Statistiek Epidemiologie en Economie

Jan van den Broek

Inhoudsopgave

1 Correlatie **1**
- 1.1 Covariantie . . . 1
- 1.2 Correlatie . . . 2

2 Het lineaire model 1: Regressie Analyse **5**
- 2.1 Een onderzoek . . . 5
- 2.2 De populatie: een model voor het data genererend proces. . . . 5
- 2.3 De steekproef . . . 7
  - 2.3.1 Kwadraatsommen . . . 7
  - 2.3.2 t-toets . . . 11
  - 2.3.3 r^2 . . . 12

3 Het lineaire model 2: Anova **13**
- 3.1 Een onderzoek . . . 13
- 3.2 De populatie: een model voor het data genererend proces . . . 15
- 3.3 De steekproef . . . 16

4 Lineaire modellen 3: Uitbreidingen **21**
- 4.1 Lineaire regressie met meerdere onafhankelijk continue variabelen . . 21
- 4.2 Anova met meer dan 1 factor . . . 23
- 4.3 Lineaire regressie met indicator variabelen . . . 24

5 Logistische regressie **27**
- 5.1 Introductie . . . 27
- 5.2 De populatie . . . 27
- 5.3 De steekproef . . . 28
- 5.4 Het logistische regressie model . . . 29

6 Survival analyse **33**
6.1 Overleving 33
6.2 Conditionaliteit: Hazard 34
6.3 Censurering 34
6.4 Schatten van de survivalfractie 36
6.5 Standard errors 37
6.6 Prognose 39

Hoofdstuk 1

Correlatie

1.1 Covariantie

We hebben te maken met twee continue variabelen die ëven belangrijk"zijn, dat wil zeggen het onderzoek is opgezet om over beide iets te weten te komen. Dat iets is in veel gevallen hun samenhang. Op blz. 131 van het boek[1] staan de resultaten van een onderzoek naar bot activiteit bij paarden (tabel 10.1). Er zijn twee verschillende maten om de bot activiteit te meten: wBAP en PICP. Figuur 10.2 op blz 131 laat het verband tussen die twee maten zien in een zogenaamde scatterplot.

Een veel voorkomend verband is een lineair verband. Daar hebben we het hier over. De vraag is, of er tussen twee variabelen een lineair verband bestaat. Als je figuur (1.1) bekijkt dan zie je dat het verband niet volledig is. Was dat wel het geval dan zouden alle punten op een rechte lijn liggen.

Een maat voor het lineaire verband is de zogenaamde covariantie

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

. Deze formule lijkt veel op die voor de variantie (vervang x door y en je hebt de formule voor de variantie van y). In de populatie wordt deze aangegeven met σ_{xy} Dat deze formule een maat is voor het lineaire verband, kan met figuur(1.1) uit gelegd worden:

1. Als x groter is dan het gemiddelde $\bar{x}$ en y is groter dan het gemiddelde $\bar{y}$, dan

[1] Als er verwezen wordt naar het boek dan wordt bedoeld: Statistics for Veterinary and Animal science van Petrie & Watson.

is $(x_i - \bar{x})(y_i - \bar{y})$ positief. Dus deze punten hebben een positieve bijdrage aan de covariantie.

2. Als x kleiner is dan het gemiddelde $\bar{x}$ en y is kleiner dan het gemiddelde $\bar{y}$ dan is $(x_i - \bar{x})(y_i - \bar{y})$ positief. Dus deze punten hebben een positieve bijdrage aan de covariantie.

3. Als x groter is dan het gemiddelde $\bar{x}$ en y is kleiner dan het gemiddelde $\bar{y}$ dan is $(x_i - \bar{x})(y_i - \bar{y})$ negatief. Dus deze punten hebben een negatieve bijdrage aan de covariantie.

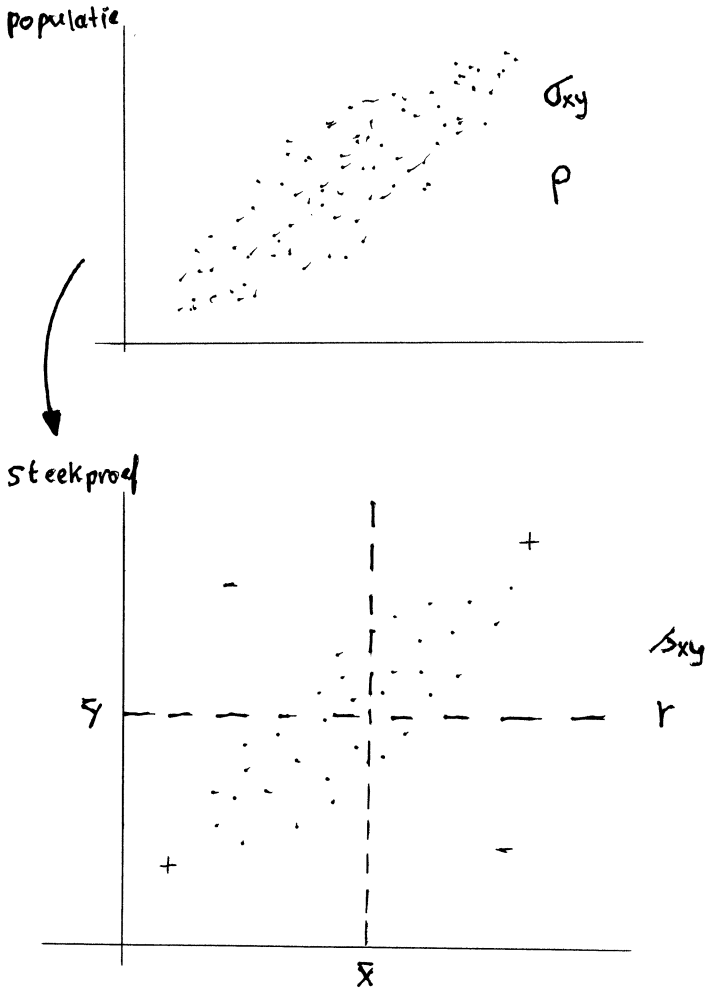
4. Als x kleiner is dan het gemiddelde $\bar{x}$ en y is groter dan het gemiddelde $\bar{y}$ dan is $(x_i - \bar{x})(y_i - \bar{y})$ negatief. Dus deze punten hebben een negatieve bijdrage aan de covariantie.

Als het aantal punten met een positieve bijdrage groter is dan het aantal punten met een negatieve bijdrage (stijgende puntenwolk) dan heet het verband positief; Als het aantal punten met een positieve bijdrage kleiner is dan het aantal punten met een negatieve bijdrage (dalende puntenwolk) dan heet het verband negatief. Zijn er ongeveer evenveel punten met een positieve bijdrage als punten met een negatieve bijdrage dan is de covariantie ongeveer nul.
Dus covariantie nul, geen lineair verband, covariantie positief een positief verband, covariantie negatief een negatief verband.
Het nadeel van de covariantie is dat deze maat afhangt van de eenheid waarin de variabelen gemeten zijn. Stel de waarde voor de y-as is gemeten in meters. Als deze schaal word veranderd naar centimeter dan wordt $y_i - \bar{y}$ 100 keer groter en dus de covariantie ook. De absolute grootte van de covariantie zegt dus niets.

1.2 Correlatie

Om de maat voor het lineaire verband onafhankelijk te maken van de meeteenheid (dus van de schaal waarop gemeten is) deelt men de covariantie door de standaard afwijking van x, s_x, (met als eenheid de eenheid van de x-as) en door de standaard afwijking van y, s_y (gemeten in de eenheid van de y-as). Zo krijgt men een maat voor het lineaire verband die schaal onafhankelijk is. Deze wordt de correlatie coëfficient genoemd:

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



Figuur 1.1: Lineaire samenhang

In de populatie wordt deze aangegeven met ρ. De correlatie coëfficient ligt altijd tussen -1 en 1. Bij -1 is er sprake van een volledig negatief lineair verband, bij 1 van een volledig positief lineair verband, en bij nul van geen lineair verband. Dat lineair is belangrijk want als de punten bijv. op een cirkel liggen is het verband volledig maar de correlatiecoëfficient is nul.
Om te toetsen of er sprake is van een lineair verband, kan de nulhypothese $H_0 : \rho = 0$ tegen het alternatief $H_1 : \rho \neq 0$ getoetst worden. Voor deze toetsing kan een t-toets gebruikt worden. Een t-toets kijkt altijd naar de afstand tussen wat je vindt in je onderzoek (hier de steekproefcorrelatie r) en de nulhypothese (die zegt dat $\rho = 0$) uitgedrukt in standard errors dus: $t = \frac{r-0}{se(r)}$. Deze t-waarde heeft onder de nulhypothese een studentverdeling met een aantal vrijheidsgraden. Om de t-waarde uit te kunnen rekenen moeten we de standard error van r weten. Deze is $se(r) = \frac{1}{\sqrt{\frac{n-2}{1-r^2}}}$. Als we het onderzoek erg vaak herhalen en in ieder onderzoek r bepalen en van al deze correlatiecoëfficiënten de standaardafwijking bepalen, dan zou dat $se(r)$ zijn. Dus de t-toets wordt ($test_{10}$ in het boek)

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

Deze t-toets heeft een student verdeling met $n-2$ vrijheidsgraden. (Om een lineair verband te bepalen, dus de lijn, heb je twee punten nodig dus hou je er $n-2$ over).

Hoofdstuk 2

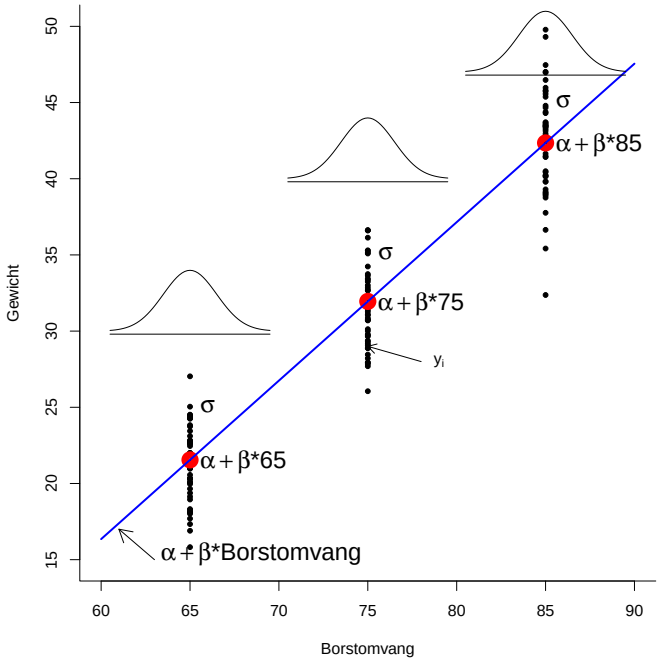
Het lineaire model 1: Regressie Analyse

2.1 Een onderzoek

In tabel 10.2 op blz. 135 van het boek, staat het levend gewicht van 66 schapen en hun borstomvang. Het gewicht van een schaap is vaak lastig te bepalen dus zou het handig zijn als we door de borstomvang te bepalen ook iets konden zeggen over het gewicht. Dan moet er echter wel een relatie zijn tussen het gewicht en de borstomvang. Bij regressie analyse wordt ervan uitgegaan dat die relatie lineair is, dus gegeven wordt door een rechte lijn. De vraag is nu hoe die lijn eruitziet. Dus de situatie is dat we een continue afhankelijke variabele (y) hebben en een andere continue variabele waarmee de afhankelijke variabele proberen lineair te beschrijven (x). Deze laatste heet ook wel de onafhankelijke variabele. Deze twee variabelen zijn dus niet gelijkwaardig zoals bij de correlatie: de een (onafhankelijke) wordt gebruikt om de ander (afhankelijke) lineair te beschrijven.
We kunnen een grafiek maken door de borstomvang op de x-as te zetten en het gewicht op de y-as. Deze z.g. scatter plot staat op blz 130.

2.2 De populatie: een model voor het data genererend proces.

De schapen uit het onderzoek vormen een steekproef uit een populatie. Hoe ziet de populatie eruit. Eerst wordt er een beschrijving gegeven van deze populatie. Anders



Figuur 2.1: Het regressie model in de populatie

gezegd we geven een beschrijving van de populatie die de data in de steekproef gegenereerd heeft. Deze beschrijving heet een model. Uiteraard is dit model een ruwe benadering van het data generend proces.

De populatie is de groep individuen waaruit de steekproef genomen is. In het geval van het voorbeeld zijn de individuen schapen. Er worden aan ieder schaap twee dingen gemeten : het gewicht en de borstomvang. Deze twee zijn niet gelijkwaardig. Het onderzoek gaat over het gewicht en de borstomvang wordt gebruikt om daar iets over te weten te komen: kan de borstomvang gebruikt worden om het gewicht lineair te beschrijven.

De algemene vorm van een rechte lijn is

$$y = \alpha + \beta x$$

y is de afhankelijke variabele en x is de onafhankelijke. α is het intercept van de lijn, het is de y coördinaat van het punt waar de lijn de y-as snijdt. β is de regressie coëfficient ofwel de helling van de lijn. Het geeft aan met hoeveel de y-variabele verandert als de x variabele met 1 toeneemt.
Als we in de populatie naar alle schapen kijken met een borstomvang van bijvoorbeeld

65 dan zouden we van al die schapen het gemiddelde gewicht kunnen bepalen. Dit gemiddelde is een punt op de lijn: het gemiddelde gewicht van alle schapen met een borstomvang van 65 cm is dus $\alpha + \beta \cdot 65$ (zie figuur 2.1). Dus het lineaire model, de rechte lijn, geldt voor de gemiddelden. De variantie van de gewichten van alle schapen met borstomvang 65 is σ^2 en dus is de standaard deviatie σ. Omdat de standaard deviatie in het algemeen aangeeft hoe de getallen rond het gemiddelde liggen en omdat het gemiddelde nu een punt op de lijn is, geeft deze standaard deviatie dus aan hoe de punten rond de lijn liggen. Als we van de gewichten van alle schapen met borstomvang 65 een histogram zouden maken dan zouden we iets vinden dat goed op een normale verdeling lijkt (zie figuur 2.1). Hetzelfde kunnen we nu doen bij een andere borstomvang bijv. 75 : het gemiddelde gewicht van alle schapen met borstomvang x_2 is dus $\alpha + \beta \cdot 75$. De variantie van de gewichten van alle schapen met borstomvang 75 is σ^2. Als we van de gewichten van alle schapen met borstomvang 75 een histogram zouden maken dan zouden we iets vinden dat goed op een normale verdeling lijkt.

Dit kunnen we voor iedere mogelijke borstomvang doen en krijgen dan een model voor de populatie waaruit de steekproef afkomstig is.

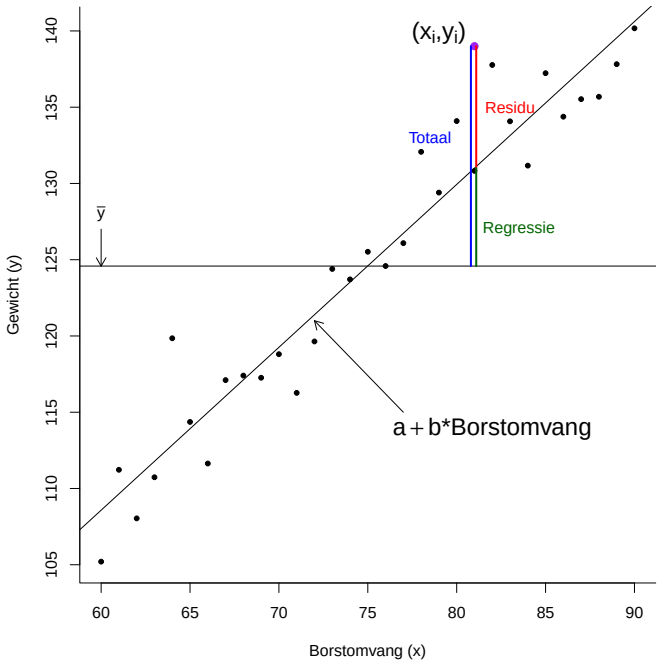
Elke waarneming uit de populatie kan nu worden beschreven als een punt op de lijn bij een zekere borstomvang plus een afwijking van de lijn die het residu wordt genoemd: $y_i = \alpha + \beta x_i + \epsilon_i$. Zie de waarneming y_i in figuur 2.1. De afstand tussen y_i en de rechte lijn (in y-richting) is het residu wat bij dit punt hoort.
Dus:

> Het lineaire model voor de regressie situatie is
>
> $$y_i = \alpha + \beta x_i + \epsilon_i$$
>
> waarbij y_i normaal verdeeld is met gemiddelde $\alpha + \beta x_i$ en variantie σ^2. Dit betekent dat de residuen ϵ_i een normale verdeling hebben met gemiddelde nul en met variantie σ^2.

2.3 De steekproef

2.3.1 Kwadraatsommen

In de steekproef moet de lijn geschat worden. De lijn wordt zo geschat dat de geschatte lijn het best bij de punten past. Die lijn wordt het best passend genoemd, waarvoor geldt dat de afstand van de punten tot de lijn (in y-richting) het kleinst is. Die



Figuur 2.2: Het regressie model in de steekproef

afstanden heten de residuen in dit geval de steekproef residuen. Zie de rode lin in figuur 2.2. Dus die lijn heet het best passend waarvoor geldt dat de residuen het kleinst zijn. Omdat het niet uitmaakt of de residu boven of onder de lijn ligt – de afstand tot de lijn blijft het zelfde – kunnen we ook naar de residuen in het kwadraat kijken. Nu zijn er net zoveel residuen als waarnemingen en dus ook net zoveel residuen in het kwadraat die allemaal op zijn kleinst moeten zijn. Als alle residuen in het kwadraat op zijn kleinst moeten zijn, dan moet dus de som van de residuen in het kwadraat op zijn kleinst zijn. De residuen gekwadrateerd en gesommeerd heet de residukwadraatsom. Een residu is de afstand van een punt tot de lijn (rode lijn in figuur 2.2) dus $y_i - (a + bx_i)$. De residu kwadraatsom is dan

$$SS_{res} = \sum_i \left[y_i - (a + bx_i)\right]^2$$

Dus die lijn is de best passende lijn bij de data punten in de steekproef, waarvoor de residu kwadraatsom op zijn kleinst zijn. Deze methode wordt ook wel kleinste kwadraten methode genoemd. De waarden a en b waarvoor geldt dat de residu kwadraatsom op zijn kleinst is kunnen worden uitgerekend met :

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Als a en b op deze manier uitgerekend worden dan is de residu kwadraatsom op zijn kleinst. Daarom heten a en b kleinste kwadraten schatters. De uit de steekproef geschatte lijn wordt nu

$$y_i = a + bx_i + e_i$$

Als we geen rekening zouden houden met de borstomvang (de x), dan zouden we de gewichten het beste kunnen samenvatten met het gemiddelde (zie figuur (2.2)). Voor een waarneming y_i zitten we er dan $(y_i - \bar{y})$ naast, de blauwe lijn in figuur 2.2. Deze afwijking heet de totale afwijking. Als we deze afwijkingen kwadrateren en sommeren, dan krijgen we een kwadraatsom:

$$SS_{Totaal} = \sum_i (y_i - \bar{y})^2.$$

Deze heet de totale kwadraatsom SS_{tot}. Delen we deze door $n-1$ dan krijgen we de variantie van de gewichten. Die $n-1$ zijn de vrijheidsgraden.

Als we wel rekening houden met de borstomvang dan kunnen we de gewichten het best samenvatten met een punt op de lijn. Voor een bepaald gewicht y_i zitten we er dan $y_i - (a + bx_i)$, het residu stukje naast. Deze residuen gekwadrateerd en gesommeerd geeft de residukwadraatsom: $SS_{res} = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$. Hier horen ook vrijheidsgraden bij. Voor dat we de residukwadraatsom kunnen uitrekenen moeten we eerst de lijn bepalen. Dan moeten we twee dingen weten nl. a en b. Dat gaat dus ten koste van 2 informatieve waarnemingen. Dus houden we er nog $n - 2$ over en dat is het aantal vrijheidsgraden voor de residukwadraatsom. Als de residukwadraatsom door het aantal vrijheidsgraden wordt gedeeld dan krijgt men de residuvariantie. In een regressieanalyse geeft dit aan wat de variantie van de punten om de lijn is.
Als we niet op de borstomvang letten dan zitten we er het totale stuk naast, letten we wel op de borstomvang dan zitten we er het residu stuk naast. In de situatie van figuur (2.2) is het totale stuk groter dan het residustuk dus we zijn er wat mee opgeschoten, namelijk het stuk $(a + bx_i) - \bar{y}$. Deze afwijking heet de door de regressie verklaarde afwijking of ook wel kortweg de regressie afwijking. Deze afwijkingen gekwadrateerd en gesommeerd geeft ook een kwadraatsom namelijk de regressie kwadraatsom:

$$SS_{reg} = \sum_{i-1}^{n}((a + bx_i) - \bar{y})^2$$

. Deze heeft 1 vrijheidsgraad. Wat voor de afwijkingen uit figuur(2.2) geldt namelijk Totale afwijking is gelijk aan de regressie afwijking plus de residu afwijking, dat geldt ook voor de kwadraatsommen:

$$SS_{Tot} = SS_{reg} + SS_{Res}$$

en ook voor de vrijheidsgraden: $n - 1 = 1 + (n - 2)$ Deel de kwadraatsommen door de vrijheidsgraden en men heeft de varianties ook wel gemiddelde kwadraatsommen genoemd (Mean Square): MS_{tot}, MS_{reg} and MS_{res}. Dan kan alles in een tabel gezet worden die anova tabel genoemd wordt.

Name	SS	df	MS	F
Regression	SS_{reg}	df_{regres}	MS_{reg}	$\frac{MS_{reg}}{MS_{res}}$
Residual	SS_{res}	df_{res}	MS_{res}	
Total	SS_{total}	df_{Total}		

Om de hypothese $H_0 : \beta = 0$ tegen $H_1 : \beta \neq 0$ te toetsen, kan men als volgt redeneren. Als de alternatieve hypothese waar is en de lijn goed bij de punten wolk past, zullen de residuen klein zijn vergeleken met de regressie stukjes. Anders gezegd de residu

variantie – de variantie van de punten om de lijn– zal dan klein zijn in vergelijking met de door de regressie verklaarde variantie. Dus dan zal $F = \frac{MS_{reg}}{MS_{res}}$ groot zijn. Dit getal F zegt hoeveel keer groter de regressie variantie is vergeleken met de residuvariantie. Als de nulhypothese waar is en de lijn helemaal niet bij de puntenwolk past dan zullen de residustukjes groot zijn in vergelijking met de door de regressie verklaarde stukjes, dus zal de residu variantie groot zijn vergeleken met de regressie variantie en dus zal F klein zijn. Dus F is hier de toetsingsgrootheid. Om de p-waarde voor de uitkomst van de toetsingsgrootheid te bepalen gebruikt men dat F een Fisher verdeling heeft met 1 en $n-2$ vrijheidsgraden.
Voor het schapen onderzoek wordt de anova tabel (zie blz. 136):

Name	SS	df	MS	F
regression	3972.930	1	3972.930	562.133
Residual	452.342	64	7.068	
Total	4425.272	65		

De uitkomst van de toetsingsgrootheid is dus 526.113, de door de regressie verklaarde variantie is 562 maal zo groot als de residu variantie. Om de p-waarde te bepalen gebruikt men de Fischer verdeling met 1 en 64 vrijheidsgraden en vindt dan (afgerond) 0.000.

2.3.2 t-toets

Om de hypothese $H_0 : \beta = 0$ tegen $H_1 : \beta \neq 0$ te toetsen, kan men ook een t-toets gebruiken. De toetsingsgrootheid van de t-toets bepaalde de afstand tussen datgene wat je vond in je onderzoek (b) en de nulhypothese ($\beta = 0$) uitgedrukt in standard errors: $t = \frac{b-0}{se(b)}$. Om dit uit te kunnen rekenen moeten we weten hoe die standard error uit gerekend wordt :

$$se(b) = \sqrt{\frac{MS_{res}}{(n-1)s_x^2}}$$

waarin s_x^2 de variatie van de x-en (de borstomvang) is. Deze t-toetsingsgrootheid heeft een student verdeling met $n-2$ vrijheidsgraden.
In het geval het eerste aantal vrijheidsgraden voor de noemer van de F-toetsingsgrootheid 1 is, geldt dat er een verband is tussen de F-toetsingsgrootheid en de t-toetsingsgrootheid: $t = \sqrt{F}$

2.3.3 r^2

Met de variantieanalyse voor de regressie wordt bekeken of er sprake is van een lineair verband tussen de afhankelijke en de onafhankelijke variabele. Als de nulhypothese wordt verworpen dan is een lineair verband aangetoond dat wil zeggen er is aangetoond dat de regressiecoéfficient geen nul is. Er is dan dus sprake van een rechte lijn. Maar hoe goed past de lijn bij de punten. Het kan een lijn zijn waarvoor geldt dat de punten er strak omheen liggen (past goed) of een lijn waarvoor geldt dat de punten er ruimer om heen liggen (past minder goed). Hier wil men een maat voor hebben. Als de lijn goed past dan vormen de residuen stukjes een klein gedeelte van de totale stukken, zie 2.2. In dat geval vormen de door de regressie verklaarde stukjes een groot deel van de totale stukjes. De door de regressie verklaarde kwadraatsom maakt een groot deel uit van de totale kwadraatsom. In dat geval zal de verhouding $\frac{SS_{reg}}{SS_{totaal}}$ groot zijn, en dicht bij 1 liggen. Als de lijn minder bij de punten past dan zullen de residuen groot zijn t.o.v het totaal en zullen de door de regressie verklaarde stukjes een klein gedeelte vormen van het totaal. De door de regressie verklaarde kwadraatsom maakt maar een klein deel uit van de totale kwadraatsom. Dus de verhouding $\frac{SS_{reg}}{SS_{totaal}}$ zal klein zijn, dicht bij nul liggen. Dus is $\frac{SS_{reg}}{SS_{totaal}}$ een maat voor hoe strak de punten om de lijn liggen. Ligt dit getal dicht bij 1 dan liggen de punten strak om de lijn, ligt dit getal dicht bij nul dan liggen de punten ruim om de lijn. Als nu de formules voor de regressie en totale kwadraatsommen gebruikt worden en de formule voor de richtingscoëfficient wordt gebruikt in de formule voor de regressie kwadraat som, dan blijkt dat $\frac{SS_{reg}}{SS_{totaal}} = r^2$. Dus de correlatiecoëficient in het kwadraat geeft aan welk gedeelte de regressiekwadraatsom uitmaakt van de totale kwadraatsom. Grofweg zegt men wel: r^2 geeft de fractie verklaarde variatie van de totale variatie.

Hoofdstuk 3

Het lineaire model 2: Anova

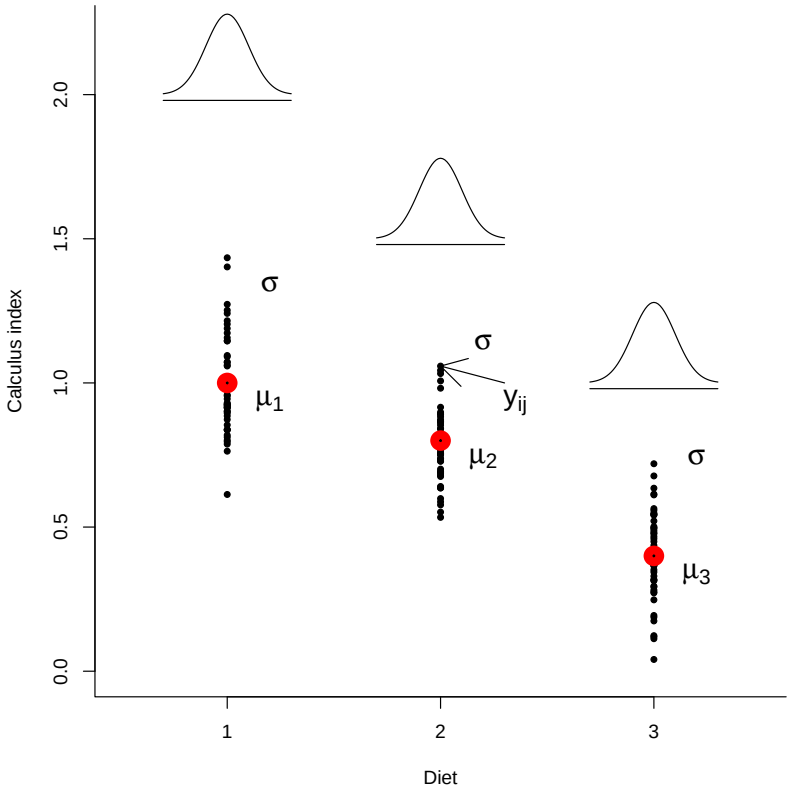
3.1 Een onderzoek

In hoofdstuk 8.6.4 wordt een onderzoek besproken: honden krijgen verschillende diëten voer, "bekleed"met verschillende stoffen die de opbouw van tandsteen zouden moeten beïnvloeden. Er wordt een "tandsteen opeenhopings index"gemeten die calculus index wordt genoemd. Er is een aselecte steekproef genomen van 26 honden die in 3 groepen verdeeld is:

	aantal waarnemingen	Gemiddelde
Controle groep	$n_1 = 9$	1.09
P_2O_7	$n_2 = 9$	0.75
HMP	$n_3 = 8$	0.44

De honden worden over de groepen verdeeld door middel van "randomisatie"dat wil zeggen de honden worden over de groepen verdeeld door middel van loting. Men zegt dan dat de behandelingsgroepen gerandomiseerd zijn. Een onderzoek waarbij de behandelingen gerandomiseerd zijn noemt men een gerandomiseerde clinical trial.

We hebben dus te maken met een continue variabele en een groepsindeling. De onderzoeksvraag is of de diëten de opbouw van tandsteen beïnvloeden ofwel is er verschil tussen de groepen voor wat betreft de gemiddelde tandsteen index.



Figuur 3.1: De populatie

3.2 De populatie: een model voor het data genererend proces

We hebben een aselecte steekproef uit een populatie. Als we een beschrijving zouden hebben van die populatie dan zouden we de data uit de steekproef beter begrijpen. Zo'n beschrijving heet een model. Dus eerst geven we een beschrijving van de populatie die de data van de steekproef gegenereerd heeft. Dit model is natuurlijk slechts een benadering van het vaak erg gecompliceerde data genererend proces.

De populatie is de groep individuen waar de steekproef aselect uit is genomen. Maar dat is niet alles. Het is de populatie die je zou hebben gekregen als het experiment uitgevoerd in de steekproef, in de hele populatie gedaan zou zijn. Dus alle honden zouden verloot worden over 3 groepen: een controle groep, een P_2O_7 groep en een HMP groep. Daarna zouden we van alle honden in de populatie de variabele meten waar we in geïnteresseerd zijn. Het algemeen gemiddelde in de populatie is μ. Het populatie gemiddelde in de eerste groep is μ_1, in de tweede groep is dat μ_2 etc, (zie figuur(3.1)). In het algemeen is het gemiddelde van groep i μ_i. De variantie in iedere groep is σ^2. Tot slot, maken we van alle getallen in de populatie per groep een histogram, dan vinden we iets dat veel lijkt op een normale verdeling. Zo'n populatie bestaat niet echt! Er bestaat geen populatie honden die we over 3 groepen gerandomiseerd hebben en welke groepen we verschillende soorten voer gaven. Dat deden we alleen in de steekproef. Dus deze beschrijving van de populatie is het model voor het data genererend mechanisme, voor het mechanisme dat de data in de steekproef gegenereerd heeft.

De hypotheses die bij de onderzoeksvraag horen zijn nu $H_0 : \mu_1 = \mu_2 = \mu_3$ en de alternatieve hypothese zegt dat minstens twee groepsgemiddelden niet gelijk zijn aan elkaar. We willen dus weten of de groepsgemiddelden van elkaar verschillen.

In het geval van de lineaire regressie beschreven we een model waarbij de groepsgemiddelden uit figuur(3.1) op een rechte lijn liggen. We hadden een model voor de gemiddelden namelijk dat die gemiddelden op een rechte lijn liggen. In dit geval beschrijven we het model rechtstreeks met de groepsgemiddelden. We spreken dan van een variantie analyse model of anova model. Ook dit model is niet exact, er zijn residuen. Als de waarnemingen uit groep 1 beschreven worden met hun gemiddelde (μ_1) dan beschrijven we niet precies de waarnemingen. Sommige waarnemingen liggen boven dat gemiddelde en sommige liggen eronder zoals in figuur(3.1) te zien is. Hoeveel een waarneming boven of onder het groepsgemiddelde ligt heet het residu. Dus een waarneming wordt beschreven als het groepsgemiddelde plus een residu, ofwel

$$y_{ij} = \mu_i + \epsilon_{ij}$$

Waarbij y_{ij} de j^{de} waarneming is uit groep i. Dus, bijvoorbeeld y_{23} is de derde waarneming uit groep twee.

Dit model gebruikt men als de groepen verschillend zijn. Als H_0 waar is dan zijn de groepen gelijk en hoeft er geen rekening gehouden te worden met verschillende gemiddelden. In dat geval is er slecht 1 gemiddelde μ en het model is dan

$$y_{ij} = \mu + \epsilon_{ij}$$

Meestal schrijft men deze twee modellen in een keer als:

$$y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$$

waarin $(\mu_i - \mu)$ de groepseffecten genoemd worden. Als de groepsgemiddelden dicht bij elkaar liggen dan lijken ze veel op elkaar en zullen ze ook erg veel op het algemene gemiddelde, μ, lijken. Het verschil tussen de groepsgemiddelde en het algemeen is erg klein dus de afwijkingen $(\mu_i - \mu)$ zullen dicht bij nul liggen. Dan zegt het model dat je een waarneming het best beschrijft met het algemeen gemiddelde plus een residu. Dit is de situatie van de nulhypothese. In de situatie dat de groepsgemiddelden niet gelijk aan elkaar zijn en dus ver uit elkaar liggen, zal een of meer van die groepsgemiddelden ook niet op het algemene gemiddelde lijken. Dus, is de afwijking $(\mu_i - \mu)$ groot (positief or negatief). Dan zegt het model dat een waarneming het best te beschrijven is met het het groepsgemiddelde plus een residu.

Dus de afwijkingen $(\mu_i - \mu)$ laten zien of er verschillen zijn tussen de groepen en worden daarom groepseffecten genoemd.

Het lineaire model voor de anova is

$$y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$$

met y_{ij} normaal verdeeld met gemiddelde μ_i en variantie σ^2

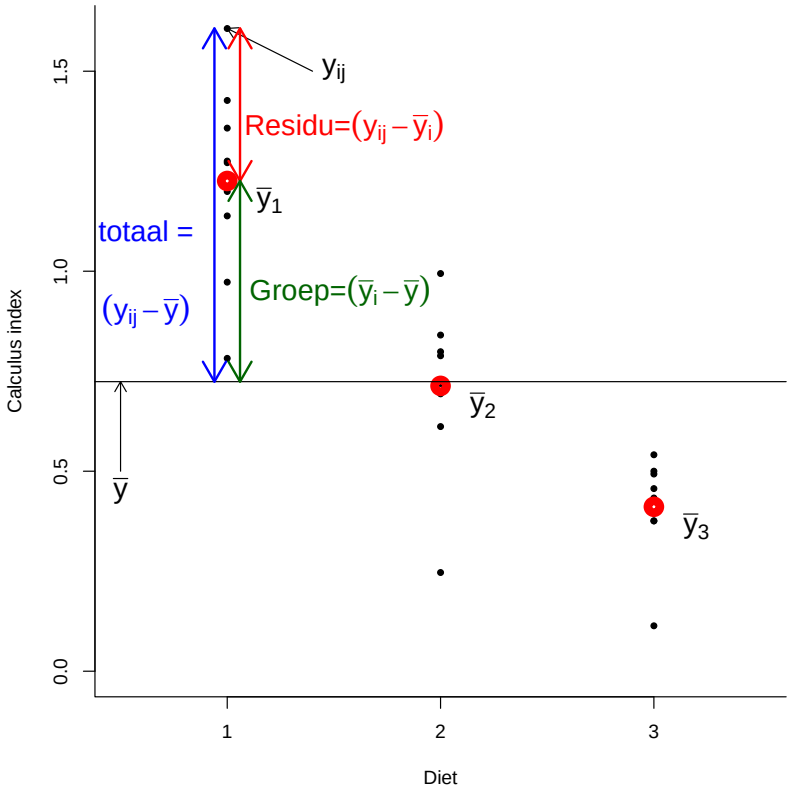
3.3 De steekproef

In de steekproef moeten we het model schatten door de populatie gemiddelden te vervangen door de steekproefgemiddelden

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + e_{ij}$$

waarbij het residu in de steekproef $e_{ij} = y_{ij} - \bar{y}_i$. Schrijf nu het model op in termen van afwijkingen door $\bar{y}$ naar de linkerkant van het = teken te brengen:

$$y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + e_{ij}$$



Figuur 3.2: De steekproef

$(y_{ij} - \bar{y})$ is de afwijking tussen een waarneming en het algemeen gemiddelde. Deze afwijking heet **de totale afwijking**. Zie figuur 3.2.
$(\bar{y}_i - \bar{y})$ is de afwijking tussen het groepsgemiddelde en het algemeen gemiddelde. Deze **afwijkingen tussen groepen** hebben we de groepseffecten genoemd en ze geven aan of er veel of weinig verschillen zijn tussen de groepen.
$e_{ij} = (y_{ij} - \bar{y}_i)$ zijn de afwijkingen tussen een waarneming en z'n groepsgemiddelde en zijn dus **de residuele afwijkingen**. Dus

$$Totale\ afwijking = afwijking\ tussen\ groepen + residu\ afwijking$$

Kwadrateer nu de afwijkingen en tel ze op over alle waarnemingen:

$$\sum_{alle\ waarnemingen} (y_{ij} - \bar{y})^2 = \sum_{all\ waarnemingen} (\bar{y}_i - \bar{y})^2 + \sum_{all\ waarnemingen} e_{ij}^2$$

De afwijkingen gekwadrateerd en gesommeerd heten kwadraatsommen. De totale afwijking gekwadrateerd en gesommeerd heet de totale kwadraatsom SS_{Totaal}, de tussen groepen afwijking gekwadrateerd en gesommeerd heet de tussen groepen kwadraatsom SS_{Group} en de residuele afwijking gekwadrateerd en gesommeerd heet de residu kwadraatsom SS_{Res}. Dus

$$SS_{Totaal} = SS_{Groep} + SS_{Res}$$

Kwadraatsommen zijn gebaseerd op een aantal informatieve waarnemingen ook wel vrijheidsgraden (degrees of freedom) genoemd. De totale kwadraatsom heeft $df_{totaal} = n - 1$ vrijheidsgraden. Wordt de totale kwadraatsom gedeeld door het aantal vrijheidsgraden dan krijgt men de totale variantie, de variantie van de n getallen in het onderzoek.

De totale kwadraatsom hangt af van de waarnemingen min het algemeen gemiddelde $(y_{ij} - \bar{y})$. Voordat dit uitgerekend kan worden moet het gemiddelde bepaald worden. Dat wordt "betaald"met 1 informatieve waarneming (zie de teksten van lijn 1). We houden dan nog $n - 1$ informatieve waarnemingen over. Ofwel de vrijheidsgraden zijn het aantal waarnemingen min een. De groepen kwadraatsom hangt af van de groepsgemiddelden min het algemeen gemiddelde $(\bar{y}_i - \bar{y})$; het aantal vrijheidsgraden is dan het aantal groepen min een: $df_{Groep} = aantal\ groepen - 1$. Wat voor de afwijkingen en de kwadraatsommen geldt, geldt ook voor het aantal vrijheidsgraden namelijk totaal = tussen groepen + residu. Het aantal vrijheidsgraden voor de residu kwadraatsom is dan gelijk aan het totaal aantal vrijheidsgraden min dat aantal voor de groepen kwadraatsom: $df_{residu} = df_{Totaal} - df_{groep}$.

Stel er zijn geen verschillen tussen de groepen dus $\mu_1 = \mu_2 = \mu_3$. De waarnemingen binnen een groep zijn niet allemaal gelijk aan elkaar; ze verschillen. Het data genererend mechanisme is kennelijk zodanig dat de waarnemingen binnen een groep van elkaar verschillen. Als er geen systematische verschillen zijn tussen de groepen (populatiegemiddelden zijn gelijk) dan zullen dit soort verschillen ook tussen de groepen zitten. De steekproefgemiddelden zullen nooit precies gelijk zijn aan elkaar. Het data genererend mechanisme dat zorgt voor de verschillen tussen de waarnemingen binnen een groep zorgt dan ook voor verschillen tussen de groepen m.a.w de variantie tussen de groepen is gelijk aan de residu variantie (=de variantie tussen de waarnemingen binnen de groepen) ofwel $F = \frac{MS_{Group}}{MS_{res}} \approx 1$. Als de groepsgemiddelden veel van elkaar verschillen dan zal de tussen groepen variantie veel groter zijn dan de residu variantie.
Om de nulhypothese $H_0 : \mu_1 = \mu_2 = \mu_3$ te testen tegen het alternatief dat minstens twee groepsgemiddelden niet gelijk zijn aan elkaar, kan men gebruiken dat de toetsingsgrootheid F een zogenaamde Fisher verdeling heeft met df_{group} and df_{res} als vrijheidsgraden. Dit kan men gebruiken om p-waardes te berekenen. Men kan tot slot de resultaten van deze berekeningen in een tabel zetten, de zogenaamde anova tabel.
Dus samenvattend:

1. Schrijf het model in de populatie op: $y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$

2. Schat het model in de steekproef: $y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + e_{ij}$

3. Schrijf het model in termen van afwijkingen: $y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + e_{ij}$

4. Kwadrateer de afwijkingen en sommeer ze over alle waarnemingen om de kwadraatsommen te krijgen : $SS_{Totaal} = SS_{Group} + SS_{Res}$

5. Deel de kwadraatsommen door de vrijheidsgraden om varianties te krijgen MS_{Totaal}, MS_{Group} and MS_{res}. Bepaal F: hoeveel keer groter is de variantie tussen de groepen vergeleken met de residuvariatie.

6. Zet alles in een anova tabel:

Naam	SS	df	MS	F
groeps	SS_{group}	df_{group}	MS_{group}	$\frac{MS_{Group}}{MS_{res}}$
Residu	SS_{res}	df_{res}	MS_{res}	
Totaal	SS_{total}	df_{Totaal}		

Dus het begint met het lineaire model in de populatie, dit wordt geschat in de steekproef, dan wordt het geschatte model geschreven in termen van afwijkingen en hieruit worden de kwadraat sommen verkregen en dus uiteindelijk de anova tabel. De anova tabel volgt dus uit het gebruik van het lineaire model. De anova-tabel zegt dus wat over dat lineaire model: welk model is volgen de geobserveerde data het beste? Is dat het model met de groepseffecten ($y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$) of kan de data net zo goed beschreven worden met een gemiddelde en een residu ($y_{ij} = \mu + \epsilon_{ij}$)
Voor het honden voorbeeld wordt de anova tabel:

Naam	SS	df	MS	F
groep	1.805	2	.902	6.668
Residu	3.112	23	.1358	
Totaal	4.917	25		

Dus de variantie tussen de groepen is 6.7 keer groter dan de residu variantie. De p-waarde die daarbij hoort is 0.005. Dus volgens de data is het lineaire model met groepseffecten erin beter dan een model zonder.

Hoofdstuk 4

Lineaire modellen 3: Uitbreidingen

4.1 Lineaire regressie met meerdere onafhankelijk continue variabelen

Vaak zijn er in een onderzoek meer dan 1 onafhankelijke variabelen. Op blz. 153 van het boek, staat een beschrijving van een onderzoek onder Marokkaanse werkezels. Omdat weegschalen zeldzaam zijn, zochten ze een andere methode om het lichaamsgewicht te bepalen. Lichaamsgewicht is dus weer de continue afhankelijke variabele. Maar in plaats van 1 onafhankelijke variabele hebben ze er nu 6, zie blz 153 en 154.

Het lineaire model voor de populatie, waaruit de steekproef afkomstig is, is een uitbreiding van het geval met 1 onafhankelijke variabele:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_6 x_{6i} + \varepsilon_i$$

waarbij de lichaamsgewichten y_i normaal verdeeld zijn met variantie σ^2.
Om dit model te krijgen worden de onafhankelijke variabelen met een β vermenigvuldigd en bij elkaar opgeteld (optellen, vandaar een lineair model). Dit model lijkt dus erg veel op het model met maar één variabele.
β_1 stelt nu de verandering voor in de gemiddelde y-variabele als x_1 met 1 verandert terwijl de andere variabelen constant gehouden worden. Dat is een voordeel van dit uitgebreide model. Als bijvoorbeeld x_2 niet in het model zit, dan weet je niet wat deze doet als x_1 verandert. Misschien verandert x_2 wel mee en schat je dus niet het echte effect van x_1 maar het 'verstrengelde effect van x_1 en x_2 samen. Als ze alle twee in het model zitten heb je daar geen last van want dan kan je er voor zorgen dat de een constant blijft als de ander verandert.

In de steekproef noemen we het intercept a en de richtingscoëfficiënten noemen

we b_1, b_2 etc. Het schatten van de parameters verloopt nagenoeg hetzelfde als in het geval dat er maar 1 onafhankelijke variabele is: men wil de residuen zo klein mogelijk hebben. Een residu is nu de afstand van een punt tot het (hyper)vlak dus $y_i - (a + b_1x_{1i} + b_2x_{2i} + \cdots + b_6x_{6i})$. De residukwadraatsom is dan

$$SS_{res} = \sum_i \left[y_i - (a + b_1x_{1i} + b_2x_{2i} + \cdots + b_6x_{6i})\right]^2$$

Dit geeft de waarden voor a en de b's waarvoor de residu kwadraatsom op z'n kleinst is.

De totale kwadraatsom wordt op dezelfde wijze uitgerekend als in het geval van 1 onafhankelijke variabele ($\sum_i (y_i - \bar{y})^2$) en omdat

$$SS_{Total} = SS_{regres} + SS_{Res}$$

kan de regressiekwadraatsom nu bepaald worden door het verschil te nemen van de totale en de residu kwadraatsom. De regressiekwadraatsom gaat in dit geval over 6 onafhankelijke variabelen dus is gebaseerd op 6 vrijheidsgraden. Het totaal aantal vrijheidsgraden is $n-1$ zodat het aantal vrijheidsgraden voor het residu $(n-1)-6 = n-7$ is. Nu kan weer de anova tabel gemaakt worden, zie blz. 147. De F-toets in deze anova tabel toetst in één keer of er een verband is tussen de afhankelijke en de 6 onafhankelijke variabelen. De nulhypothese luidt dus $H_0 : \beta_1 = \beta_2 = \cdots = \beta_6 = 0$ De alternatieve hypothese luidt dat er minstens 1 β ongelijk aan nul is.

Net als in het geval dat er sprake was van 1 onafhankelijke variabele, kunnen we ook hier voor iedere onafhankelijke variabele een t-toets gebruiken. Deze t-toets toetst of de onafhankelijke variabele kan worden gebruikt om de afhankelijke variabele lineair te beschrijven, **gegeven dat de andere variabelen in het model dat ook al doen**. Zo kan bijvoorbeeld de nulhypothese $H_0 : \beta_1 = 0$ worden getoetst. De t-toets is weer een soort afstandsmaat: het is de afstand tussen wat je vindt in het onderzoek (b_1) met de nulhypothese $\beta_1 = 0$ uitgedrukt in standard errors ofwel $t = \frac{b_1 - 0}{se(b_1)}$. Natuurlijk kunnen we in plaats van een t-toets ook de F-toets per variabele uitrekenen want de F-toets is in het geval dat er 1 variabele getoetst aan de t-toets in het kwadraat.

Als er niet te veel onafhankelijke variabelen zijn, dan kan de volgende procedure gebruikt worden: Kijk naar de tabel met de t-toetsen en vind de minst significante. Dat is dus die variabele met de grootste p-waarde, groter dan de onbetrouwbaarheid α. Laat deze uit het model dat wil zeggen: doe de analyse overnieuw zonder deze variabele. Kijk dan weer naar de tabel van de t-toetsen en laat de minst significante uit het model, enzovoort. Dit gaat zo door tot er allen nog onafhankelijke variabelen in het model zitten die volgens de t-toetsen significant zijn.

4.2 Anova met meer dan 1 factor

Bij het werkcollege over variantie-analyse, keken we naar de verschillen tussen groepsgemiddelden. Het voorbeeld ging daar over honden die verschillende diëten kregen, "bekleed"met verschillende stoffen die de opbouw van tandsteen zouden moeten beïnvloeden. Er wordt een "tandsteen-opeenhopings-index"gemeten. Het lineaire model voor deze situatie was:

$$y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$$

waarin $(\mu_i - \mu)$ de groeps effecten genoemd worden.

De groepsindeling zoals boven gebruikt noemt men ook wel een factor (of een categorische variabele). Deze factor geeft dus aan welke behandeling de hond gehad heeft en voor iedere hond is deze factor bekend. Het gaat daarbij dus om de factor behandeling met 3 niveaus. Buiten deze factor kan men ook nog geïnteresseerd zijn in een ander factor bijvoorbeeld ras. Stel men wil, naast de behandeling, ook de invloed van twee rassen bekijken. De bovenstaande waarnemingen zouden dan voor ras A kunnen zijn en men zou zo ook waarnemingen van ras B kunnen hebben verdeeld over 3 dieet groepen.

Een waarneming kan nu worden aangegeven met y_{ijk}. Dit is de k^{de} waarneming uit behandelingsgroep i en ras j. De gemiddelden voor de behandelingsgroepen worden aangegeven met μ_i en de gemiddelden voor de rasgroepen met μ_j.
De groepseffecten voor de behandeling zijn $(\mu_i - \mu)$, precies als in het geval van 1 factor. Deze groepseffecten worden kortweg behandelingseffecten genoemd. De groepseffecten voor de rassen zijn nu $(\mu_j - \mu)$, kortweg raseffecten genoemd. De interpretatie is ook hetzelfde: Als de groepsgemiddelden voor de rassen dicht bij elkaar liggen dan lijken ze veel op elkaar en zullen ze ook erg veel op het algemene gemiddelde lijken. Het verschil tussen de groepsgemiddelde en het algemeen gemiddelde is erg klein dus de afwijkingen $(\mu_j - \mu)$ zullen dicht bij nul liggen. Dit is de situatie van de nulhypothese die zegt dat er geen verschillen zijn tussen de ras gemiddelden in de populatie. In de situatie dat de groepsgemiddelden niet gelijk aan elkaar zijn en dus ver uit elkaar zullen liggen, zal een of meer van die groepsgemiddelden ook niet op het algemene gemiddelde lijken. Dus is de afwijking $(\mu_j - \mu)$ groot (positief of negatief).

Het lineaire model is het lineaire model met 1 factor waarbij we nu de ras effecten optellen. De analyse verloopt daarna nagenoeg hetzelfde als voor het geval er 1 factor aanwezig is.

1. Schrijf het model in de populatie op: $y_{ijk} = \mu + (\mu_i - \mu) + (\mu_j - \mu) + \epsilon_{ijk}$. De y_{ijk} zijn normaal verdeeld met gemiddelde $\mu + (\mu_i - \mu) + (\mu_j - \mu)$ en variantie σ^2

2. Schat het model in de steekproef: $y_{ijk} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + e_{ijk}$

3. Schrijf het model in termen van afwijkingen: $y_{ijk} - \bar{y} = (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + e_{ijk}$. De totale afwijking, $y_{ijk} - \bar{y}$, is nu op te splitsen in 3 stukken: een tussen behandelinsgroepen afwijking $(\bar{y}_i - \bar{y})$, een tussen rasgroepen afwijking $(\bar{y}_j - \bar{y})$ en een residu afwijking e_{ijk}.

4. Kwadrateer de afwijkingen en sommeer ze over alle waarnemingen om de kwadraatsommen te krijgen: $SS_{Total} = SS_{behandeling} + SS_{ras} + SS_{Res}$

5. Deel de kwadraatsommen door de vrijheidsgraden om varianties te krijgen. De vrijheidsgraden voor de behandelingskwadraatsom is aantal behandelingsgroepen min een, en voor de ras kwadraatsom wordt dit: aantal rasgroepen min een. Bepaal nu de toetsingsgrootheid F: hoeveel keer groter is de variantie tussen de groepen vergeleken met de residuvariatie.

6. Zet nu het een en ander in een anova tabel.

Name	SS	df	MS	F
Behandeling	$SS_{behandeling}$	$df_{behandeling}$	$MS_{behandeling}$	$\frac{MS_{behandeling}}{MS_{residu}}$
Ras	SS_{ras}	df_{ras}	MS_{ras}	$\frac{MS_{ras}}{MS_{residu}}$
Residual	SS_{residu}	df_{residu}	MS_{residu}	
Total	SS_{totaal}	df_{Totaal}		

Om te toetsen of er verschil is tussen populatiegemiddelden van de behandelingsgroepen gebruikt men als toetsingsgrootheid $F = \frac{MS_{behandeling}}{MS_{residu}}$ Deze heeft een Fisher verdeling met $df_{behandeling}$ en df_{residu} vrijheidsgraden.
Om te toetsen of er verschil is tussen populatiegemiddelden van de rassen gebruikt men als toetsingsgrootheid $F = \frac{MS_{ras}}{MS_{residu}}$. Deze heeft een Fisher verdeling met df_{ras} en df_{residu} vrijheidsgraden.

Een en ander kan geheel analoog worden uitgebreid naar gevallen met 3 of meer factoren

4.3 Lineaire regressie met indicator variabelen

Stel dat in het voorbeeld uit hoofdstuk 8.6.4, honden slechts twee verschillende di¨eten kregen, een controle dieet en voer bekleed met $P2O7$. Vervolgens werd de calculusindex ("tandsteen-opeenhopings-index") gemeten. De controle groep bestaat uit 9 honden en de $P2O7$ groep uit 9. De data set zou er als volgt uit kunnen zien:

Groep	tandsteen index
1	0.49
1	1.05
2	0.53
ect	etc

De controle-groep is aangegeven met een 1 en de $P2O7$-groep met 2. Vervolgens kunnen we de analyse doen met een lineair model voor een categoriale variabele en kunnen we een anova tabel of een t-toets krijgen.
Er kan van die groepsvariabele ook een zogenaamde indicator variabele gemaakt worden: De controle groep wordt aangegeven met een nul en de behandelingsgroep $P2O7$ met een 1. Deze "nul-een-variabele wijst de behandelingsgroep aan: een behandelingsgroep indicator. Als we nu de calculus-index zouden analyseren met een gewone regressie waarin we doen alsof behandelingsindicator de continue onafhankelijke variabele is, wat zou er dan gebeuren? Noem die behandelingsindicator $G2$ omdat het een indicator is van groep 2, de $P2O7$ groep. Het regressie model is dan $y_i = \alpha + \beta G2_i + \epsilon_i$. Voor alle waarnemingen uit de controle groep geldt dat $G2 = 0$ dus wordt het model $y_i = \alpha + \epsilon_i$. Alle 9 waarnemingen uit de controle groep worden met een constante α beschreven. De beste schatter voor die constante is dan het gemiddelde van de controle groep zeg $\bar{y}_1$. Voor de waarnemingen uit de $P2O7$ groep geldt het model $y_i = \alpha + \beta + \epsilon_i$, want $G2$ is 1 voor die waarnemingen. Alle 9 waarnemingen uit de $P2O7$ worden met een constante $\alpha + \beta$ beschreven. De beste schatter voor die constante is dan het gemiddelde van de controle groep, zeg $\bar{y}_2$. Dus de schatter voor α is $\bar{y}_1$, de beste schatter voor $\alpha + \beta$ is $\bar{y}_2$. Dat betekent dat $\bar{y}_2 - \bar{y}_1$ een schatter voor β is. Dus α is het gemiddelde van groep1 en wordt geschat met $\bar{y}_1$, β is het verschil in gemiddelde tussen groep 2 en 1 en wordt geschat met $\bar{y}_2 - \bar{y}_1$. Het verschil tussen de gemiddelden van groep 1 en 2, β, is te beschouwen als het groepseffect. Dit is dus een andere manier om een groepseffect te definiëren dan we bij de variantie analyse gedaan hebben. Daar definieerden we een groepseffect als een groepsgemiddelde min het algemeen gemiddelde $(\mu_i - \mu)$.
In het geval van een categoriale variabele met 3 niveau's (3 groepen) werkt het op dezelfde manier:

Groep	G2	G3
1	0	0
2	1	0
3	0	1

Als er 3 groepen zijn dan maken we er nog een indicator bij: de groep 3 indicator. Deze heeft de waarde 1 voor iedere waarneming uit groep 3 en gelijk aan nul voor

de waarnemingen uit groep 1 en 2. Het lineaire model wordt nu: $y_i = \alpha + \beta_1 G2_i + \beta_2 G3_i + \epsilon_i$. Voor alle waarnemingen uit de controlegroep geldt dat $G2 = 0$ en $G3 = 0$ dus wordt het model $y_i = \alpha + \epsilon_i$. De beste schatter voor de constante α is weer het gemiddelde van de controlegroep zeg $\bar{y}_1$. Voor de waarnemingen uit de $P2O7$ groep geldt het model $y_i = \alpha + \beta_1 + \epsilon_i$, want $G2$ is 1 voor die waarneming en $G3 = 0$. Dus alle 9 waarnemingen uit de $P2O7$ worden met een constante $\alpha + \beta_1$ beschreven. De beste schatter voor die constante is dan het gemiddelde van de $P2O7$ groep zeg $\bar{y}_2$. Dat betekent $\bar{y}_2 - \bar{y}_1$, het verschil in groepsgemiddelden tussen groep 2 en q, een schatter voor β_1 is.
Volgens eenzelfde redenering geldt dat de beste schatter voor β_2 het verschil $\bar{y}_3 - \bar{y}_1$ is.

Voor een lineair model met een categoriale variabele met 3 niveaus worden twee indicator variabelen gebruikt voor een model met een categoriale variabele met 4 niveaus worden 3 indicator variabelen gemaakt etc. In het model met twee indicator variabelen is α het gemiddelde van de groep die voor alle indicatoren een nul heeft, die niet geïndiceerd is de zogenaamde referentiegroep; β_1 is het verschil in gemiddelde tussen groep 2 en de groep die niet geïndiceerd is en β_2 is het verschil in gemiddelde tussen groep 3 en de groep die niet geïndiceerd is.
Deze 0-1 codering wordt vaak gebruikt in statistische programmatuur. Vandaar dat je in de uitvoer bij een analyse van bijv. 3 groepen, voor die groepsvariabele maar 2 regels ziet; iedere regel bevat het verschil in groepsgemiddelden vergeleken met de eerste groep, de niet geïndiceerde groep (= de referentie groep). Voor het voorbeeld uit hoofdstuk 8.6.4 zou de volgende uitvoer gegeven kunnen worden, waarbij groep 1 de controle-groep is, groep 2 de P_2O_7-groep en groep 3 de HMP-groep is

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0889     0.1226   8.881 6.83e-09 ***
groep2       -0.3422     0.1734  -1.974  0.06056 .
groep3       -0.6514     0.1787  -3.644  0.00135 **
---
```

Dus het gemiddelde in de controle groep is 1.0889; het verschil in gemiddelden tussen de P_2O_7-groep en de controle-groep is -0.3422 en het verschil in gemiddelden tussen de HMP-groep en de controle-groep is -0.6514

Als er meerdere categorische variabelen zijn dan is het intercept het gemiddelde van de groep waarbij alle indicatoren nul zijn.

Hoofdstuk 5

Logistische regressie

5.1 Introductie

Soms meet men een afhankelijke variabele die de aanwezigheid van een kenmerk aangeeft. Denk bijvoorbeeld aan de aanwezigheid van een ziekte bij een individu. Deze afhankelijke variabele kan dus als een indicator variabele gezien worden: de variabele heeft een 1 voor een ziek individu en een nul voor een niet ziek individu, dus de variabele wijst een ziek individu aan. Op bladzijde 156 en verder van het boek wordt een voorbeeld besproken. Bij 1000 keizersnedes bij koeien onder veldcondities werd er gemeten of er complicaties optraden. Dit is dus een 0-1 variabele (een indicator): een koe met keizersnede heeft een 1 als uitkomst als er een complicatie is en een 0 als dat niet zo is. Deze afhankelijke variabele wil men relateren aan een of meer andere variabelen: de onafhankelijke variabelen. Een van de variabelen uit het voorbeeld is het type van de koe: melk of vlees. Deze kan ook als een 0-1 variabele gecodeerd worden: een melkkoe krijgt een 0 en een vleeskoe een 1. Deze variabele kan dus ook gezien worden als een indicator.

5.2 De populatie

Men kan deze 1000 koeien opvatten als een steekproef uit een populatie (de populatie van alle koeien waarbij een keizersnee wordt uitgevoerd). Stel dat we de hele populatie konden bekijken. Deze populatie kan worden verdeeld in twee groepen: de melkkoeien en de vleeskoeien. We zouden van alle koeien kunnen bepalen of er complicaties ontstaan. De fractie vleeskoeien waarbij dit gebeurd wordt aangegeven met π_1, en bij de melkkoeien met π_0. Dus als $\pi_1 = 0.3$ dan betekent dat dat 30% van de

vleeskoeien met een keizersnede complicaties heeft gekregen. Deze beschrijving van de populatie vormt het model waarvan we aannemen dat het de waarnemingen uit het onderzoek gegenereerd heeft. Dit model is een erg eenvoudige weergave van het data genererend proces.

De vraag is nu of het type van de koe gerelateerd is aan het voorkomen van complicaties. Komen er bij de vleeskoeien vaker of minder vaak complicaties voor vergeleken met de melkkoeien. Deze vraag kan men beantwoorden met de odds ratio. In plaats van naar een fractie te kijken– het aantal gevallen op het totaal– kan men kijken naar de odds: $\frac{\pi_1}{1-\pi_1}$, het aantal koeien met complicaties per koe zonder complicatie.

Stel, als voorbeeld, dat de fractie complicaties $\frac{1}{3}$, dan is de odds $\frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$. Dus het aantal koeien met complicatie per koe zonder complicatie is een half. Ofwel voor iedere koe met een complicatie zijn er twee zonder. De verhouding koe met complicatie: koe zonder complicatie: $\pi_1 : 1 - \pi_1$ is $1 : 2$.

Men kan de odds ook uitrekenen voor de melkkoeien: $\frac{\pi_0}{1-\pi_0}$, De verhouding koeien met complicatie: koe zonder complicatie bij de melkkoeien. De odds ratio is nu verhouding van de odds van beide groepen: $\omega = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}$. Komt hier bijvoorbeeld 3 uit dan is de verhouding koeien met een complicatie versus koeien zonder complicatie, bij de vleeskoeien 3 keer groter vergeleken met de melkkoeien. Ofwel het aantal koeien met complicaties per koe zonder complicatie is bij de vleeskoeien 3 maal groter dan bij de melkkoeien. De odds ratio is dus een maat voor de samenhang tussen soort koe en complicaties.

5.3 De steekproef

We hebben een steekproef van 1000 koeien. Deze kunnen we ook in twee groepen verdelen: de vleeskoeien en de melkkoeien. Stel er zijn $(c+d)$ vleeskoeien (zie tabel 5.1)in de steekproef waarvan er (d) complicaties hebben. De steekproeffractie vleeskoeien met complicaties is dan: $P_1 = \frac{d}{c+d}$. Als we een goede aselecte steekproef hebben dan zal deze steekproeffractie veel lijken op de populatiefractie π_1. Hetzelfde geldt voor de groep melkkoeien: stel er zijn $a + b$ melkkoeien waarvan b met complicaties. De steekproeffractie melkkoeien met complicaties is dan $P_0 = \frac{b}{a+b}$. Deze P_1 en P_0 zijn schattingen voor de populatiefracties π_1 en π_0. Met behulp hiervan kan de populatie oddsratio $\omega = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}$ geschat worden door de populatie fracties te vervangen door de

Tabel 5.1: Aantallen en fracties in de steekproef

	Complicatie		
	nee	ja	
melkkoeien	a	b	$P_0 = \frac{b}{a+b}$
vleeskoeien	c	d	$P_1 = \frac{d}{c+d}$

steekproeffracties:

$$OR = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}$$

Als voor P_1 en P_0 respectievelijk $\frac{d}{c+d}$ en $\frac{b}{a+b}$ ingevuld wordt dan geldt $OR = \frac{ad}{bc}$.

5.4 Het logistische regressie model

Het regressie model relateert de afhankelijke variabele aan een onafhankelijke variabel. Het is een model waarbij de gemiddelden van de afhankelijke variabele worden gerelateerd aan de onafhankelijke variabele met een lineair model. Zoiets kan ook als de afhankelijke variabele een "0-1-variabele is. De kans op complicaties wordt dan gerelateerd aan de onafhankelijke variabele (type koe). Het lineaire model: $\pi = \alpha + \beta \cdot type$ werkt niet omdat π tussen 0 en 1 moet liggen maar de uitkomst van $\alpha + \beta \cdot type$ kan alles zijn, van groot positief tot groot negatief. Het lineaire model wordt niet toegepast op π maar op een transformatie van π. De transformatie is de logaritme van de odds (de log-odds): $\ln \frac{\pi}{1-\pi}$.

Het logistische regressie model relateert de logaritme van de odds aan de onafhankelijke variabele op een lineaire manier. In de populatie wordt dat: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot type$ waarin type het type van de koe is: type=1 voor de vleeskoeien en type=0 voor de melkkoeien. Type is hier dus een indicator variabele.

Dit model voor de groep melkkoeien wordt dan: $\ln\left(\frac{\pi_0}{1-\pi_0}\right) = \alpha + \beta \cdot 0 = \alpha$, dus is α de log-odds in de groep waarvoor de indicator variabele nul is (melkkoeien), ofwel, het is de logaritme van de verhouding complicaties- geen complicaties in de groep waarvoor de indicator nul is. Het model in de groep vleeskoeien wordt: $\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha + \beta \cdot 1$. Dus $\alpha + \beta$ is de log-odds in de groep waarvoor de indicator variabele 1 is. Dat betekent dat β het verschil is in log-odds tussen de groep met indicator waarde 1 en

de groep met indicator waarde 0: $\ln\left(\frac{\pi_1}{1-\pi 1}\right) - \ln\left(\frac{\pi_0}{1-\pi_0}\right) = \ln\left(\frac{\frac{\pi_1}{1-\pi 1}}{\frac{\pi_0}{1-\pi_0}}\right) = \ln(\omega) = \beta$. Dus β is de complicatie log-odds ratio voor de vleeskoeien (indicator=1) versus de melkkoeien (indicator=0) en e^β is dan de complicatie odds ratio voor de vleeskoeien versus de melkkoeien.

In de steekproef moeten we het model schatten. Dat doen we weer door de populatiefracties te vervangen door de steekproeffracties. De zo verkregen schatters voor α en β noemen we a en b. Dus $a = \ln\left(\frac{P_0}{1-P_0}\right)$, de logaritme van de verhouding complicaties-geen complicaties in de steekproef, en $b = \ln(OR)$, de logaritme van de steekproef odds ratio.

De onafhankelijke variabele het *type* van de koe is een groep indicator. Deze kan slechts twee waarden aannemen. In dit voorbeeld is er ook een continue onafhankelijke variabele: de hoeveelheid verdoving. Als er een continue onafhankelijke variabele wordt gebruikt dan is het model hetzelfde: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot sedation$. De log-odds wordt nu lineair beschreven met behulp van een continue onafhankelijk variabele. Merk op dat het model rechts van het =-teken een gewoon regressie model is. Als de onafhankelijke variabele uitkomst 0 heeft dan wordt het model in de populatie: $\ln\left(\frac{\pi_0}{1-\pi_0}\right) = \alpha$. Dus α is de log odds voor alle individuen waarvoor de onafhankelijke variabele nul is. Omdat het model rechts van het =-teken een regressie model is, kan β gezien worden als de verandering in datgene wat links van het=-teken staat als de onafhankelijke variabele met 1 verandert, ofwel β is de verandering in log odds als de onafhankelijke variabele met 1 verandert: $\beta = \ln\left(\frac{\pi_1}{1-\pi_1}\right) - \ln\left(\frac{\pi_0}{1-\pi_0}\right) = \ln\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}\right) = \ln(\omega)$. Dus β is de log oddsratio voor het verschil in onafhankelijke variabele van 1 en dus is e^β de odds ratio voor een verandering van 1 in de onafhankelijke variabele. Deze oddsratio geldt over de gehele schaal voor de onafhankelijke variabele, dus als de onafhankelijke variabel bijvoorbeeld verandert van 1 naar 2 of als deze verandert van bijvoorbeeld 8 naar 9 of van 13 naar 14.

Dit model moet geschat worden in de steekproef. Echter nu kunnen we niet meer simpel de populatiefracties vervangen door de steekproeffracties. Er is een andere methode nodig: de maximum likelihood methode. De schatters die we daarmee krijgen noemen we weer a en b.

Als er geen relatie is tussen de onafhankelijke variabele en wel of geen complicaties dan is de oddsratio gelijk aan 1 ofwel de log-oddsratio is gelijk aan nul. De nulhypothese zegt dat er geen relatie is in de populatie: $H_0 : \beta = 0$ De alternatieve hypothese is $H_1 : \beta \neq 0$. De toetsingsgrootheid is in principe hetzelfde als in de

regressie situatie: je kijkt naar de afstand tussen wat je vind in je onderzoek (b) en de hul hypothese ($\beta = 0$) uitgedrukt in standard errors:

$$t = \frac{b - 0}{se(b)}$$

Vaak wordt deze toetsingsgrootheid aangegven met z omdat de verdeling ervan met een standaard normale verdeling benaderd wordt.

Deze toets wordt, in geval van de maximum likelihood methode, de Wald toets genoemd.

Net als met de lineaire regressie kunnen er meer dan 1 onafhankelijke variabele gebruikt worden. In het boek op blz 157 staat een tabel van het voorbeeld met wel of geen complicaties na keizersnedes bij koeien waar 4 onafhankelijke variabelen gebruikt worden. Dit model in de populatie heeft dus 4 beta's. De geschatte beta's met behulp van de steekproef worden aangegeven met b's. In tabel 11.1 staan in de kolom waarboven b_i staat, de schattingen van het model. Dit zijn dus de log-oddsratio's. Daarachter staan de standard errors. Deel je de schattingen door de standard errors dan krijg je de Wald toetsingsgrootheden. Deze worden gebruikt om een p-waarde te krijgen. De log-oddsratio's kunnen omgezet worden in oddsratio's door de anti-log te nemen van de b_i's (e^{b_i}). De tabel geeft ook het 95% tweezijdige betrouwbaarheidsinterval voor de oddsratio's.

Hoofdstuk 6

Survival analyse

6.1 Overleving

In een onderzoek had men 112 katten die de diagnose melkkliertumor hadden gekregen. Deze katten werden vanaf dat moment - het moment van diagnose - gevolgd en de onderzoekers stelden vast hoelang het dier nog in leven bleef. Dit is een typisch voorbeeld van overlevingsdata: vanaf een van tevoren goed vastgesteld moment meet men de tijdsduur die verstrijkt totdat er een bepaalde gebeurtenis optreedt. De gebeurtenis in dit voorbeeld is dus sterfte en de tijdsduur is de tijd tussen de diagnose en de sterfte. De term overlevingsanalyse wordt ook gebruikt als er geen sprake is van sterfte maar van een andere gebeurtenis. Zo is in een onderzoek naar de uitdrijftijden bij schapen ook een overlevingsanalyse gedaan. De gebeurtenis in dit onderzoek was de geboorte en de tijdsduur is de tijd tussen het begin van de partus en de geboorte.

In het melkkliertumoronderzoek kan men na 1 week uitrekenen welke fractie van de katten nog in leven is. Dit is 0.93 omdat 8 van de 112 katten in de eerste week dood gegaan zijn (zie tabel). Dus 93 % van de katten is na 1 week nog in leven. Dit kan men ook voor de tweede week en de volgende weken uitrekenen. Zo krijgt men voor iedere week een overlevingsfractie. In de populatie noemt men deze overlevingsfracties op de verschillende tijdstippen, de overlevingsfunctie aangegeven met: $\psi(i) = Pr(T > i)$. Hierin betekent Pr "de kans op". Hier staat dus de fractie katten die een overlevingstijd groter dan week i hebben. De overlevingsfracties in de steekproef worden aangegeven met S(i).

Er zijn twee kenmerken aan survivaldata die deze data bijzonder maken. Het eerste is wat men noemt conditionaliteit. Als je bijvoorbeeld wil weten hoe groot de kans is dat een kat week 3 overleeft, dan moet je er van uitgaan dat die kat ook week

twee heeft overleefd. Het heeft geen zin te vragen naar de kans op overleving van week 3 voor katten die voor week 3 niet meer in leven waren. Dus het is de kans dat een kat week 3 overleeft gegeven (onder de voorwaarde) dat de kat aan het begin van week 3 nog in leven was.

Een tweede kenmerk is de zogenaamde cencurering: aan het eind van het onderzoek is de gebeurtenis nog niet opgetreden bij een kat.

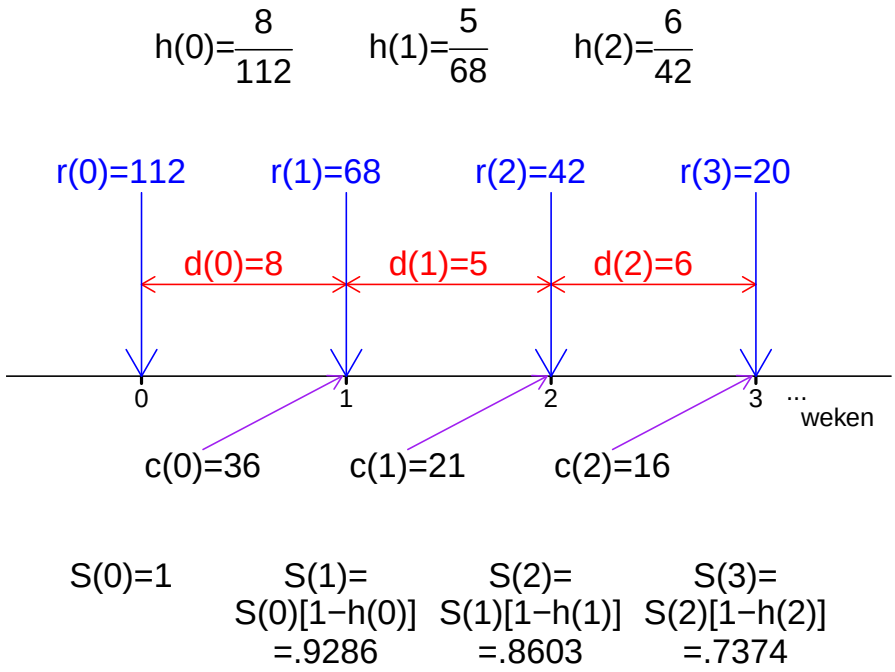
6.2 Conditionaliteit: Hazard

Het sterfterisico op tijdstip i is de kans dat een kat doodgaat gedurende tijdstip i gegeven dat deze kat nog in leven is tot tijdstip i. Dus bijvoorbeeld gegeven dat een kat nog leefde voordat de derde week begon, wat is dan de kans dat deze kat gedurende deze week sterft. Dit sterfterisico wordt in de literatuur aangeduid met "hazard". Voor overlevingsanalyse is het sterfterisico het belangrijkste om naar te kijken. Dat is ook wel logisch: het is alleen zinvol om te spreken over de kans dat een kat in de derde week dood gaat als bekend is dat deze kat voor de derde week nog in leven was. Om de "hazard " uit te kunnen rekenen voor bijvoorbeeld week 3 kijk je eerst naar het aantal katten dat week 3 gehaald heeft en vervolgens bepaal je de fractie katten die dood gaat gedurende die week, (figuur 6.1). Bijvoorbeeld, week 3 begint met 42 levende katten. Dit aantal heet het aantal at risk en wordt aangegeven met $r(3)$. Gedurende week 3 gaan er 6 dood. Dit aantal wordt aangegeven met $d(3)$, de "hazard " voor week 3 is dan $6/42$. De "geschatte hazard " voor week i wordt aangegeven met $h(i)$. De "hazard" wordt gebruikt om voor ieder tijdstip de overlevingsfractie te berekenen.

6.3 Censurering

Een probleem dat zich vaak voordoet bij het meten van overlevingstijden is, dat de gebeurtenis waarin men geinteresseerd is zich niet voordoet bij een aantal dieren in het onderzoek. Dit noemt men censurering. Er zijn verschillende oorzaken voor censurering:

1. Men verliest een dier uit het oog door bijvoorbeeld verhuizing.
2. Een dier kan doodgaan aan een andere oorzaak dan de oorzaak (gebeurtenis) waar het onderzoek overgaat.
3. Een dier kan nog steeds in leven zijn aan het eind van het onderzoek.



Figuur 6.1: Schema voor het berekenen van de overlevingsfracties voor de eerste 3 tijdstippen. c =censurering, d =dood, h =hazard en s =overlevingsfractie

Tabel 6.1: Berekenen van survivalfractie

Als een dier op tijdstip i in het onderzoek zit en na dit tijdstip gecensureerd wordt, dan is alleen bekend dat zijn overlevingstijd groter zal zijn dan i. Tot en met dit tijdstip wordt het dier meegeteld als een dier die het risico loopt te overlijden aan de ziekte waar het onderzoek over gaat. Het dier is "at risk" tot het tijdstip van censurering. Daarna wordt het dier niet meer meegeteld. Dit aantal wordt aangegeven met $c(i)$. In het melkkliertumor onderzoek begint men met 112 katten. Er gaan er in de eerste week 8 dood en er zijn 36 dieren gecensureerd. Op het volgende tijdstip zijn er dan nog 68 dieren over. Men gaat er vanuit dat de 36 gecensureerde katten in de eerste week nog "at risk" waren, de week daarop tellen ze niet meer mee. Er wordt dus verondersteld dat censurering plaats vindt aan het eind van de week, vlak voordat de volgende week begint(zie figuur 6.1)). Met deze aanname heet de berekening van de hazard de Kaplan-Meier schatting

6.4 Schatten van de survivalfractie

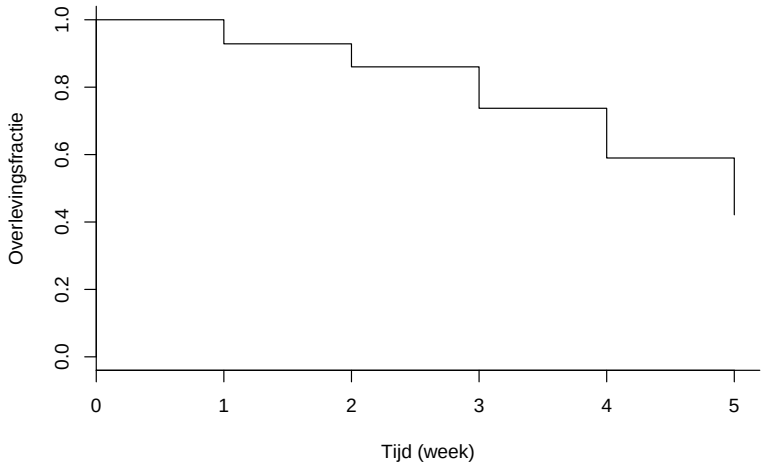
Het aantal katten dat in een week overlijdt, geven we aan met d. Het aantal katten dat de week levend begint, dus in die week "at risk" zijn, geven we aan met r. De "hazard " voor deze week wordt dan geschat met: $h(i) = \frac{d(i)}{r(i)}$, en omdat de censurering wordt geacht plaats te vinden aan het eind van het tijdstip gaat het hier over de Kaplan-Meier schatting.

Nu de overlevingskans. De geschatte kans voor een dier om tot week i te overleven (overlevingsfractie) is gelijk aan de geschatte kans dat week $i-1$ overleefd wordt, maal de geschatte kans dat het dier niet dood gaat gedurende die week :

$$S(i) = S(i-1) \times (1 - h(i-1))$$

$S(0) = 1$ omdat met 100 % van de dieren wordt gestart. Dus, bijvoorbeeld $S(1) = S(0)(1 - \frac{8}{112}) = .9286$ en $S(2) = S(1)(1 - 5/68)) = .8603$. Zie figuur 6.1.
In tabel 6.1 staat de berekening met melkkliertumor als voorbeeld.

Tijd (weken)	at risk (r_i)	dood (d_i)	gecensureerd (c_i)	hazard (h_1)	survival
1	112	8	36	0.0714	0.9285
2	68	5	21	0.735	0.8603
3	42	6	16	0.1429	0.7374
4	20	4	2	0.2	0.5899
5	14	4	19	0.2857	0.4214



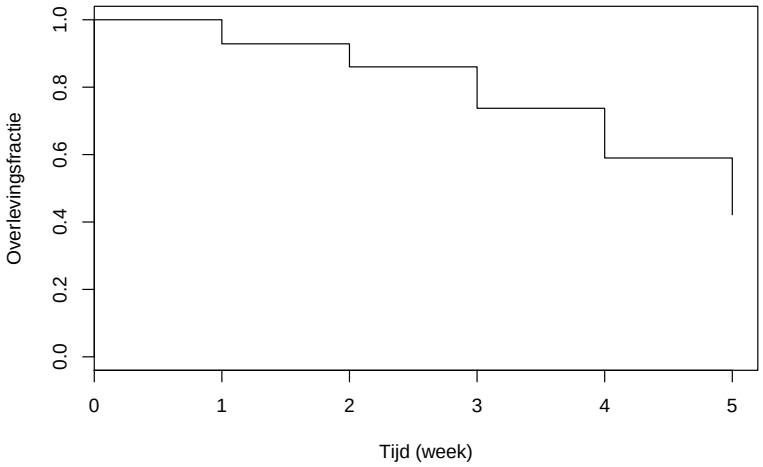
Figuur 6.2: De Kaplan-Meier survival functie

Aan de formules voor de hazard en de survivalfunctie,kan je zien dat als er voor een tijdstip geen sterftegevallen zijn, dat dan $h(i) = 0$ en $1-h(i) = 1$. De overlevingsfractie voor dat tijdstip is dan hetzelfde als de overlevingsfractie van het vorige tijdstip. Daarom kijkt men met deze methode alleen naar tijdstippen waarop er katten overlijden.

Van de overlevingsfracties kan men een grafiek maken. Op de horizontale as komt de tijd te staan en op de verticale as de overlevingsfracties. De tijd as begint met 0. De overlevingsfractie is daar 1. Vervolgens tekent men een horizontale lijn totdat op de tijd as 1 staat. Daar zakt de lijn loodrecht naar beneden tot 0.9286. Vervolgens loopt de lijn weer horizontaal totdat op de tijd as 2 staat. Dan zakt de lijn weer loodrecht naar beneden tot 0.8603, enzovoort. In figuur (6.2) is het resultaat te zien voor de melkkliertumor bij katten.

6.5 Standard errors

De overlevingsfracties zijn schattingen van de populatiefracties met behulp van de steekproef. Om een indruk te krijgen hoe precies we de populatiefracties kunnen schatten, kunnen we de standard error van de steekproef overlevingsfractie ($S(i)$) bepalen. Om de standard error voor de overlevingsfractie op tijdstip i te bepalen in



Figuur 6.3: De survival functie met de standard errors

het geval er geen censurering is, gebruikt men formules $\sqrt{\frac{S(i)[1-S(i)]}{n}}$. Vergelijk dit met de standard error van een steekproeffractie. (zie de formule op blz 51 of in 9.3.1 op blz. 108 van het boek).

Als er sprake is van censurering klopt deze formule niet meer en gebruikt men een andere, de zogenaamde formule van Greenwood:

$$se[S(i)] = S(i)\sqrt{\sum_{alle\ tidstippen \leq i} \frac{d(i)}{r(i)\Big(r(i)-d(i)\Big)}}$$

Eerst wordt voor ieder tijdstip $\frac{d(i)}{r(i)\Big(r(i)-d(i)\Big)}$ berekend. Vervolgens bepaalt men voor ieder tijdstip de som van deze getallen tot en met dit tijdstip. Hieruit wordt de wortel genomen en daarna met S(i) vermenigvuldigd. Op deze wijze wordt voor ieder tijdstip de standard error bepaald. In tabel 6.2 staan de berekeningen. Wat opvalt is dat de "standard error" voor week 5 ongeveer 3.7 keer groter is dan die voor week 1. Dat komt omdat naarmate de tijd verstrijkt er steeds minder katten in het onderzoek overblijven en de schattingen van de populaties daarmee minder precies worden. Men kan met de "standard errors" een betrouwbaarheidsinterval voor de overlevingsfractie

Tabel 6.2: Berekenen van de standard error voor de survivalfracties

Tijd	at risk ($r(i)$)	dood ($d(i)$)	$\frac{d(i)}{r(i)\left(r(i)-d(i)\right)}$	$\sqrt{\sum_{alle\ tijdstippen \leq i} \frac{d(i)}{r(i)\left(r(i)-d(i)\right)}}$	se
1	112	8	.0006868	0.0262	0.0243
2	68	5	0.0011671	0.0431	0.0370
3	42	6	0.0039683	0.0763	0.0563
4	20	4	0.0125	0.1354	0.0799
5	14	4	0.028517	0.2165	0.0912

per tijdstip berekenen met:

$$S(i) \pm 1.96\, se[S(i)]$$

voor een 95% betrouwbaarheidsinterval (1.96 komt uit de standaard normale verdelingstabel). Dit betrouwbaarheidsinterval kan samen met de Kaplan-Meier overlevingsfracties in een grafiek gezet worden (figuur (6.3)).

6.6 Prognose

Uit de overlevingsduurcurve is af te lezen na hoeveel tijd 90%, 80%...10% nog in leven is. Een veelgebruikt percentiel is het 50-ste percentiel, ook wel aangeduid met de mediane overlevingsduur (ongeveer 5 weken, zie figuur 6.3).

Een andere vorm van prognose is de vierweekse overlevingskans: na 4 weken is in het bovenstaand geval 59% van de katten nog in leven.

Referentie: Inleiding tot de medische statistiek, J.C. van Houwelingen e.a.