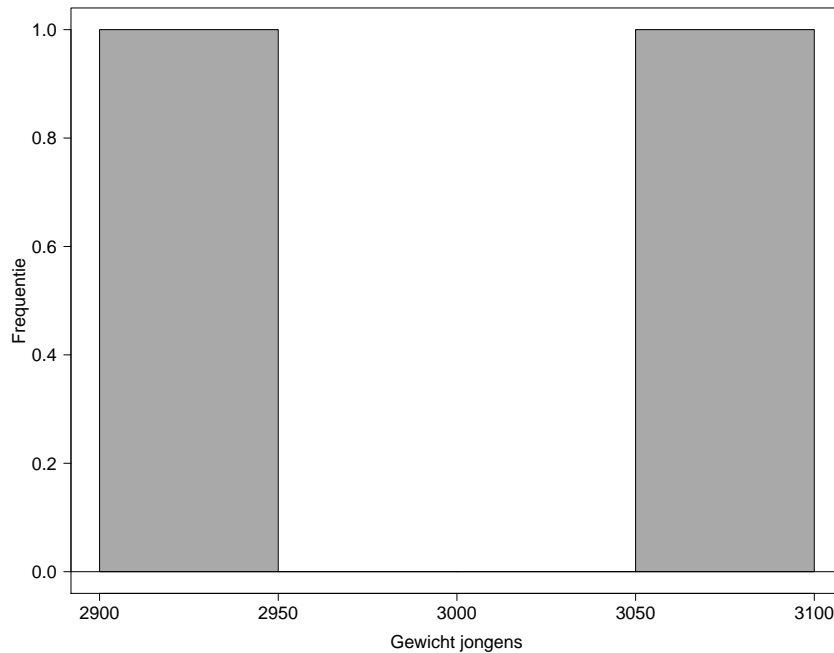
3. Zie figuur 2 en 3 voor de 2 histogrammen:

```
> submeisjes <- subset(opgave1, subset = Geslacht == "meisje")
> Hist(submeisjes$Gewicht, xlab = "Gewicht meisjes", cex.axis = 1.5,
+       cex.lab = 1.5, ylab = "Frequentie", scale = "frequency",
+       breaks = "Sturges", col = "darkgray", las = 1)
```



Figuur 2:

```
> subjongens <- subset(opgave1, subset = Geslacht == "jongen")
> Hist(subjongens$Gewicht, xlab = "Gewicht jongens", cex.axis = 1.5,
+       cex.lab = 1.5, ylab = "Frequentie", scale = "frequency",
+       breaks = "Sturges", col = "darkgray", las = 1)
```



Figuur 3:

NB: merk op dat het bij boxplots en scatterplots ook mogelijk is om in 1 plaatje de twee verschillende groepen aan te geven. In beide gevallen kun je de hele dataset gebruiken en de knop 'plot by groups' gebruiken.

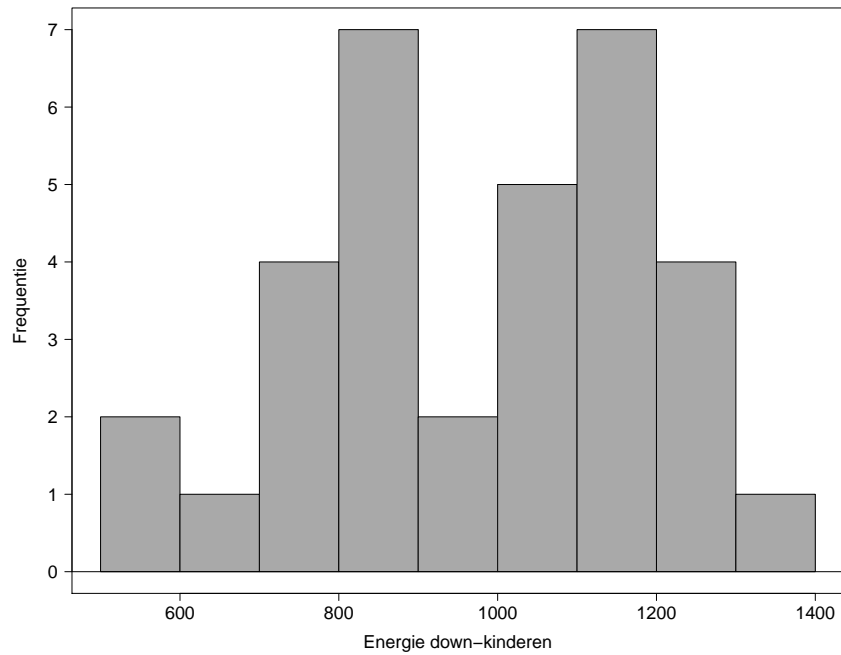
Opgave 2

4. Numeriek-continu.
5. Gebruik weer twee keer 'Subset Active Data Set'. Denk aan het dubbele =teken en aanhalingstekens rondom Down-syndroom. Na de eerste keer is je deeldataset de actieve dataset. Je krijgt de oorspronkelijke dataverzameling weer terug via 'Data' - 'Active Data Set' - 'Select Active Data Set'. Je ziet nu ook dat het handig is alle dataverzamelingen waarmee je werkt verschillende, informatieve namen te geven. Zie figuren 4 en 5.

```

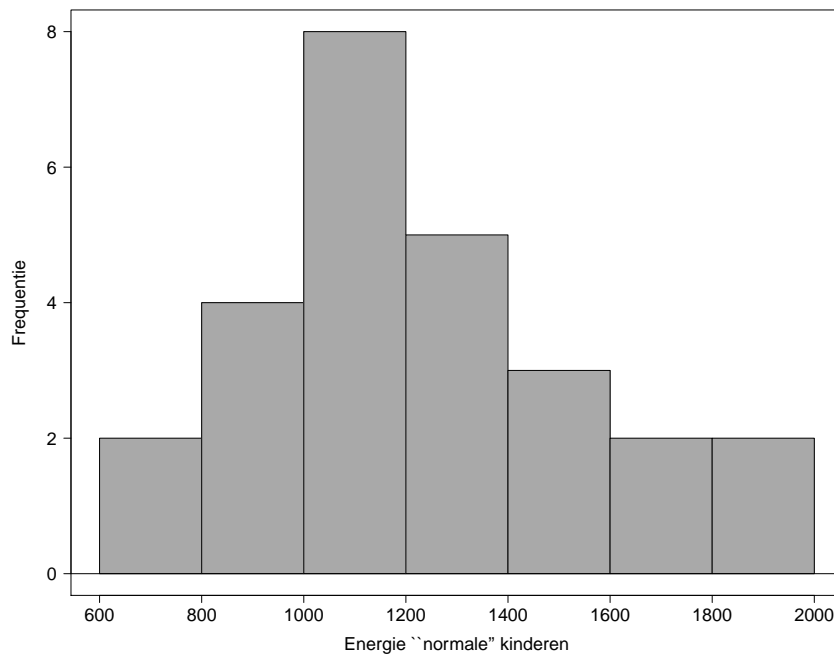
> subdown <- subset(down, subset = groep == "Down-syndroom")
> Hist(subdown$energie, xlab = "Energie down-kinderen", cex.axis = 1.5,
+       cex.lab = 1.5, ylab = "Frequentie", scale = "frequency",
+       breaks = "Sturges", col = "darkgray", las = 1)

```



Figuur 4:

```
> subnormaal <- subset(down, subset = groep == "Normaal")
> Hist(subnormaal$energie, xlab = "Energie ``normale'' kinderen",
+       cex.axis = 1.5, cex.lab = 1.5, ylab = "Frequentie", scale = "frequency",
+       breaks = "Sturges", col = "darkgray", las = 1)
```



Figuur 5:

6. Gebruik weer de oorspronkelijke dataset.

```
> numSummary(down[, "energie"], groups = down$groep, statistics = c("mean",
+   "sd", "quantiles"), quantiles = c(0, 0.25, 0.5, 0.75, 1))
```

	mean	sd	0%	25%	50%	75%	100%	n
Down-syndroom	976.394	209.6163	544	804.00	1018	1152.00	1363	33
Normaal	1214.115	321.7358	688	1008.25	1149	1408.75	1953	26

7. Kijk naar energie van de twee groepen apart. Tamelijk symmetrisch: gemiddelde en mediaan liggen bij elkaar in de buurt, spreiding rondom het gemiddelde links niet heel anders dan rechts. Het is van belang dit te controleren om te weten of je een t-toets mag doen. Het mag in dit geval.

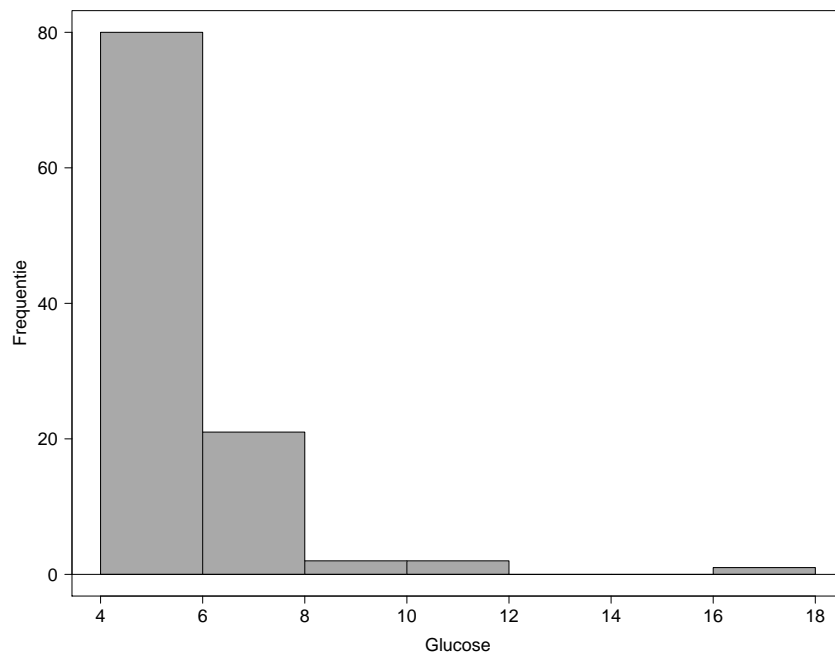
Opgave 3

```
8. > numSummary(glucose[, "glucose"], statistics = c("mean", "sd",
+           "quantiles"), quantiles = c(0.5))
```

```
      mean      sd    n
5.759717 1.506129 106
```

9. Zie figuur 6.

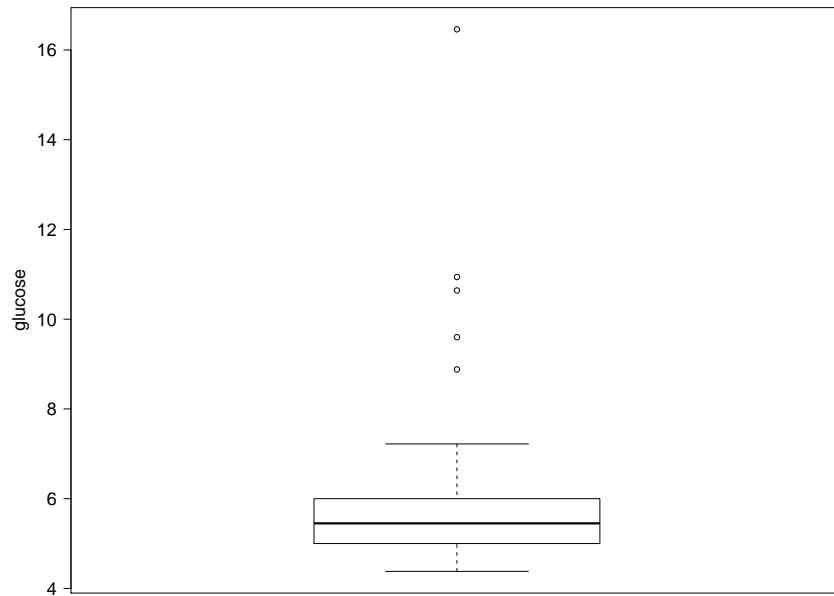
```
> Hist(glucose$glucose, xlab = "Glucose", cex.axis = 1.5, cex.lab = 1.5,
+       ylab = "Frequentie", scale = "frequency", breaks = "Sturges",
+       col = "darkgray", las = 1)
```



Figuur 6:

10. Zie figuur 7.


```
> boxplot(glucose$glucose, cex.axis = 1.5, cex.lab = 1.5, ylab = "glucose",
+         las = 1)
```



Figuur 7:

11. Het referentie-interval geeft aan waarbinnen naar verwachting 95 % van je waarnemingen ligt. Let op: dit is wat anders dan een betrouwbaarheidsinterval, dat aangeeft hoe goed je schatter is. Ook als je alle parameters van een verdeling kent en niets schat, dan nog zijn de waarnemingen gespreid rondom het (echte) gemiddelde. Dat druk je uit met het referentieinterval. In de praktijk schat je het referentieinterval meestal met behulp van het geschatte gemiddelde en de geschatte standaarddeviatie; als je ervan uitgaat dat je data normaal verdeeld zijn, schat je het als: $\text{gemiddelde} \pm 1.96 \times \text{SD}$.

Op grond van normale verdeling: (2.807704, 8.71173).

Op grond van percentielen: (4.4775, 9.99).

12. Methode 2), want de data zijn duidelijk niet normaal verdeeld.

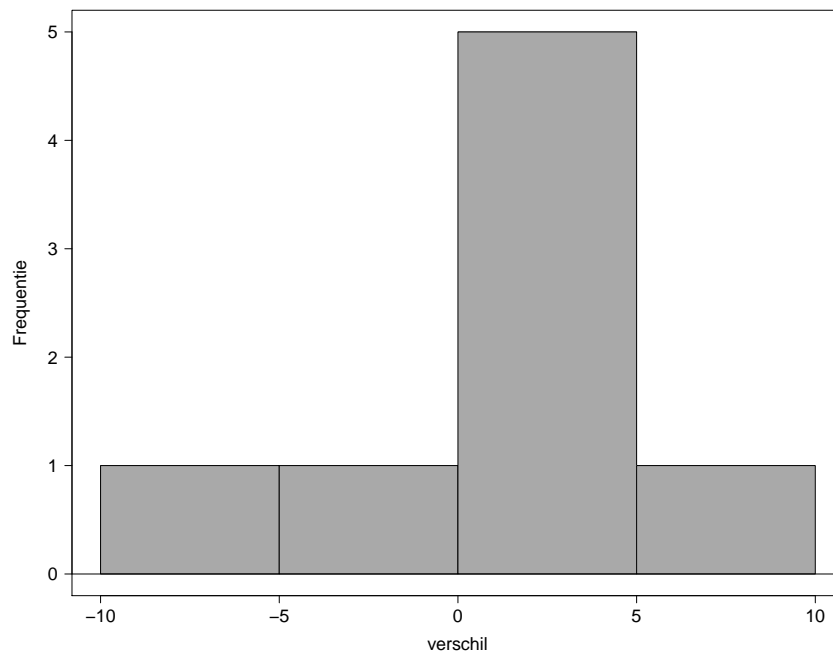
Opgave 4

13. Numeriek-continu.

14. -

15. Zie figuur 8.

```
> Hist(bloeddruk$verschil, xlab = "verschil", cex.axis = 1.5, cex.lab = 1.5,  
+       ylab = "Frequentie", scale = "frequency", breaks = "Sturges",  
+       col = "darkgray", las = 1)
```



Figuur 8:

Opgaven met toetsen en schatten

Toetsen

Opgave 1

16. -

```
17. > numSummary(down[, "energie"], groups = down$groep, statistics = c("mean",  
+       "sd", "quantiles"), quantiles = c(0.5))
```

	mean	sd	n
Down-syndroom	976.394	209.6163	33
Normaal	1214.115	321.7358	26

Voor de histogrammen zie figuren 5 en 6.

18. Energieinname per groep is redelijk normaal verdeeld:
- gemiddelde en mediaan liggen bij elkaar in de buurt, dus tamelijk symmetrisch;
 - spreiding rondom het gemiddelde links en rechts vergelijkbaar.
- Ongepaarde t-toets omdat de 2 groepen onafhankelijk zijn.
19. Nullhypothese: verschil in gemiddelde energieinname van Down en niet-Down kinderen is nul.
 Alternatief: verschil is ongelijk aan nul.

Two Sample t-test

```
data: energie by groep
t = -3.4247, df = 57, p-value = 0.001147
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-376.7194 -98.7235
sample estimates:
mean in group Down-syndroom      mean in group Normaal
                976.394                1214.115
```

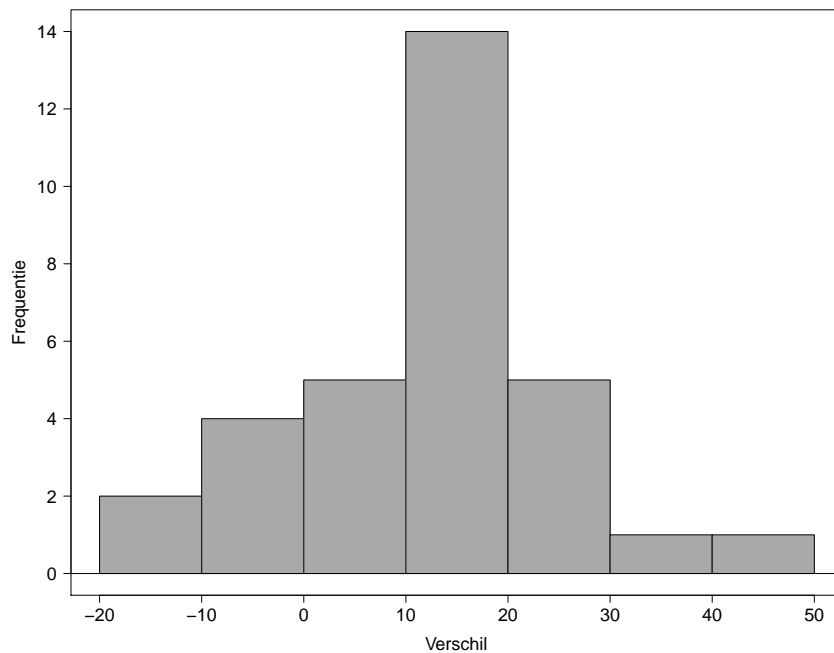
Vrijheidsgraden: $57 = 59$ (aantal personen) - 2
 Betrouwbaarheidsinterval = confidence interval. 0 ligt niet binnen dit interval.
 Bij deze alpha moet je de nulhypothese verwerpen: er is een significant verschil tussen gemiddelde energieinname van kinderen met en zonder Down.

Opgave 2

20. Numeriek-continue data. Gepaarde gegevens, want dezelfde persoon wordt telkens 2x gemeten.
21. mean sd n
 14.375 13.48536 32

Zie p. 5 van het dictaat:
 $SEM(\text{verschil}) = \frac{sd}{\sqrt{n}} = \frac{13,48536}{\sqrt{32}} = 2,383897$.
 $BI(\text{verschil}) = 14,375 \pm t_{[31],0,05} \times 2,383 = 14,375 \pm 2,039513 \times 2,383 = (9,51484; 19,23516)$.
 Zie afbeelding 9 voor het histogram.

```
> Hist(wao$verschil, xlab = "Verschil", cex.axis = 1.5, cex.lab = 1.5,
+       ylab = "Frequentie", scale = "frequency", breaks = "Sturges",
+       col = "darkgray", las = 1)
```



Figuur 9:

22. De gepaarde t-toets gaat er net als de ongepaarde t-test vanuit dat de data normaal verdeeld is. De input is in dit geval niet de percentages, maar het verschil in percentages. De verschilvariabele moet dus ongeveer normaal verdeeld zijn. Gezien het histogram van de vorige vraag is dit een redelijke aanname.

23. Paired t-test

```
data: wao$arts1 and wao$arts2
t = 6.03, df = 31, p-value = 1.127e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 9.51301 19.23699
sample estimates:
mean of the differences
      14.375
```

Betrouwbaarheidsinterval is gelijk aan het bovenstaande. Conclusie:

verwerp nulhypothese dat artsen gemiddeld hetzelfde percentage concluderen.

24. Zie dictaat p. 11:

$$T = \frac{\bar{X} - \mu_0}{SEM} = \frac{14,375 - 0}{2,383897} = 6.030042.$$

0 zit niet in het BI; dat is equivalent aan de nulhypothese verwerpen.

Opgave 3

25.

	borstv	
groep	ja	nee
Down-syndroom	28	14
Normaal	29	8

	borstv			
groep	ja	nee	Total	Count
Down-syndroom	66.7	33.3	100	42
Normaal	78.4	21.6	100	37

Bij “normale” kinderen lijkt het percentage borstgevoeden hoger te zijn.

26. Chi-kwadraat toets. Aanname is dat toetsingsgrootheid

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

normaal verdeeld is onder de nulhypothese.

```
> .Table <- xtabs(~groep + borstv, data = borstvoeding)
> .Table
```

	borstv	
groep	ja	nee
Down-syndroom	28	14
Normaal	29	8

```
> .Test <- chisq.test(.Table, correct = FALSE)
> .Test
```

Pearson's Chi-squared test

```
data: .Table
X-squared = 1.3428, df = 1, p-value = 0.2465
```

```
> remove(.Test)
> fisher.test(.Table)
```

Fisher's Exact Test for Count Data

```
data: .Table
p-value = 0.3171
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1730478 1.6836068
sample estimates:
odds ratio
 0.555889

> remove(.Table)
```

Beide toetsen verwerpen niet de nulhypothese dat de percentages in beide groepen gelijk zijn (neem $\alpha = 0.05$). De odds ratio is gedefinieerd als $\frac{p_1(1-p_1)}{p_2(1-p_2)}$; nulhypothese is dat de odds ratio gelijk is aan 1.

```
27.          borstv8
groep        ja  nee
Down-syndroom 19  23
Normaal       27  10
```

```
          borstv8
groep        ja  nee Total Count
Down-syndroom 45.2 54.8   100     42
Normaal       73.0 27.0   100     37
```

Pearson's Chi-squared test

```
data: .Table
X-squared = 6.221, df = 1, p-value = 0.01262
```

Fisher's Exact Test for Count Data

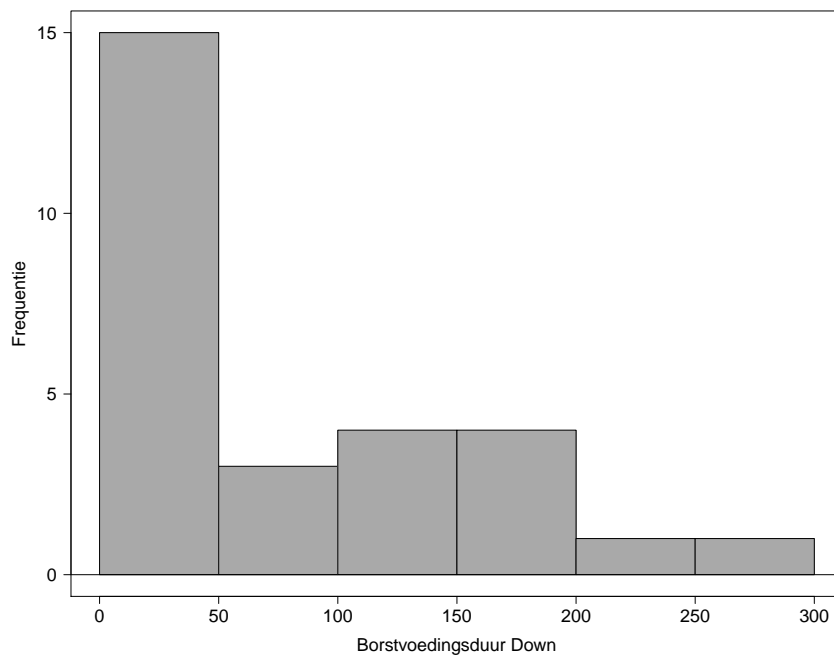
```
data: .Table
p-value = 0.02162
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1055302 0.8646113
sample estimates:
odds ratio
 0.3107606
```

Beide toetsen wijzen nu wel op een significant verschil.

```
28.               mean      sd 0% 25% 50% 75% 100%  n NA
Down-syndroom 78.96429 83.39353  1   6  36 150  270 28 16
Normaal       77.48276 45.10038  4  42  90 120  180 29  8
```

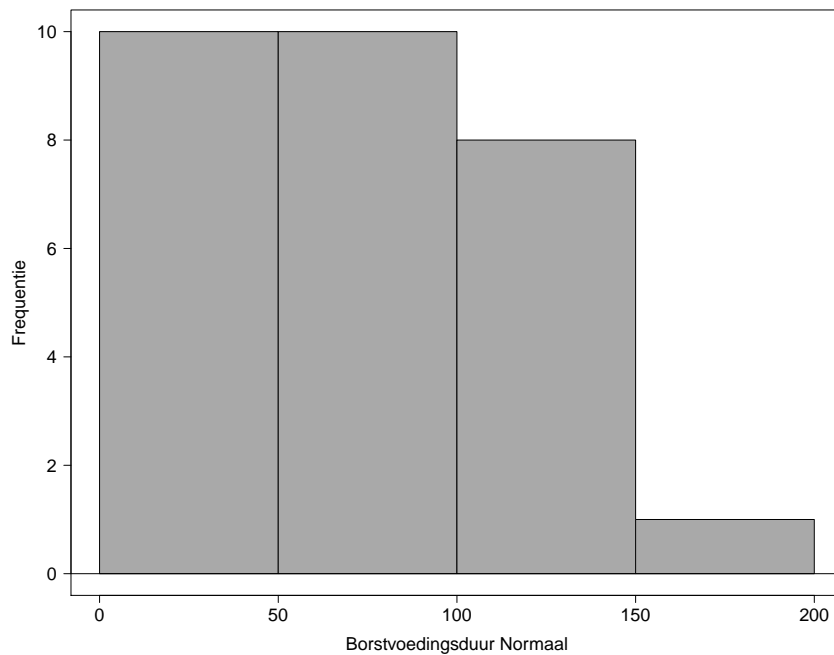
Gemiddelde en mediaan liggen ver uit elkaar, zeker in de Down-groep.
Verdeling is niet normaal.

```
> subdown <- subset(borstvoeding, subset = groep == "Down-syndroom")
> Hist(subdown$bvduur, xlab = "Borstvoedingsduur Down", cex.axis = 1.5,
+      cex.lab = 1.5, ylab = "Frequentie", scale = "frequency",
+      breaks = "Sturges", col = "darkgray", las = 1)
```



Figuur 10:

```
> subnormaal <- subset(borstvoeding, subset = groep == "Normaal")
> Hist(subnormaal$bvduur, xlab = "Borstvoedingsduur Normaal", cex.axis = 1.5,
+      cex.lab = 1.5, ylab = "Frequentie", scale = "frequency",
+      breaks = "Sturges", col = "darkgray", las = 1)
```



Figuur 11:

29. Vanwege niet-normale verdeling en kleine aantallen kun je beter een niet-parametrische toets gebruiken, in dit geval Wilcoxon's rangteken-toets (zie dictaat paragraaf 4.2).
30. Je vindt Wilcoxon's rangteken-toets via 'Statistics' - 'Nonparametric tests' - 'Two-sample Wilcoxon test'.

Down-syndroom	Normaal
36	90

Wilcoxon rank sum test with continuity correction

```
data: bvduur by groep
W = 357, p-value = 0.4354
alternative hypothesis: true location shift is not equal to 0
```

Je kunt de nulhypothese (er zit geen verschil tussen de verdelingen van de borstvoedingsduur van beide groepen) niet verwerpen. Het

lijkt erop dat relatief veel moeders van Down-kinderen vroegtijdig met de borstvoeding moeten stoppen. Het is mogelijk dat dit veroorzaakt wordt door de gemiddeld slechtere gezondheid van Down-kinderen: borstvoeding is voor veel baby's moeilijker te leren dan flesvoeding en Downkinderen hebben vaak gezondheidscomplicaties. De overgebleven moeders van Down-kinderen gaan wel lang door, waardoor het gemiddelde van de Down-groep bijna gelijk is aan dat van de normale groep. Dit zou ook weer te maken kunnen hebben met de gemiddeld slechtere gezondheid van Down-kinderen: borstvoeding is gezonder dan flesvoeding, wat juist voor ongezonde kinderen extra belangrijk is. Deze data helpen hypothesen te formuleren die je in een vervolgonderzoek verder kunt bestuderen.

Opgave 4

31. -

32. man vrouw
 38 73

 man vrouw
34.23423 65.76577

 rookt rookt niet
 7 104

 rookt rookt niet
6.306306 93.693694

 trap roltrap
 87 24

 trap roltrap
78.37838 21.62162

 lopen fiets auto bus trein tandem duikboot
 5 48 0 11 45 1 1

 lopen fiets auto bus trein tandem duikboot
4.5045045 43.2432432 0.0000000 9.9099099 40.5405405 0.9009009 0.9009009

	mean	sd	0%	25%	50%	75%	100%	n	NA
Alcoholgebruik	8.127273	10.931907	0	1.25	5	10	80	110	1
Leeftijd	19.000000	1.328020	17	18.00	19	19	25	111	0
Lengte	175.936937	8.893592	160	168.00	175	181	204	111	0

33. Rookgedrag

Geslacht	rookt	rookt niet
man	4	34
vrouw	3	70

Rookgedrag				
Geslacht	rookt	rookt niet	Total	Count
man	10.5	89.5	100	38
vrouw	4.1	95.9	100	73

Pearson's Chi-squared test

data: .Table
X-squared = 1.7415, df = 1, p-value = 0.1869

Chi-kwadraat-toets; p-waarde is 0,19, dus nulhypothese dat percentages gelijk zijn kun je niet verwerpen.

34. Ongepaarde t-toets.

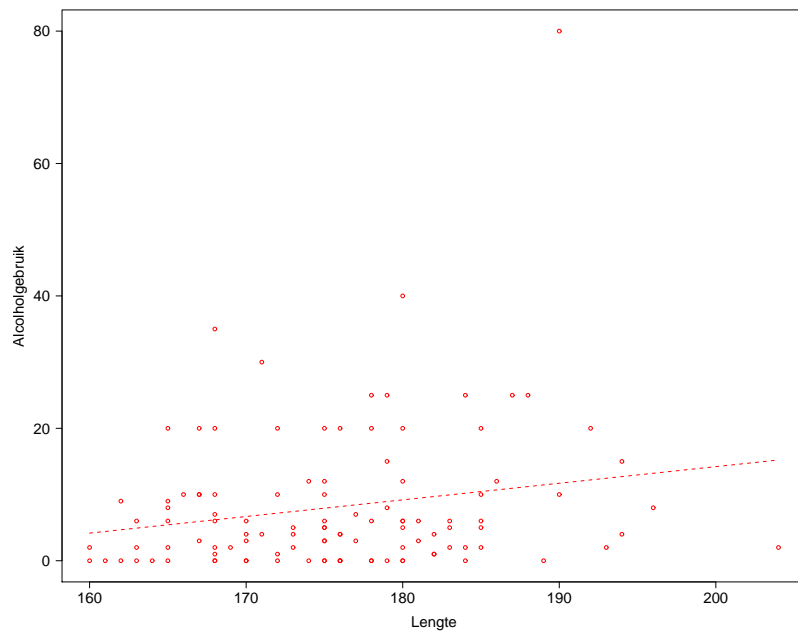
Two Sample t-test

data: Lengte by Geslacht
t = 8.1255, df = 109, p-value = 7.561e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
8.664679 14.255292
sample estimates:
mean in group man mean in group vrouw
183.4737 172.0137

De p-waarde is heel klein en dus mag je concluderen dat er wel degelijk een verschil in lengte is tussen mannen en vrouwen.

35. Zie figuur 12.

```
> scatterplot(Alcoholgebruik ~ Lengte, xlab = "Lengte", cex.axis = 1.5,
+             cex.lab = 1.5, ylab = "Alcoholgebruik", reg.line = lm, smooth = FALSE,
+             labels = FALSE, boxplots = FALSE, span = 0.5, data = enquete,
+             las = 1)
```



Figuur 12:

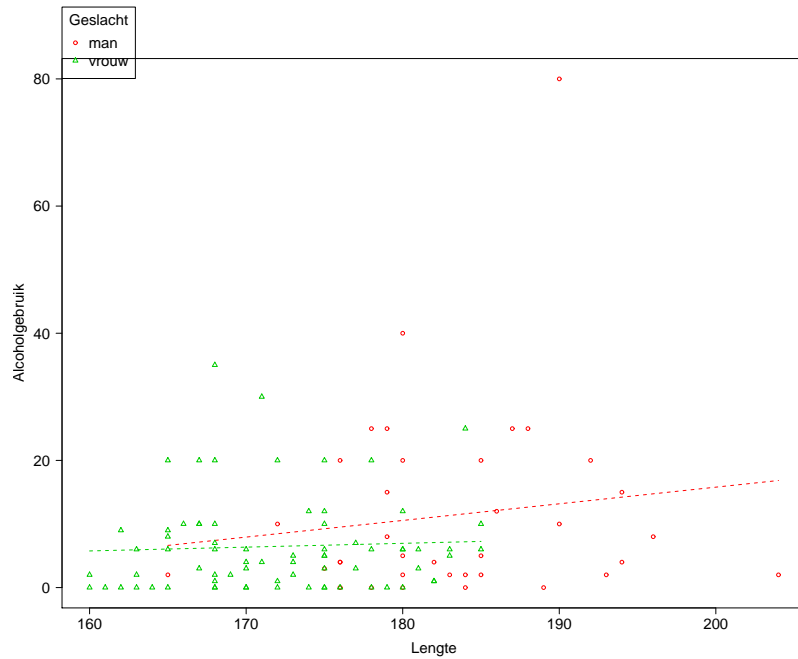
36. Pearson's product-moment correlation

```
data: enquete$Alcoholgebruik and enquete$Lengte
t = 2.1251, df = 108, p-value = 0.03586
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01361054 0.37356948
sample estimates:
      cor
0.2003414
```

Positieve correlatie (0,20), associatie significant (p-waarde <0,05).

37. Het 'ware' verband gaat via het geslacht (zie figuur 13). Voor de vrouwen verdwijnt de associatie lengte-alcoholgebruik, voor de mannen wordt de correlatiecoëfficiënt kleiner (0.1318951).

```
> scatterplot(Alcoholgebruik ~ Lengte | Geslacht, reg.line = lm,
+ smooth = FALSE, labels = FALSE, boxplots = FALSE, span = 0.5,
+ by.groups = TRUE, data = enquete, las = 1, cex.axis = 1.5,
+ cex.lab = 1.5)
```



Figuur 13:

Toetsen en regressie

38. -.

```
39.          mean          sd  0%  50% 100%   n
gewicht 1291.9437 386.301203 530 1260 2382 160
zwschduu  29.0125   1.987342  23   29   31 160
```

```
meisje jongen
  66      94
```

```
meisje jongen
41.25 58.75
```

```
eenling meerling
  111      49
```

eenling	meerling
69.375	30.625

geen sectio	sectio
95	65

geen sectio	sectio
59.375	40.625

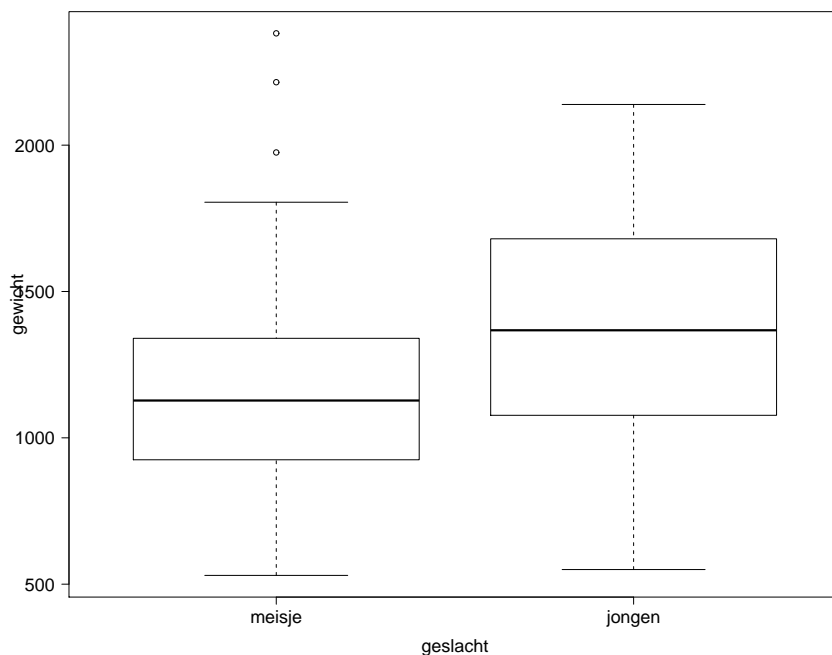
40. Let op dat je de goede percentages neemt.

	sectio	
meerl	geen	sectio
eenling	54	57
meerling	41	8

	sectio			
meerl	geen	sectio	sectio	Total Count
eenling	48.6	51.4	100	111
meerling	83.7	16.3	100	49

Van de eenlingen is 51.4% met een keizersnede geboren, van de meerlingen 16.3%.

```
41. > boxplot(gewicht ~ geslacht, ylab = "gewicht", xlab = "geslacht",
+           cex.axis = 1.5, cex.lab = 1.5, data = zwangerschap, las = 1)
```



Figuur 14:

Jongens lijken gemiddeld iets meer te wegen dan meisjes.

```
42.           mean  n
meisje 1175.682 66
jongen 1373.574 94
```

We hebben een continue uitkomst en twee niet-gepaarde groepen. De groepen zijn groot en gewicht is geen extreem scheef verdeelde variabele. De ongepaarde t-toets is dus de meest geschikte methode. Dat geeft de volgende uitvoer:

Two Sample t-test

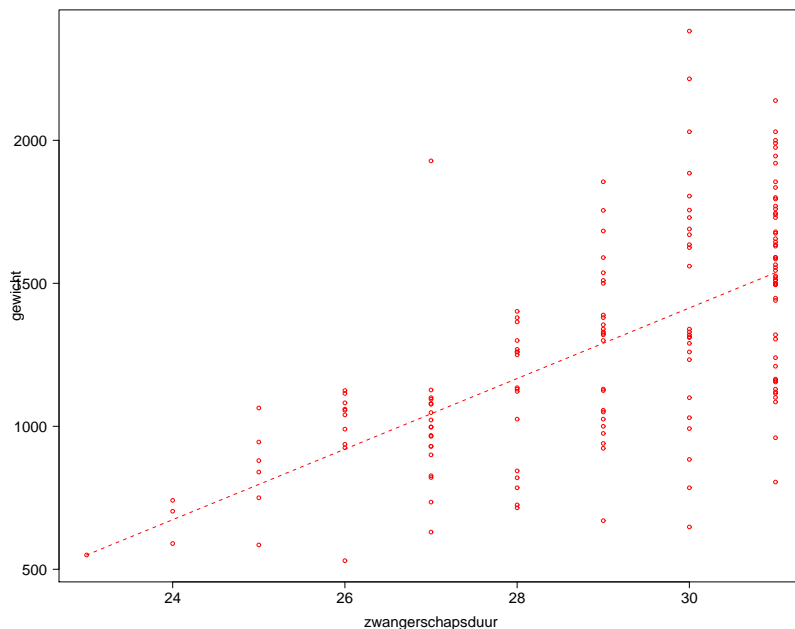
```
data:  gewicht by geslacht
t = -3.2868, df = 158, p-value = 0.001249
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -316.81023  -78.97507
sample estimates:
```

mean in group meisje	mean in group jongen
1175.682	1373.574

De p-waarde die bij de t-test hoort is 0.001. Deze is veel kleiner dan 0.05. De jongens wegen gemiddeld 198 gram meer dan de meisjes (95% CI (78.9 - 317 gram) en dat is een statistisch significant verschil.

43. Vul $y=ax+b$ in voor 24 weken en 30 weken en bereken daaruit x en y.

```
> scatterplot(gewicht ~ zwschduu, xlab = "zwangerschapsduur", cex.axis = 1.5,
+ cex.lab = 1.5, ylab = "gewicht", reg.line = lm, smooth = FALSE,
+ labels = FALSE, boxplots = FALSE, span = 0.5, data = zwangerschap,
+ las = 1)
```



Figuur 15:

44. De uitvoer met de regressiecoëfficiënten is:

Call:

```
lm(formula = gewicht ~ zwschduu, data = zwangerschap)
```

Residuals:

Min	1Q	Median	3Q	Max
-765.81	-188.55	17.30	193.29	968.19

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2288.52	347.44	-6.587	6.33e-10 ***
zwschduu	123.41	11.95	10.329	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 299.4 on 158 degrees of freedom

Multiple R-squared: 0.4031, Adjusted R-squared: 0.3993

F-statistic: 106.7 on 1 and 158 DF, p-value: < 2.2e-16

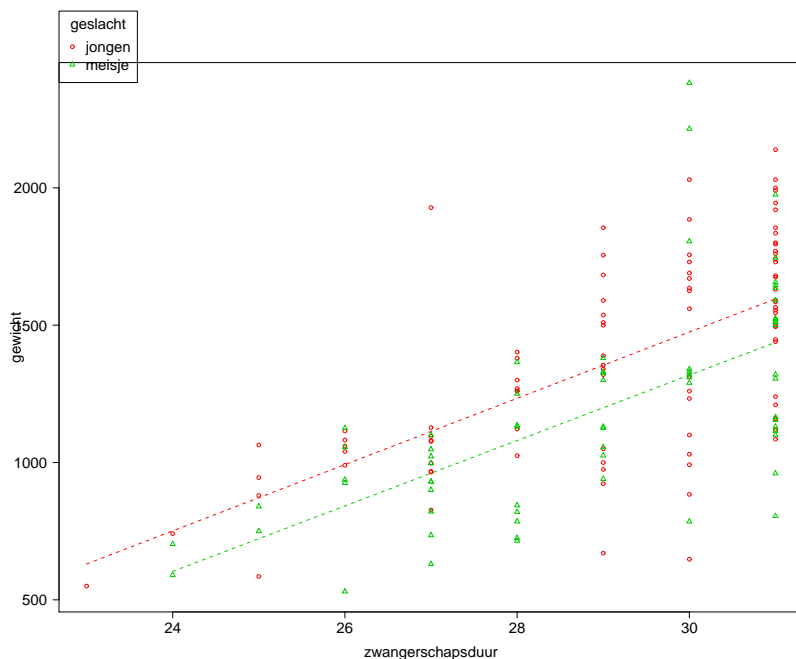
De bijbehorende regressievergelijking wordt dan: gemiddeld gewicht
= $-2288.522 + 123.411 \times \text{zwangerschapsduur}$.

45. De regressiecoëfficiënt 123.41 is de richtingscoëfficiënt van de regressielijn. Als de zwangerschapsduur met een week toeneemt, neemt het gemiddelde geboortegewicht met 123.41 gram toe. De p-waarde die bij deze richtingscoëfficiënt hoort is 0.000. De richtingscoëfficiënt verschilt dus significant van 0, er is een significante relatie tussen zwangerschapsduur en geboortegewicht.


```

46. > scatterplot(gewicht ~ zwschduu | geslacht, , xlab = "zwangerschapsduur",
+   ylab = "gewicht", cex.axis = 1.5, cex.lab = 1.5, reg.line = lm,
+   smooth = FALSE, labels = FALSE, boxplots = FALSE, span = 0.5,
+   by.groups = TRUE, data = zwangerschap, las = 1)

```



Figuur 16:

```

47. Call:
lm(formula = gewicht ~ zwschduu + geslacht, data = zwangerschap)

```

Residuals:

Min	1Q	Median	3Q	Max
-826.42	-156.17	23.13	156.82	1062.69

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2280.58	336.97	-6.768	2.45e-10 ***
zwschduu	120.00	11.63	10.315	< 2e-16 ***
geslacht[T.jongen]	155.11	46.82	3.313	0.00115 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 290.4 on 157 degrees of freedom

Multiple R-squared: 0.4421, Adjusted R-squared: 0.435
F-statistic: 62.21 on 2 and 157 DF, p-value: < 2.2e-16

De regressievergelijking wordt nu:

gemiddeld geboortegewicht = $-2280.58 + 120.00 \times \text{zwangerschapsduur}$
+ $155.11 \times \text{geslacht}$.

48. Meisje=0, jongen=1. Om de regressievergelijking voor de meisjes te krijgen vullen we geslacht=0 in.

gemiddeld geboortegewicht = $-2280.57 + 120.00 \times \text{zwangerschapsduur}$

Voor de jongens vullen we geslacht=1 in:

gemiddeld geboortegewicht = $-2125.46 + 120.00 \times \text{zwangerschapsduur}$.

49. De regressiecoëfficiënt voor zwangerschapsduur geeft aan dat het gemiddelde geboortegewicht 120 gram stijgt als de zwangerschapsduur 1 week toeneemt, gecorrigeerd voor geslacht (dat wil zeggen dat je nu de invloed van geslacht direct bekijkt, en niet via een eventueel verschillende zwangerschapsduur). De coëfficiënt voor geslacht geeft aan dat bij gelijke zwangerschapsduur jongens gemiddeld 155.11 gram zwaarder zijn bij de geboorte.