FACULTEIT PSYCHOLOGIE EN PEDAGOGISCHE WETENSCHAPPEN

STATISTIEK II

Prof. dr. Thierry Marchant

Academiejaar 2020–2021

UNIVERSITEIT GENT

Voorwoord

Deze cursus in een herziene versie van de cursus Statistiek II in 2019-2020. Er zijn niet veel verschillen maar ze zijn belangrijk en het wordt dus afgeraden om de oude cursus te gebruiken.

Een grondige kennis van (en inzicht in) de cursus Statistiek I is noodzakelijk om deze cursus te kunnen studeren. Elementaire vertrouwdheid met het softwarepakket `RStudio` of de programmeertaal `R` wordt verondersteld.

De cursus bevat 122 oefeningen in de marge. Je vindt de oplossingen van die oefeningen aan het einde van elk hoofdstuk. De oefeningen maken deel uit van de examenstof.

In deze cursus wordt veel belang gehecht aan het softwarepakket `RStudio` (in mindere maat aan SPSS en Excel). Dit maakt ook deel uit van de examenstof.

Deze cursus bevat waarschijnlijk een aantal fouten. Meldingen van fouten kunnen naar `Thierry.Marchant@UGent.be` gestuurd worden. Ik ben dank verschuldigd aan mijn collega's van de vakgroep data-analyse en aan veel studenten die enkele fouten in de vorige versie hebben verbeterd. Met speciale aandacht bedank ik Kevin De Gruyter die de hele cursus nauwlettend heeft gelezen en die veel fouten heeft verbeterd.

Inhoudsopgave

Deel I

Cursus

Hoofdstuk 0

Prolegomena

0.1 Statistiek, psychometrie en methodologie

Een onderzoeker wenst één of meerdere onderzoeksvragen te beantwoorden. Deze vragen hebben betrekking op een bepaalde populatie. Hiertoe zal de onderzoeker data verzamelen in een specifieke steekproef die min of meer representatief zal zijn voor de beoogde populatie. Het is de taak van de methodologie om te bepalen hoe de steekproef samengesteld wordt en welk soort van onderzoeksdesign het meest geschikt is om een antwoord te bieden op de onderzoeksvragen. Bv.

- observationeel onderzoek (inclusief vragenlijsten)
- experimenteel onderzoek.

Het is de taak van de psychometrie om te bepalen hoe de variabelen (bv. attitude t.o.v. vreemdelingen) het best kunnen gemeten worden. Eenmaal de data werd verzameld is het de taak van de statistiek (beschrijvend of inductief) om met verschillende technieken de data te analyseren teneinde een antwoord te kunnen formuleren op de vooropgestelde onderzoeksvragen.

De beschrijvende statistiek bestaat uit een aantal technieken om de gegevens (data, waarnemingen) in een steekproef te beschrijven, te ordenen, te presenteren en samen te vatten.

De inductieve statistiek bestaat uit technieken om observaties (op het niveau van de steekproef) te veralgemenen naar de populatie.

0.2 Variabelen

Een *variabele* is een eigenschap die bij de elementen van de populatie of van de steekproef variëert.

Een variabele kan numeriek of niet numeriek zijn. Dit is gewoon een keuze die de onderzoeker maakt. Hij/zij kan het geslacht coderen als man/vrouw of als

0/1. Hij kan de sociabiliteit coderen als asociaal/sociaal/zeer sociaal of gebruik maken van een vragenlijst met numerieke scores.

Een variabele kan continu of discreet zijn. Continu betekent dat er tussen elke twee willekeurige waarden een derde waarde ligt. Als een variabele niet continu is, dan is ze discreet. Er bestaan verschillende technieken om discrete en continue variabelen te analyseren. Discrete variabelen met heel veel mogelijke waarden (bv. lonen) worden vaak op dezelfde manier geanalyseerd als continue variabelen. Dit is in principe niet correct, maar vanuit een pragmatisch standpunt werkt dit goed.

Een variabele wordt meestal aangeduid door een hoofdletter. bv. X of Y. De waarnemingen (de geobserveerde scores of waarden) van een variabele worden door de overeenkomende kleine letter gerepresenteerd. Bv. x of y. De successieve waarnemingen van X in een steekproef worden aangeduid door bv. $x_1, x_2, \ldots, x_n$, waar n de steekproefgrootte is.

0.3 Meetniveaus

We onderscheiden vijf meetniveaus.

- Absolute schaal. De variabele wordt gemeten door gewoon objecten (of mensen) te tellen. Bv. gezinsgrootte, klasgrootte, aantal inwoners, enz. De meeteenheid is vast. Het nulpunt is vast. De variabele is discreet.

- Ratioschaal. Om de variabele te meten moet je eerst een meeteenheid kiezen. Dan moet je het aantal meeteenheden tellen tussen het te meten object en het vaste nulpunt. Bv. leeftijd, reactietijd, lengte, gewicht, oppervlakte, enz. De variabele is continu.

- Intervalschaal. Om de variabele te meten moet je eerst een meeteenheid en een referentiepunt kiezen. Dan moet je het aantal meeteenheden tellen tussen het te meten object en het referentiepunt. De variabele is continu.

- Ordinale schaal. Je kan de te meten objecten ordenen, maar je kan geen meeteenheid definiëren. De waarde van de variabele bij een object is gewoon zijn plaats of rangnummer in de ordening. Bv. uitslag van een wielerwedstrijd, mate van instemming met een bepaalde uitspraak (Likert schaal). De variabele kan continu of discreet zijn.

- Nominale schaal. De te meten objecten kunnen niet geordend worden. bv. postcode, haarkleur, geslacht, enz. De variabele is noch discreet noch continu. Inderdaad om te bepalen of een variabele discreet of continu is, moet je twee willekeurige waarden kiezen en nagaan of er waarden *ertussen* liggen. Dit is onmogelijk indien je de waarden niet kunt ordenen. Binnen de familie van de nominale variabelen onderscheidt men soms de dichtome variabelen, dat zijn variabelen die slechts twee waarden kunnen aannemen (bv. gescheiden of niet). En als die twee waarden 0 en 1 zijn, dan spreekt men van 0-1 variabelen.

De statistische technieken voor variabelen van interval- of ratiomeetniveau zijn identiek. Veel onderzoekers groeperen die twee meetniveaus onder de noemer *continue* variabelen. Dit is niet 100% correct omdat ordinale variabelen ook continu kunnen zijn. Maar in de praktijk is dit OK.

Er zijn specifieke statistische analyses voor nominale variabelen, voor ordinale variabelen en voor variabelen van absoluut meetniveau. Sommige technieken voor nominale variabelen worden ook gebruikt voor ordinale variabelen (met weinig mogelijke waarden). Nominale en ordinale variabelen worden soms gegroepeerd onder de noemer *categorische variabelen*.

0.4 Zinvolheid

Een bewering of uitspraak is zinvol indien haar waarheidswaarde onafhankelijk is van de meetschaal die je gebruikt. M.a.w., indien de bewering correct is met een bepaalde schaal dan blijft ze correct met een andere schaal; indien de bewering fout is met een bepaalde schaal dan blijft ze fout met een andere schaal.

Voorbeeld: de gemiddelde leeftijd in groep A is groter dan in groep B. Stel dat deze bewering juist is wanneer we de leeftijd in jaar uitdrukken. Dan is ze ook correct als we de leeftijd in maanden of eeuwen of seconden uitdrukken. Deze bewering is dus zinvol.

Voorbeeld: de gemiddelde temperatuur in Gent in Februari is dubbel zo groot als in Helsinki. Stel dat deze bewering juist is wanneer we de temperatuur in graden Celsius uitdrukken. Ze is fout indien we de temperatuur in graden Fahrenheit uitdrukken. Deze bewering is dus zinloos.

Voorbeeld: de gemiddelde score op de Likert schaal "intrinsieke motivatie" is groter in groep 1 dan in groep 2. Stel dat deze bewering juist is wanneer we de vijf niveaus van deze schaal coderen d.m.v. 1, 2, 3, 4 en 5. Ze hoeft niet correct te zijn indien we de vijf niveaus coderen d.m.v. 0, 2, 3, 4 en 6. Deze bewering is dus zinloos.

Om zinloze beweringen te vermijden moet je voorzichtig zijn bij het manipuleren van scores. Bij nominale en ordinale variabelen mag je de scores niet optellen of met elkaar vermenigvuldigen of van elkaar aftrekken of door elkaar delen. Dus geen gemiddelde, variantie, covariantie, correlatie, enz. Bij variabelen van intervalmeetniveau mag je de scores optellen en uit elkaar aftrekken. De scores door elkaar delen of met elkaar vermenigvuldigen is riskant. Logaritmes van scores zijn verboden. Je mag wel de afwijkingen (bv. $x_1 - x_2$ of $x_i - \bar{x}$) met elkaar vermenigvuldigen of door elkaar delen. Je mag ook de logaritme van een afwijking berekenen. Bij variabelen van ratiomeetniveau zijn er bijna geen restricties. Bij variabelen van absoluut meetniveau is er geen restrictie.

Hoofdstuk 1

Data manipulatie

De analyse van data of gegevens gebeurt meestal m.b.v. een statistische softwarepakket. In deze cursus zullen we focussen op het softwarepakket `RStudio` dat je al bij het vak Statistiek I hebt leren kennen. Dit softwarepakket is een gebruikersvriendelijke implementatie van de `R` programmeertaal. In het vervolg zal ik dus meestal van `R` spreken en niet van `RStudio`. Als je `RStudio` op je computer nog niet geïnstalleerd hebt, dan doe het nu!

Om gegevens m.b.v. `R` te kunnen analyseren moeten we eerst een databestand aanmaken. Dit kan op veel verschillende manieren en met veel verschillende softwarepaketten. We beperken ons tot drie paketten: `R`, Excel en SPSS.

1.1 De data in R

De simpelste functie om data in `R` te stoppen is `c`. Dit is een functie om vectoren[1] aan te maken. Vb.

1. Typ dit commando en alle andere commando's van de cursus in `R` (behalve waar anders vermeld).

```
> leeftijd <- c(18, 22, 17, 19, 19)
```

Het commando `c(18, 22, 17, 19, 19)` creëert een object dat bestaat uit vijf getallen en het pijltje ' `<-` ' kent de naam `leeftijd` toe aan dit object. Het object '`leeftijd`' wordt in het geheugen van `R` gestopt en kan achteraf gebruikt worden. Vb.

2. Probeer al deze commando's te begrijpen. Als het niet lukt, gebruik dan de hulpfunctie van `R`.

```
> leeftijd
[1] 18 22 17 19 19
> mean(leeftijd)
[1] 19
> length(leeftijd)
[1] 5
> min(leeftijd)
```

[1] Een `R` vector is gewoon een reeks objecten dat als één samengesteld object beschouwd wordt door `R`.

```
[1] 17
> max(leeftijd)
[1] 22
> median(leeftijd)
[1] 19
> leeftijd[1]
[1] 18
> leeftijd[2]
[1] 22
> leeftijd[3]
[1] 17
> leeftijd[6]
[1] NA
```

1.1.1 R en de meetniveaus

Een *string* is een reeks tekens. Bv. Statistiek, intrinsieke_motivatie, ABC1$ en 2018 zijn allemaal strings. Maar deze laatste string kan ook geïnterpreteerd worden als een getal en niet zomaar als een reeks van vier tekens zonder betekenis. Om verwarringen te vermijden moet je strings altijd tussen aanhalingstekens aan `R` doorgeven. Bv.

```
roker <- c("ja", "neen", "ja")
postcode <- c("9000", "2500", "8400")
```

Als je een vector aanmaakt met het commando

```
roker <- c("ja", "neen", "ja")
```

dan weet `R` automatisch dat de drie waarden van de variabele `roker` van ordinaal of nominaal meetniveau zijn: `R` weet dat strings niet numeriek zijn. Als je `R` vraagt om het gemiddelde van de vector te berekenen, dan krijg je een foutmelding:

```
> mean(roker)
[1] NA

Warning message:
In mean.default(roker) :  argument is not numeric or logical:
  returning NA
```

Als je een vector aanmaakt van tramnummers in Gent met het commando `tramnummer <- c(1, 21, 22, 4, 22, 21, 1, 4)`, dan kan `R` niet weten dat die getallen de waarden van een nominale variabele zijn. Als je `R` vraagt om het gemiddelde van de vector te berekenen, dan krijg je *geen* foutmelding:

```
> mean(tramnummer)
[1] 12
```

Om dit te vermijden gebruik je de functie `factor` om `R` te zeggen dat de getallen in de vector als niet numeriek beschouwd moeten worden.[2] Bv.

```
> tramnummer <- factor( c(1, 21, 22, 4, 22, 21, 1, 4) )
```

Als je dan probeert het gemiddelde te berekenen, krijg je een foutmelding:

```
> mean(tramnummer)
[1] NA

Warning message:
In mean.default(tramnummer) :
  argument is not numeric or logical:  returning NA
```

Als je de naam van de vector typt, dan krijg je de vector te zien, maar ook de lijst van de verschillende waarden in de vector. Die waarden worden "levels" genoemd in het `R` jargon.

```
> tramnummer
[1] 1  21 22 4  22 21 1  4
Levels: 1 4 21 22
```

Als je een vector wil aanmaken met waarden van een ordinale variabele, dan gebruik je ook het commando `factor` maar je gebruikt bovendien de argumenten `levels` en `ordered`. Voorbeeld: je wil een vector aanmaken met de uitslagen van een groep atleten. Je gebruikt dit commando:

```
> uitslag <- factor( c("brons", "goud", "goud", "brons", "zilver",
"brons", "brons", "brons"), levels = c( "brons", "zilver", "goud" ),
ordered = TRUE)
```

Als je de naam van de vector typt, dan krijg je de vector te zien, maar ook de lijst van de levels en hun volgorde. Voor sommige analyses is het belangrijk dat `R` de volgorde van de levels kent.

```
> uitslag
[1] brons  goud   goud   brons  zilver brons  brons  brons
Levels: brons < zilver < goud
```

1.1.2 Data frames

We gebruiken een fictief voorbeeld met 8 variabelen waargenomen in een steekproef van $n = 30$ FPPW studenten. De variabelen zijn

`score` – score op het examen statistiek II

`iq` – intelligentie-quotiënt

[2] Een andere mogelijkheid is dat je de tramnummers als strings doorgeeft: `tramnummer <- c("1", "21", "22", "4", "22", "21", "1", "4")`. Het gebruik van `factor` is explicieter en aangeraden.

`motivatie` – gemeten op een Likert schaal van 1 (zeer laag) tot 7 (zeer hoog)

`geslacht` – geslacht van de student: man of vrouw

`roken` – de student rookt regelmatig: ja of neen

`opleiding` – psychologie, pedagogische wetenschappen, sociaal werk

`gewicht` – gewicht van de student in kg

`lengte` – lengte van de student in cm

We gaan de data voor dit voorbeeld in `R` stoppen. Met het commando '`c`', maken we een vector aan met de scores van de studenten en we geven deze vector de naam `score`.

```
> score <- c(16, 10, 11, 14, 8, 18, 13, 9, 11, 10, 5, 14, 11, 11,
0, 18, 19, 18, 9, 6, 4, 18, 9, 20, 3, 6, 11, 6, 16, 18)
```

Probeer niet al die commando's in `R` zelf te typen. Ze zijn veel te lang. Binnenkort krijg je wel eenvoudige `R` oefeningen. We maken nieuwe vectoren aan met de andere variabelen.

```
> iq <- c(127, 125, 138, 104, 118, 132, 121, 120, 82, 103, 145,
119, 109, 111, 128, 133, 128, 94, 86, 119, 126, 106, 90, 119,
116, 133, 119, 106, 139, 122)
> motivatie <- factor( c(4, 2, 1, 6, 5, 5, 5, 6, 1, 6, 5, 2, 7, 5,
5, 5, 6, 5, 5, 2, 3, 7, 1, 1, 1, 3, 2, 7, 2, 6),
levels = c( 1, 2, 3, 4, 5, 6, 7 ), ordered = TRUE)
> geslacht <- c("V", "V", "V", "M", "M", "V", "V", "V", "M", "M",
"M", "V", "V", "V", "V", "M", "V", "M", "M", "M", "M", "M", "V",
"V", "V", "V", "V", "M", "M", "M")
```

Heb je de aanhalingstekens opgemerkt? Zonder aanhalingstekens zou `R` denken dat `V` en `M` de namen van twee variabelen zijn, zoals `iq` of `motivatie`. Heb je ook de functie `factor` opgemerkt bij het aanmaken van de vector `motivatie`? We moeten deze functie (met het argument `ordered = TRUE`) gebruiken om `R` te zeggen dat de cijfers 1 t.em. 7, de waarden van een ordinale variabele zijn.

```
> roken <- c("Neen", "Neen", "Neen", "Neen", "Ja", "Neen", "Ja",
"Ja", "Neen", "Ja", "Neen", "Ja", "Neen", "Ja", "Ja", "Ja",
"Neen", "Ja", "Neen", "Neen", "Neen", "Neen", "Neen", "Ja",
"Neen", "Neen", "Neen", "Neen", "Neen", "Neen")
> opleiding <- c("psy", "psy", "psy", "psy", "psy", "ped", "psy",
"psy", "psy", "ped", "ped", "psy", "soc", "soc", "ped", "ped",
"psy", "ped", "psy", "psy", "ped", "ped", "ped", "psy", "psy",
"psy", "psy", "psy", "ped", "psy")
> gewicht <- c(69, 64, 96, 76, 78, 75, 74, 51, 80, 76, 88, 73,
83, 86, 73, 67, 53, 64, 90, 67, 48, 59, 46, 59, 80, 104, 53, 82,
61, 69)
```

```
> lengte <- c(158, 170, 180, 156, 176, 174, 162, 147, 168, 170,
169, 187, 164, 169, 174, 159, 170, 163, 163, 166, 147, 173, 156,
178, 162, 195, 150, 154, 182, 158)
```

3. Maak een vector aan in R, met naam `sport` *en met de waarden: voetbal, basketbal, voetbal, basketbal, zwemmen, voetbal, badminton, voetbal.*

Wat zijn de scores van de eerste student op alle 8 variabelen? Je kijkt naar het eerste element in alle 8 vectoren en je vindt 16, 127, 4, V, Neen, psy, 69 en 158. Voor jou is het gemakkelijk omdat je weet dat alle 8 vectoren betrekking hebben op dezelfde 30 studenten. Maar `R` weet het niet. We hebben het niet expliciet aan `R` gezegd. Voor `R` zijn er 8 vectoren met lengte 30, maar ze hebben misschien betrekking op 8 verschillende steekproeven met grootte 30. We gaan nu de 8 vectoren in een grotere structuur zetten (een data frame) zodat het duidelijk voor `R` zal zijn dat de 8 vectoren betrekking hebben op dezelfde 30 studenten. We gebruiken het commando '`data.frame`' en tussen haakjes de naam van de 8 vectoren. We geven de naam `myData` aan dit data frame.

4. Maak een vector aan in R, met naam `onderwijs` *en met de waarden: ASO, ASO, TSO, ASO, BSO, TSO, TSO, ASO.*

```
> myData <- data.frame(score, iq, motivatie, geslacht, roken,
opleiding, gewicht, lengte)
```

Om te begrijpen wat er gebeurd is, typ je '`myData`' en je krijgt als output

```
> myData
   score  iq motivatie geslacht roken opleiding gewicht lengte
1     16 127         4        V  Neen       psy      69    158
2     10 125         2        V  Neen       psy      64    170
3     11 138         1        V  Neen       psy      96    180
4     14 104         6        M  Neen       psy      76    156
5      8 118         5        M    Ja       psy      78    176
6     18 132         5        V  Neen       ped      75    174
7     13 121         5        V    Ja       psy      74    162
8      9 120         6        V    Ja       psy      51    147
9     11  82         1        M  Neen       psy      80    168
10    10 103         6        M    Ja       ped      76    170

...

27    11 119         2        V  Neen       psy      53    150
28     6 106         7        M  Neen       psy      82    154
29    16 139         2        M  Neen       ped      61    182
30    18 122         6        M  Neen       psy      69    158
```

Je ziet dat `R` een grote tabel heeft aangemaakt met de 8 vectoren. `R` heeft ook een nummer aan elke student toegekend (eerste kolom links). Vanaf nu verwijst `myData` naar deze tabel. Indien je een specifieke kolom van deze tabel wil raadplegen typ je gewoon `myData` gevolgd door '`$`' en de naam van de variabele. Bv.

```
> myData$gewicht
 [1]  69  64  96  76  78  75  74  51  80  76  88  73  83  86  73
[16]  67  53  64  90  67  48  59  46  59  80 104  53  82  61  69
```

We doen hetzelfde met het geslacht.

```
> myData$geslacht
 [1] V V V M M V V V M M M V V V V M V M M M M M V V V V V M M M
Levels: M V
```

Merk op dat R een extra regel output heeft geproduceerd: 'Levels: M V'. De reden is simpel: R heeft begrepen dat de variabele geslacht niet numeriek is en geeft de lijst weer van alle verschillende waarden (of niveaus) van deze variabele. Indien we het gewicht van de tiende student willen weten, typen we

5. Maak een data frame aan in R, met naam mijnEersteDataFrame *en met de twee vectoren* sport *en* onderwijs*.*

```
> myData$gewicht[10]
[1] 76
```

Met het commando 'dim' (dimensies) krijgen we de grootte van het data frame.

```
> dim(myData)
[1] 30  8
```

Dus dim(myData)[1] geeft de steekproefgrootte en dim(myData)[2] geeft het aantal variabelen. Een andere techniek om de steekproefgrootte te raadplegen is

```
> length(myData$score)
[1] 30
```

of

6. Hoeveel individuen zijn er in mijnEersteDataFrame*? Gebruik de functies* dim *en* length *(twee werkwijzen).*

```
> length(myData$geslacht)
[1] 30
```

Een data frame is een zeer belangrijk object in R want alle datasets worden gestopt in de vorm van een data frame. Alle statistische analyses die we in deze cursus zullen uitvoeren zullen dus toegepast worden op data frames. Je hebt al kunnen opmerken dat een data frame in R aanmaken niet zo handig is. Het typen van lange commando's zoals

7. Probeer deze commando's uit: names(myData) *en* names(myData)[2]*. Probeer de naam* iq *om te zetten naar* intelligentie*.*

8. In welke type onderwijs zit de vijfde individu? Gebruik een R *commando.*

```
> geslacht <- c("V", "V", "V", "M", "M", "V", "V", "V", "M", "M",
"M", "V", "V", "V", "V", "M", "V", "M", "M", "M", "M", "M", "V",
"V", "V", "V", "V", "M", "M", "M")
```

is zeer omslachtig en het is moeilijk om geen fout te maken. Bij voorkeur zullen we dus andere softwarepaketten gebruiken om data frames aan te maken: Excel en SPSS.

Als je een data frame aangemaakt hebt of als je een data frame gewijzigd hebt, dan wens je misschien het op te slaan. Dit doe je met de functie 'write.csv'. Bv. als je het data frame myData wil opslaan, typ je

```
> write.csv(myData, file = 'myData.csv', row.names = FALSE)
```

Let op, dit is het enige commando (in deze cursus) waar het enkele aanhalingsteken gebruikt wordt in R. Bij andere commando's wordt het dubbele aanhalingsteken gebruikt. Het data frame wordt dan opgeslaan in de 'working directory', dat is de directory[3] waar R alle bestanden leest en opslaat. Om de working directory te bepalen, kies je "Set Working Directory ▶ Choose Directory ..." in het menu "Session".

Als je later het data frame myData opnieuw wenst te gebruiken, mag je het lezen m.b.v. het commando 'read.csv'. Bv.

```
> read.csv(file = "myData.csv")
```

Om de gegevens later te kunnen gebruiken moet je ze een naam toekennen. Bv.

```
> myData <- read.csv(file = "myData.csv")
```

Het data frame myData is dan klaar om gebruikt te worden. Je mag het bestand myData.csv ook met Excel lezen.

1.2 De data in Excel

Excel is een zeer handig softwarepakket om tabellen aan te maken en te editen. Het werken met Excel wordt in deze cursus niet beschreven. We gaan ervan uit dat je de data in een Excel bestand (met extensie "xls" of "xlsx") hebt gestopt en we gaan zien hoe je die data kunt exporteren voor verder gebruik in R. Op je computer heb je dus een Excel venster die min of meer op Fig. 1.1 lijkt. We gaan eerst het bestand met een ander formaat opslaan. Ten dien einde gebruiken we het commando "Save As ..." in het menu "File". Een venster gaat open. Nu kies je "Comma separated Values (.csv)" in het drop down menu "File Format". Je kiest ook een naam (bvb. myData.csv) voor het nieuwe bestand en een locatie (een directory). Dan click je op "Save". Op dit moment wordt een bestand aangemaakt met extensie "csv".

9. Voor het gemak kan je het bestand "myData.xlsx" downloaden op Ufora.

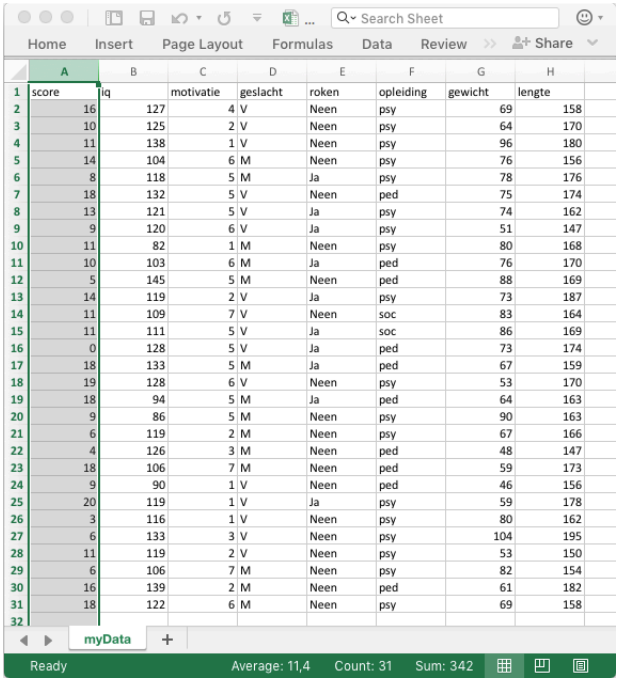
We gaan dit bestand in R importeren. Eerst moet je de working directory wijzigen (zie hierboven) want R leest (of importeert) alleen bestanden die in de working directory zitten. Dan typ je

```
data <- read.csv(file = "myData.csv")
```

En zo heb je een nieuw data frame in R aangemaakt, met naam data. Het bevat de gegevens van de Excel file.

Afhankelijk van je instellingen gebruikt Excel soms een punt en soms een komma als decimaal teken. Om dit na te gaan, maak een nieuw leeg Excel bestand aan, selecteer een cel en kies het formaat 'Number' voor deze cel. In deze cel typ je nu '1/2' en dan druk je op ENTER. Is het resultaat 0.50 of 0,50? Als je versie van Excel een komma als decimaal teken gebruikt en als je een csv bestand hebt aangemaakt, dan gebruik je de functie read.csv met een extra argument om je csv bestand in te lezen:

[3] Een directory wordt ook soms een mapje of folder genoemd.



	A	B	C	D	E	F	G	H
1	score	iq	motivatie	geslacht	roken	opleiding	gewicht	lengte
2	16	127	4	V	Neen	psy	69	158
3	10	125	2	V	Neen	psy	64	170
4	11	138	1	V	Neen	psy	96	180
5	14	104	6	M	Neen	psy	76	156
6	8	118	5	M	Ja	psy	78	176
7	18	132	5	V	Neen	ped	75	174
8	13	121	5	V	Ja	psy	74	162
9	9	120	6	V	Ja	psy	51	147
10	11	82	1	M	Neen	psy	80	168
11	10	103	6	M	Ja	ped	76	170
12	5	145	5	M	Neen	ped	88	169
13	14	119	2	V	Ja	psy	73	187
14	11	109	7	V	Neen	soc	83	164
15	11	111	5	V	Ja	soc	86	169
16	0	128	5	V	Ja	ped	73	174
17	18	133	5	M	Ja	ped	67	159
18	19	128	6	V	Neen	psy	53	170
19	18	94	5	M	Ja	ped	64	163
20	9	86	5	M	Neen	psy	90	163
21	6	119	2	M	Neen	psy	67	166
22	4	126	3	M	Neen	ped	48	147
23	18	106	7	M	Neen	ped	59	173
24	9	90	1	V	Neen	ped	46	156
25	20	119	1	V	Ja	psy	59	178
26	3	116	1	V	Neen	psy	80	162
27	6	133	3	V	Neen	psy	104	195
28	11	119	2	V	Neen	psy	53	150
29	6	106	7	M	Neen	psy	82	154
30	16	139	2	M	Neen	ped	61	182
31	18	122	6	M	Neen	psy	69	158
32								

Figuur 1.1: Een Excel venster met de gegevens van `myData`.

```
data <- read.csv(file = "myData.csv", dec = ",")
```

Met het softwarepakket Excel kan je ook statistische analyses uitvoeren, maar het is afgeraden om statistische analyses met Excel uit te voeren. Excel biedt niet veel technieken aan en is niet zeer betrouwbaar.

1.3 De data in SPSS

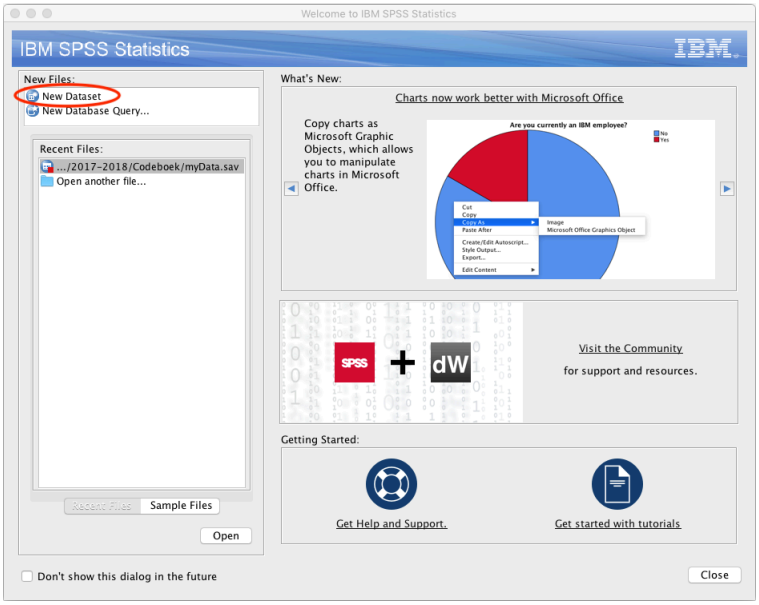
SPSS is een statistische softwarepakket dat ook handig is om data bestanden aan te maken en te editen.

Alhoewel SPSS een goede statistische softwarepakket is, is `R` op veel vlakken toch beter en we focussen dus in deze cursus op `R`. Bovendien is `R` gratis en OS onafhankelijk, i.e. het gebruik van `R` is bijna identiek op alle computers (Windows, Apple of Linux). Uiteindelijk, wie `R` kan gebruiken zal ook gemakkelijk de output van SPSS kunnen interpreteren. In deze cursus gaan we zien hoe je een databestand kunt aanmaken met SPSS.

Als je SPSS opstart,[4] gaat een venster open, dat op Fig. 1.2 lijkt. In functie van de versie van SPSS die je gebruikt, kan het venster licht anders er uitzien.

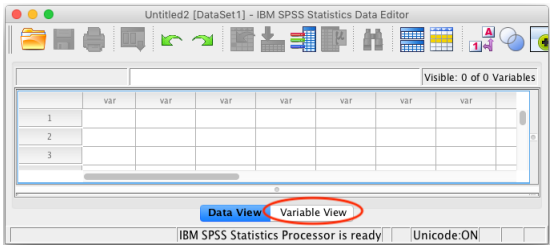
[4] Als je op Athena werkt, krijg je eerst een melding:"Om een betere werking van SPSS op Athena te bekomen, dient u uw data op te slaan als een .zsav bestand." Je mag deze melding negeren.

Om een nieuw databestand aan te maken dubbel-click je op “New Dataset”



Figuur 1.2: Het start-venster bij SPSS.

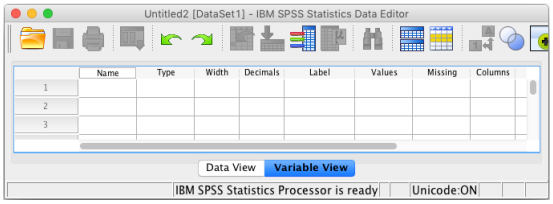
(omcirkeld in het rood op Fig. 1.2). Een nieuw venster gaat open (Fig. 1.3): de data editor van SPSS, in mode “Data View”. Dit venster lijkt op een Excel



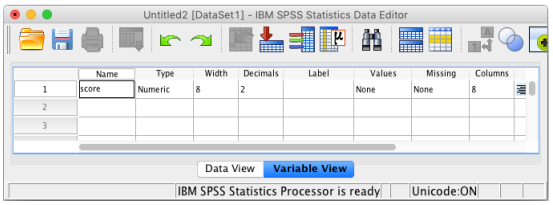
Figuur 1.3: De data-editor bij SPSS: Data View.

venster en werkt inderdaad min of meer zoals Excel, maar er zijn belangrijke verschillen. Vooraleer we onze gegevens in dit venster invullen, gaan we de namen van de variabelen definiëren. Ten dien einde click je op “Variable View” (omcirkeld in het rood op Fig. 1.3). Het uiterlijk van het venster verandert (Fig. 1.4). Dit wordt de “Variable View” genoemd. In dit venster komt elke regel van de tabel overeen met een variabele van je dataset. Typ “score” in het veld “Name”[5] van de eerste rij (Fig. 1.5) en druk “Enter”. Nadat je “Enter”

[5]Let op: de naam van een variabele in SPSS kan geen spatie bevatten. Net zoals in R. In Excel zijn spaties wel toegelaten.



Figuur 1.4: De data-editor bij SPSS: Variable View.



Figuur 1.5: De data-editor bij SPSS: Variable View.

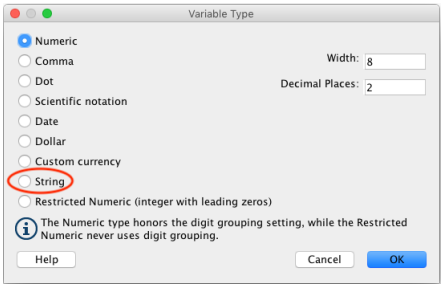
drukt, verschijnen een aantal dingen in de eerste rij. Dit zijn allemaal *default eigenschappen* van de nieuwe aangemaakte variabele. Als je statistische analyses met SPSS wenst uit te voeren, dan zijn die eigenschappen[6] belangrijk en je hoeft ze aan te passen als de default waarden niet geschikt zijn. In deze cursus gebruiken we `R` voor de statistische analyses en veel van die eigenschappen zijn dus niet relevant. Als je die eigenschappen toch aanpast, moet je weten dat ze verloren gaan als je de gegevens exporteert naar `R`. Voor ons is de eigenschap "Type" wel belangrijk. Je ziet dat de default waarde "Numeric" is en dat is in orde voor de variabele "score" want deze variabele is inderdaad numeriek.

Typ nu "iq" in het veld "Name" van de tweede rij en druk "Enter". Nadat je "Enter" drukt, verschijnen default eigenschappen van de nieuwe variabele. De default type van "iq" is "numeric" en dat is in orde. We doen nog hetzelfde voor "motivatie" en voor "geslacht". Hier is er wel een probleem: de variabele "geslacht" is niet numeriek en we gaan dus de type van deze variabele wijzigen. Click op "Numeric" naast "geslacht" en een nieuw venster gaat open (Fig. 1.6). Selecteer "String" (omcirkeld in het rood op Fig. 1.6) en click "OK". De data editor ziet er nu uit zoals op Fig. 1.7. We definiëren nog de variabelen "roken", "opleiding", "gewicht" en "lengte". Bij "roken" en "opleiding" moeten we de type aanpassen naar "String". De data editor ziet er nu uit zoals op Fig. 1.8. Je bent nu klaar om je gegevens in te vullen: je clickt op "Data View" en je kan nu de gegevens invullen in de tabel, net zoals in Excel. Als je klaar bent, mag je de gegevens opslaan. Het bestand krijgt de extensie "sav". De data editor ziet er nu uit zoals op Fig. 1.9.

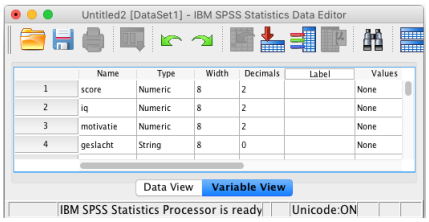
Om de data in `R` te analyseren, moet je ze exporteren. In het menu "File" kies

10. Voor het gemak kan je het bestand "myData.sav" *downloaden op Ufora.*

[6] Een van de eigenschappen is "Measure" en bepaalt het meetniveau van de variabele. SPSS onderscheidt drie meetniveaus: *scale* (interval en ratio), *ordinal* en *nominal*.



Figuur 1.6: Het aanpassen van de "type" van een variabele.



Figuur 1.7: De data-editor bij SPSS: Variable View.

je "Export ▶ CSV Data . . . ". Een venster gaat open (Fig. 1.10). In dit venster zet je "type" op "Comma delimited" en "Encoding" op "Local Encoding". Je geeft ook een naam aan het bestand (bv. myData), je kiest een directory en dan click je op "Save". Een bestand met extensie "csv" wordt aangemaakt.

Het bestand met extensie "csv" is net zoals het bestand met extensie "csv" dat we aangemaakt hebben met Excel (Rubr. 1.2). Het wordt als volgt in `R` geïmporteerd:

```
> gegevens <- read.csv(file = "myData.csv", sep = ";")
```

Let op, als je SPSS op Athena gebruikt, dan gaat waarschijnlijk je "csv" bestand een komma i.p.v. een punt gebruiken als decimaal teken. Om dit bestand in `R` te importeren, gebruik je dan het commando

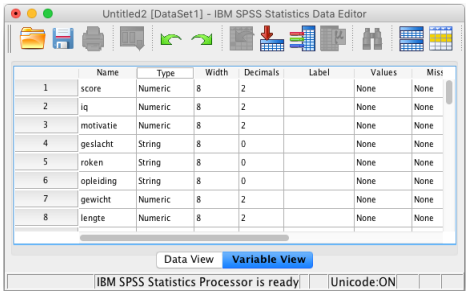
```
> gegevens <- read.csv(file = "myData.csv", sep = ";" , dec = ",")
```

1.4 Geïmporteerde data en meetniveaus

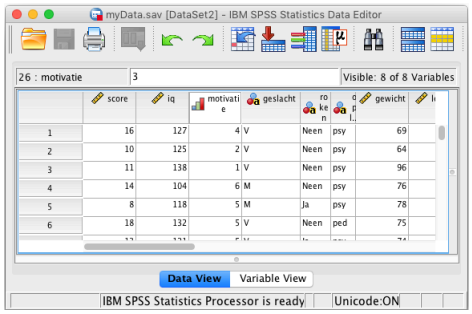
1.4.1 Numerieke variabelen

Numerieke variabelen die uit Excel of SPSS in `R` geïmporteerd worden, worden automatisch door `R` als variabelen van ratio of intervalmeetniveau beschouwd.

Dat is niet altijd correct. Bv. de variabele `motivatie` is numeriek maar ordinaal. Dit moeten we expliciet aan `R` laten weten. We doen het zo:



Figuur 1.8: De data-editor bij SPSS: Variable View.



Figuur 1.9: De data-editor bij SPSS: Data View.

```
> data$motivatie <- factor(data$motivatie,levels = c( 1, 2, 3, 4,
  5, 6, 7 ), ordered = TRUE)
```

Indien de variabele `motivatie` nominaal zou zijn, dan zouden we dit commando gebruiken:

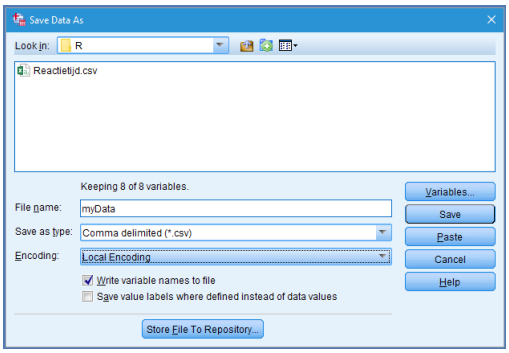
```
> data$motivatie <- factor(data$motivatie)
```

1.4.2 Niet numerieke variabelen

Niet numerieke variabelen die uit Excel of SPSS in `R` geïmporteerd worden, worden automatisch door `R` als nominale variabelen beschouwd. Dat is niet altijd correct: een niet numerieke variabele kan ook ordinaal zijn. Dit moeten we expliciet aan `R` laten weten, net zoals in vorige rubriek.

1.5 Het codeboek

Een codeboek is een document (al dan niet digitaal) dat beschrijft hoe gegevens in een data bestand gecodeerd worden. Stel bv. dat je “man” en “vrouw” hebt gecodeerd door 0 en 1. In je codeboek staat dan expliciet genoteerd wat de cijfers 0 en 1 betekenen of representeren. Het is belangrijk dat je een codeboek



Figuur 1.10: Exporteren naar CSV.

voor elk onderzoek opstelt. Dit codeboek is nuttig voor jezelf indien je jaren later je gegevens opnieuw wenst te analyseren. Het codeboek is ook nuttig voor andere navorsers (medestudent, promotor, collega, enz.) die je gegevens willen gebruiken. Dit past in het *Ghent University Policy Framework on Research Data Management*[7]. Meer informatie over het delen van data vind je bv. in [Ellis and Leek, 2017].

1.5.1 Structuur

Een codeboek bevat onderstaande elementen:

- De naam van de onderzoeker(s);
- De periode waar het onderzoek plaatsvond;
- Een korte beschrijving van het onderzoek en, indien mogelijk, een verwijzing naar een publicatie waar het onderzoek gerapporteerd wordt;
- Informatie dat toelaat om het corresponderende databestand te raadplegen. Dit kan een url (adres van het data bestand op internet) zijn of contactgegevens van de onderzoeker of ... Let op, er zijn wettelijke of ethische restricties voor het delen van gegevens.
- Indien er meerdere versies van het databestand zijn, dan zijn er ook meerdere versies van het codeboek en de koppeling moet duidelijk zijn (versienummers, datum, ...);
- Per variabele, een nauwkeurige beschrijving van het meetinstrument (met een eventuele verwijzing naar een publicatie waar het meetinstrument beschreven wordt) en van de codering.
- Enz.

[7] https://www.ugent.be/en/research/datamanagement/policies/rdm-policy.pdf

1.5.2 Codering

Het is onmogelijk om alle mogelijke coderingen en hun beschrijvingen op te sommen. Hieronder vind je gewoon een paar voorbeelden.

Voor alle variabelen van ratio meetniveau wordt de meeteenheid vermeld. Dit is soms triviaal, maar niet altijd. Voor variabelen van interval meetniveau wordt het nulpunt ook vermeld.

Ordinale en nominale variabelen worden vaak numeriek gecodeerd. Bv. 0 en 1 voor man en vrouw of de cijfers 1 t.e.m. 7 voor de motivatie, enz. Die codes moeten expliciet uitgelegd worden. Vergeet niet de volgorde (voor ordinale variabelen) te vermelden. Stelt het cijfer 7 de hoogste of de laagste motivatie voor?

Het is aangeraden om vanzelfsprekende codes te gebruiken. Bv. M en V voor het geslacht (en niet 0 en 1), zoals in het data frame `myData`. Het is ook mogelijk om categorische variabelen helemaal niet te coderen. Bv. "man" en "vrouw" voor het geslacht. Of "sterk gemotiveerd", "matig gemotiveerd", enz. in plaats van de cijfers van 1 t.e.m. 7. Het nadeel hiervan is dat alle grafieken en tabellen dan onoverzichtelijk worden. Probeer een compromis te vinden tussen bondigheid en duidelijkheid.

1.5.2.1 Missing data

Omwille van allerlei problemen gebeurt het zelden dat elke variabele van het onderzoek gemeten wordt bij elke individu in de steekproef. Vb. de respondent heeft niet alle vragen van een schriftelijke vragenlijst beantwoord; de respondent heeft geweigerd een vraag van een interview te beantwoorden; een antwoord is niet leesbaar; omwille van een technisch probleem werkte het meetinstrument niet tijdens een bepaalde periode; enz. In het algemeen wordt zo'n ontbrekend antwoord beschouwd als "missing data".[8] In `R` wordt dit gecodeerd door `NA`. Dit staat voor "Not Available". Vb.

```
gewicht <- c(69, 64, NA, 76, 78, 75, 74, 51)
```

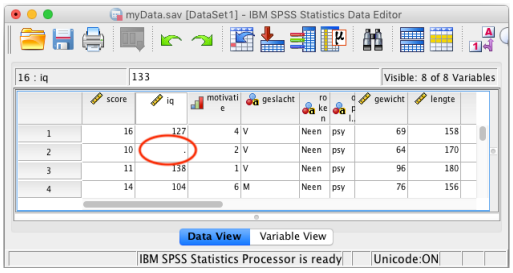
Merk op dat we geen aanhalingstekens zetten om `NA`.

In SPSS is er geen specifiek symbool voor een missing value. Een missing value is gewoon een blanco of leeg veld in de Data View van de data editor (zie Fig. 1.11). Indien een missing value voorkomt bij een numerieke variabele dan toont SPSS een leeg veld met gewoon een puntje erin (alhoewel het veld leeg is).

In Excel is er ook geen specifiek symbool voor een missing value. Een missing value is gewoon een blanco of leeg veld in de spreadsheet.

De drie softwarepaketten `RStudio`, Excel en SPSS bieden allemaal een techniek om missing values te representeren (`NA` of lege velden) maar het kan soms handig zijn om andere codes zelf te definiëren. Bij een interview kan het soms wenselijk zijn om een onderscheid te maken tussen "de respondent weet het

[8] Bij andere vakken zal je zien hoe statistische analyses uitgevoerd kunnen worden indien er missing data zijn.



Figuur 1.11: Een missing value.

niet", "de respondent wil niet antwoorden" en "de batterij van mijn opnametoestel was op en ik weet niet meer wat de respondent geantwoord heeft". Je mag drie verschillende codes kiezen voor die drie verschillende situaties. De codes in zich zijn niet belangrijk. Maar je moet hen consistent gebruiken en zeker zijn dat de codes alleen voor missing values gebruikt worden. Een code kan een string zijn indien de variabele niet numeriek is. Als de variabele numeriek is, dan moet de code ook numeriek zijn. Bv. waarden zoals 99, 999 of -9999. Je vermijdt best de code 9 of 99 indien de variabele leeftijd opgenomen wordt.

Als je data van Excel of SPSS naar `R` wil importeren, hebben we gezien dat je de functie `read.csv` moet gebruiken. Tijdens het importeren worden lege velden bij numerieke variabelen automatisch omgezet naar `NA`. Lege velden bij niet-numerieke variabelen worden omgezet naar lege strings in `R`.

1.5.2.2 Niet van toepassing

Bij sommige variabelen is de waarde NVT of "Niet van Toepassing" nuttig. Dit is bijzonder belangrijk bij conditionele vragen. Bv. bij het Sexpert onderzoek[9] krijgen de participanten de vraag "Had U seks gedurende de voorbije zes maanden?" De antwoordmogelijkheden zijn JA of NEEN. Wie JA antwoordt krijgt als volgende vraag "Had U seks in de voorbije twee weken?" Wie NEEN antwoordt krijgt de tweede vraag niet. Hoe wordt het antwoord op deze tweede vraag gecodeerd voor iemand die NEEN antwoordde op de eerste vraag? Hier moet je ook een code kiezen.

1.5.3 Voorbeelden

Ellis and Leek [2017] presenteren een voorbeeld van een onderzoek omtrent verschillende hormoonspiegels bij patiënten met diabetes (Fig. 1.12).

Sinds 1997 organiseert de Operationele Directie Volksgezondheid en Surveillance van het Wetenschappelijk Instituut Volksgezondheid (WIV) een grootschalige Gezondheidsenquête[10] onder de bevolking in België, ongeveer om de vier

[9] Samenwerking UGent en KUL — https://www.ugent.be/pp/ekgp/nl/onderzoek/onderzoeksgroepen/relatie-en-gezinsstudies/sexpert

[10] http://www.gezondheidsenquete.be

Study Design:

Experimental Question: This study looks to determine whether or not there are differences in hormone levels in individuals with diabetes relative to healthy controls.

Sample Details: 20 individuals with diabetes and 20 unrelated age- and sex-matched controls were included for study. Individuals were recruited to the study using flyers posted throughout Johns Hopkins Hospital and online recruitment through www.website.com. Informed consent was obtained from all study participants. Blood was drawn by a single phlebotomist in clinic X and all samples processed on the same day they were collected by company Y.

Code Book/Data dictionary:

Variable	Description	Units	CodingNotes	OtherNotes
Age	Age At Blood Draw	years	numerical	Taken from electronic medical record
Sex	Self-reported	'male', 'female'	2-level factor	Confirmed using electronic medical record
BMI	weight/height	kg/m^2	numerical	Measured day of blood draw
Collection Date	Date of Blood Draw	date	YYYY-MM-DD	Collection of blood by phlebotomist
Diagnosis	Individual diagnosis	'diabetes', 'control'	2-level factor	`diabetes` = Type 2 Diabetes. Confirmed by medical record.
Cortisol	Stress Hormone	µg/dL	numerical	Required fasting and to be measured in the AM (8-10am)
IGF1	Insulin-Like Growth Factor 1	ng/dL	numerical	Did not require fasting, but taken at the same time as other measures
⋮	⋮	⋮	⋮	⋮
Hormone50	Hormone Name	ng/dL	numerical	Hormone Details

Figuur 1.12: Een codeboek [Ellis and Leek, 2017].

jaar. Het document "Highlights of the Belgian Health Interview Survey 2008" [Charafeddine et al., 2011b] rapporteert bv. de bevindingen van de enquête van het jaar 2008. Honderden variabelen worden bij dit onderzoek gemeten en de databestanden zijn beschikbaar voor verder onderzoek door andere onderzoekers (mits aanvraag bij de Commissie voor de Bescherming van de Persoonlijke levenssfeer). Een codeboek beschrijft alle variabelen en hun codering. Je kan het downloaden op dit adres: https://his.wiv-isp.be/nl/SitePages/Toegang_gegevens2013.aspx.

1.6 Oplossingen

3) Maak een vector aan in R, met naam `sport` en met de waarden: voetbal, basketbal, voetbal, basketbal, zwemmen, voetbal, badminton, voetbal.

Oplossing: Het commando is `sport <- c("voetbal", "basketbal", "voetbal", "basketbal", "zwemmen", "voetbal", "badminton", "voetbal")`.

4) Maak een vector aan in R, met naam `onderwijs` en met de waarden: ASO, ASO, TSO, ASO, BSO, TSO, TSO, ASO.

Oplossing: Het commando is `onderwijs <- c("ASO", "ASO", "TSO", "ASO", "BSO", "TSO", "TSO", "ASO")`.

5) Maak een data frame aan in R, met naam `mijnEersteDataFrame` en met de twee vectoren `sport` en `onderwijs`.

Oplossing: Het commando is `mijnEersteDataFrame <- data.frame( sport, onderwijs)`.

6) Hoeveel individuen zijn er in `mijnEersteDataFrame`? Gebruik de functies `dim` en `length` (twee werkwijzen).

Oplossing:

```
> dim( mijnEersteDataFrame )[1]
[1] 8
> length(mijnEersteDataFrame$sport)
[1] 8
```

7) Probeer deze commando's uit: `names(myData)` en `names(myData)[2]`. Probeer de naam `iq` om te zetten naar `intelligentie`.

Oplossing:

```
> names(myData)
[1] "score"         "intelligentie" "motivatie"     "geslacht"
[5] "roken"         "opleiding"     "gewicht"       "lengte"
> names(myData)[2]
[1] "iq"
> names(myData)[2] <- "intelligentie"
```

We verifiëren nu of de naam veranderd is.

```
> names(myData)[2]
[1] "intelligentie"
```

Prima!

8) In welke type onderwijs zit de vijfde individu? Gebruik een R commando.

Oplossing:

```
> mijnEersteDataFrame$onderwijs[5]
[1] BSO
Levels: ASO BSO TSO
```

Hoofdstuk 2

Beschrijvende statistiek

2.1 Ordeningstechnieken

Vanaf nu zal je zelf alle R commando's van de cursus typen. Te dien einde ga je eerst het bestand myData.csv van Ufora downloaden (onder documenten → R). Je gaat het bestand dan openen met RStudio. Dit bestand bevat het data frame `myData`. ,

Om een frequentieverdeling in R te krijgen typ je bijvoorbeeld

```
> table(myData$opleiding)

ped psy soc
 10  18   2
```

11. Wat is de frequentieverdeling van `roken`*? Gebruik een R commando.*

Voor de relatieve frequentieverdeling moet je nog door de steekproefgrootte n delen:

```
> table( myData$opleiding ) / dim( myData )[1]

       ped        psy        soc
0.33333333 0.60000000 0.06666667
```

Om de relatieve frequentieverdeling te bekomen, mag je ook de functie `prop.table` gebruiken:

```
> prop.table( table( myData$opleiding ) )

       ped        psy        soc
0.33333333 0.60000000 0.06666667
```

Deze functie transformeert een frequentieverdeling in een relatieve frequentieverdeling. Indien de variabele numeriek is, dan worden de verschillende scores geordend. Bv.

```
> table(myData$gewicht)

 46  48  51  53  59  61  64  67  69  73  74  75  76  78  80  82
  1   1   1   2   2   1   2   2   2   2   1   1   2   1   2   1
 83  86  88  90  96 104
  1   1   1   1   1   1
```

12. Wat is de relatieve frequentieverdeling van `roken`*? Gebruik een R commando.*

Voor een bivariate frequentieverdeling gebruik je opnieuw de functie `table`, met als argumenten de namen van de twee variabelen. Voorbeeld:

```
> table( myData$geslacht , myData$opleiding )

   ped psy soc
M    7   7   0
V    3  11   2
```

13. Vraag de bivariate frequentieverdeling aan voor `roken` *en* `motivatie`*.*

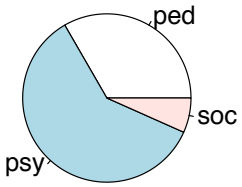
14. Vraag de relatieve bivariate frequentieverdeling aan voor `geslacht` *en* `opleiding`*.*

2.2 Grafische voorstellingen

Cirkeldiagram Voor nominale variabelen is het cirkeldiagram (pie chart in het Engels) bijzonder geschikt. De functie `pie` heeft twee argumenten nodig: `x` (een vector met de frequenties of de proporties) en `labels` (een vector met de namen van de categorieën). Voorbeeld:

```
> pie(x = c(10, 18, 2), labels = c("ped", "psy", "soc"))
```

Je vindt de output van dit commando in Fig.2.1. In plaats van twee vectoren



Figuur 2.1: Cirkeldiagram voor `opleiding`.

door te geven aan de functie `pie`, kunnen we een tabel geven:

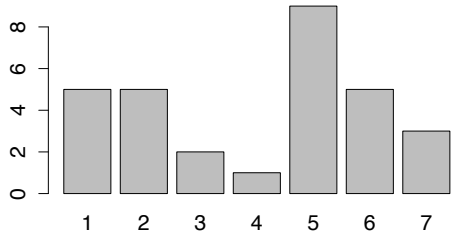
```
> pie(table(myData$opleiding))
```

De output is net dezelfde. Het is normaal want een tabel bestaat eigenlijk uit twee vectoren. R begrijpt zelf welke kolom van de tabel de labels bevat.

15. Teken een cirkeldiagram voor de frequentieverdeling van `score` *m.b.v. R.*

Lijndiagram Voor discrete variabelen is het lijndiagram of staafdiagram (bar chart in het Engels) vaak nuttig. De functie `barplot` heeft ook twee argumenten nodig: een vector met de waarden van de variabele en een vector met de corresponderende frequenties. We kunnen hier ook de twee vectoren in de vorm van een tabel doorgeven (output in Fig.2.2):

```
> barplot(table(myData$motivatie))
```



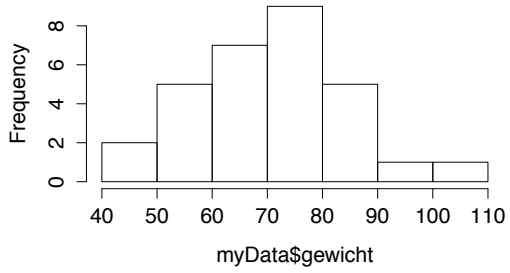
Figuur 2.2: Bar chart voor `motivatie`.

Merk op dat de grafiek in Fig.2.2 geen histogram is: de rechthoeken raken elkaar niet.

16. Teken een staafdiagram voor de frequentieverdeling van `geslacht` *m.b.v. R.*

Histogram Voor variabelen van ratio of interval meetniveau gebruiken we de functie `hist` om een histogram te representeren. Deze functie heeft slechts één argument nodig: de lijst van alle scores waarvoor je een histogram wenst. Voorbeeld (output in Fig.2.3):

```
> hist(x = myData$gewicht)
```



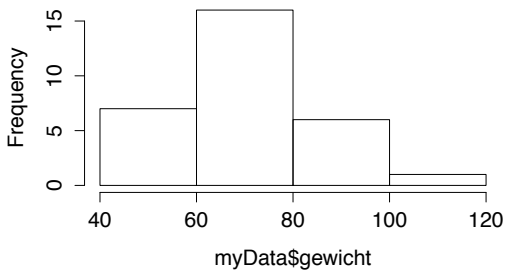
Figuur 2.3: Histogram voor `gewicht`.

Je ziet in Fig.2.3 dat R de data heeft gegroepeerd in 7 klassen en R heeft zelf de klassegrenzen bepaald. Indien gewenst kan je zelf het aantal klassen bepalen met het argument `break` (output in Fig.2.4):

```
> hist(x = myData$gewicht, breaks = 4)
```

Maar, indien R denkt dat het gewenste aantal klassen geen goede keuze is, dan zal R zelf een ander aantal bepalen.

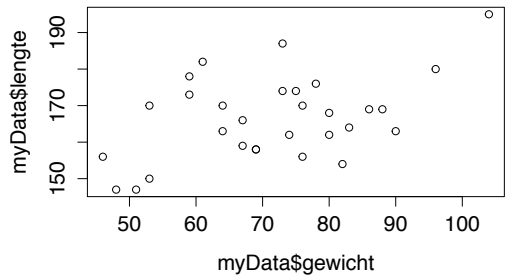
17. Teken een histogram voor de frequentieverdeling van `lengte`*, met 5 klassen, m.b.v. R.*



Figuur 2.4: Histogram voor `gewicht`, met 4 klassen.

Spreidingsdiagram Voor bivariate frequentieverdelingen van variabelen van ratio of interval meetniveau is het spreidingsdiagram (scatterplot in het Engels) de grafische voorstelling bij uitstek. Het wordt getekend door de functie `plot` met twee argumenten: `x` (de lijst van alle scores van de variabele op de horizontale as) en `y` (de lijst van alle scores van de variabele op de verticale as). Je vindt de output in Fig.2.5:

```
> plot(x = myData$gewicht, y = myData$lengte)
```



Figuur 2.5: Spreidingsdiagram voor `gewicht` en `lengte`.

18. Teken een spreidingsdiagram voor `lengte` *en* `motivatie` *m.b.v. R.*

De visuele analyse van Fig.2.5 toont een matig stijgend verband tussen de twee variabelen.

2.3 Reductietechnieken

2.3.1 Centrummaten

Het gemiddelde De formule van het rekenkundig gemiddelde (mean of average in het Engels) is

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

De R functie voor het gemiddelde is `mean`.

```
> mean(myData$score)
[1] 11.4
> mean(c(12, 14, 9, 15, 8, 11, 13, 17, 8, 16, 12, 11, 200))
[1] 26.6
```

Het gemiddelde is gevoelig aan outliers. Vooraleer je het gemiddelde berekent, is het dus aangeraden om je gegevens grafisch te representeren en visueel na te gaan of er outliers zijn. Zo ja, dan moet je beslissen of ze al dan niet verwijderd moeten worden (tikfouten). Zo neen, dan gebruik je best een andere centrummaat.

Het gemiddelde wordt nooit gebruikt met nominale of ordinale variabelen. Met variabelen van interval meetniveau moet je voorzichtig zijn. bv. $\overline{x} > \overline{y}$ is zinvol, maar $\overline{x} > 2\overline{y}$ niet. Met variabelen van ratio en absoluut meetniveau is er geen probleem.

19. Bereken het gemiddelde van `lengte` *m.b.v. R.*

De mediaan De mediaan is gedefinieerd bij variabelen van minstens ordinaal meetniveau. Om de mediaan van X (symbool md_X) te bepalen moet je alle scores ordenen van klein naar groot. Indien n oneven is, dan is de mediaan gelijk aan de waarneming met rangnummer $(n+1)/2$. Indien n even is, dan is de mediaan gelijk aan het gemiddelde van de waarnemingen met rangnummers $n/2$ en $1+n/2$. Als de variabele ordinaal is en als n even is, dan is de mediaan niet gedefinieerd (het gemidddelde mag niet berekend worden) behalve indien waarnemingen met rangnummers $n/2$ en $1+n/2$ gelijk aan elkaar zijn.

De definitie van de mediaan is een beetje omslachtig, maar het idee is simpel: 50% van de waarnemingen zijn kleiner dan md_X en 50% zijn groter dan md_X. R implementatie:

```
> median(c(10, 15, 13, 17))
[1] 14
> median(myData$score)
[1] 11
```

Laten we de mediaan van `motivatie` berekenen:

```
> median(myData$motivatie)
```

```
Error in median.default(myData$motivatie) :  need numeric data
```

R geeft een foutmelding. Hij wil de mediaan van een ordinale variabele niet berekenen. Hij heeft gelijk omdat de mediaan van ordinale variabelen niet altijd gedefinieerd is; als n even is, kan het gebeuren dat de mediaan tussen twee opeenvolgende waarden (of niveaus) valt. In dat geval moet R het middelpunt berekenen en dat bestaat eigenlijk niet bij ordinale variabelen. Maar in de praktijk gaan we toch soms de mediaan van ordinale variabelen berekenen. We gebruiken een truc. We liegen tegen R en we beweren dat de variabele van een hoger meetniveau is, met de functie `as.numeric`. We illustreren eerst het effect van deze functie:

```
> as.numeric(myData$motivatie)
 [1] 4 2 1 6 5 5 5 6 1 6 5 2 7 5 5 5 6 5 5 2 3 7 1 1 1 3 2 7 2 6
```

Deze functie heeft de niet-numerieke niveaus van de variabele `motivatie` omgezet naar numerieke niveaus.[1] De volgorde is gerespecteerd en deze nieuwe codering van de variabele is dus volledig equivalent aan de oorspronkelijke codering. Opgelet: de variabele `motivatie` werd niet gewijzigd; de functie `motivatie` maakt een nieuwe vector aan waarvan de waarden numeriek zijn. We kunnen nu de mediaan van `motivatie` berekenen:

```
> median( as.numeric( myData$motivatie ) )
[1] 5
```

Stel nu dat de uitkomst 5.5 zou zijn. Dan zou de mediaan niet gedefinieerd zijn. Maar we zouden toch weten dat het centrum van de verdeling ergens tussen 5 en 6 ligt. Deze informatie is wel relevant. Let op, bij de interpretatie van de mediaan moet je rekening houden met de omzetting die gebeurde bij de uitvoering van `as.numeric`. Bv.

```
> median( as.numeric( uitslag ) )
[1] 1
```

De mediaan is dus het eerste niveau, dat is, brons.

Nu een tricky voorbeeld. Stel dat je de vijf niveaus van een Likert schaal codeert als -2, -1, 0, 1 en 2. Je voert je gegevens in:

```
> antwoorden <- factor( c(1, -2, 2, 1, 0, 2, 2, 2),
   levels = c( -2, -1, 0, 1, 2 ), ordered = TRUE)
```

Je berekent dan de mediaan:

```
> median(as.numeric(antwoorden))
[1] 4.5
```

De mediaan ligt tussen het vierde en het vijfde niveau, dat is, tussen 1 en 2.

De mediaan is niet gevoelig aan outliers.

20. Bereken de mediaan van `geslacht` *m.b.v. R.*

De modus De modus *mo* is de score met de hoogste frequentie (of de frequentste score). Om de modus te bepalen moet je de hoogste frequentie in de frequentieverdeling opzoeken. De corresponderende waarde is de modus. Er zijn soms meerdere modi. De modus is niet gevoelig aan outliers.

R voorbeeld:

[1] Als de niveaus niet 1, 2, 3, ... zijn, dan worden ze wel omgezet naar de numerieke waarden 1, 2, 3, ... Bv.

```
> as.numeric(uitslag)
[1] 1 3 3 1 2 1 1 1
```

```
> table(myData$roken)

  Ja Neen
  10   20
```

21. Bereken de modus van `geslacht` *m.b.v. R.*

De hoogste frequentie is 20. De modus is dus "Neen".

2.3.2 Spreidingsmaten

De variantie en de standaarddeviatie De variantie van de variabele X in de steekproef wordt gegeven door de formule

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Het wordt ook soms gepresenteerd met de *sum of squares* formule:

$$s_X^2 = \frac{\mathrm{SS}_X}{n-1} \quad \text{met} \quad \mathrm{SS}_X = \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

De R functie om de variantie te berekenen is `var`. Voorbeeld:

```
> var(myData$score)
```

22. Bereken de variantie van `lengte` *m.b.v. R.*

De variantie is altijd positief of nul (want ze is een som van kwadraten). We kunnen de varianties in twee steekproeven vergelijken en vaststellen dat de variantie van X in steekproef 1 groter is dan in steekproef 2. De spreiding is dus groter in steekproef 1 dan in steekproef 2. Verder is is de variantie moeilijk te interpreteren. Ze variëert niet tussen 0 en 1 of tussen 0 en 100. Het is dus onmogelijk om te zeggen of een variantie groot of klein is. We kunnen ook niet de varianties van twee verschillende variabelen vergelijken. Voorbeeld:

$$s^2_{\texttt{score}} = 28.5 < s^2_{\texttt{lengte}} = 125.$$

Deze bewering is zinloos. Inderdaad, indien de onderzoeker de lengte in meter zou uitdrukken, dan zouden we de omgekeerde ongelijkheid bekomen:

$$s^2_{\texttt{score}} = 28.5 > s^2_{\texttt{lengte}} = 0.01.$$

In plaats van de variantie gebruikt men soms de standaarddeviatie; het is gewoon de vierkantswortel uit de variantie:

$$s = \sqrt{s^2}.$$

De standaarddeviatie is ook altijd positief of nul en kan moeilijk geïnterpreteerd worden. Er is een R functie om de standaarddeviatie te berekenen: `sd`. Voorbeeld:

```
> sd(myData$score)
[1] 5.334339
```

De variantie en de standaarddeviatie zijn gevoelig aan outliers. Vooraleer je hen berekent, is het dus aangeraden om je gegevens grafisch te representeren en visueel na te gaan of er outliers zijn. Zo ja, dan moet je beslissen of ze al dan niet verwijderd moeten worden (tikfouten). Zo neen, dan gebruik je best een andere spreidingsmaat.

De variantie en de standaarddeviatie worden nooit gebruikt met nominale of ordinale variabelen. Met variabelen van interval, ratio en absoluut meetniveau is er geen probleem.

23. Bereken de standaarddeviatie van `lengte`*. Eens met de functie* `var` *en een met de functie* `sd`*.*

De variatiebreedte Deze spreidingsmaat is zeer gemakkelijk te berekenen en te interpreteren. Het is gewoon de afwijking tussen de grootste en de kleinste score. Implementatie in R:

```
> max(myData$iq) - min(myData$iq)
[1] 63
```

De variatiebreedte is een afwijking en mag dus niet met nominale en ordinale variabelen gebruikt worden. Het is duidelijk zeer gevoelig aan outliers.

De interkwartiele afstand Om de interkwartiele afstand te definiëren moeten we eerst het 1ste en het derde kwartiel definiëren. Het idee is simpel: het eerste kwartiel (of ook percentiel 25, met symbool P_{25}) wordt zodanig gekozen dat 25% van de waarnemingen kleiner zijn dan P_{25} en 75% groter zijn dan P_{25}; het derde kwartiel (of ook percentiel 75, met symbool P_{75}) wordt zodanig gekozen dat 75% van de waarnemingen kleiner zijn dan P_{75} en 25% groter zijn dan P_{75}. De manier waarop het exact berekend wordt is een beetje omslachtig (zoals de mediaan). De interkwartiele afstand is de afwijking tussen het eerste en het derde kwartiel: $Q = P_{75} - P_{25}$.

De interkwartiele afstand kan gezien worden als een soort gecorrigeerde variatiebreedte: het is de variatiebreedte van onze steekproef nadat we de extreme scores hebben weggegooid.

Er bestaat een R functie om de interkwartiele afstand te berekenen:

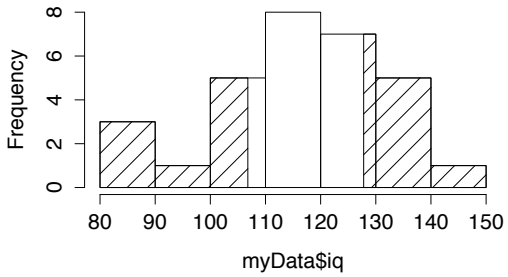
```
> IQR(myData$iq)
[1] 21
```

Fig.2.6 toont het histogram van `iq`, met de eerste en derde kwartielen. De interkwartiele afstand is de breedte van de niet gearceerde zone.

24. Bereken de interkwartiele afstand van `lengte` *m.b.v. R.*

De spreidingsmaat d Deze spreidingsmaat is bijzonder geschikt voor nominale variabelen. Indien het aantal mogelijke waarden van de variabele groot is, worden de gegevens best eerst gegroepeerd. De formule is

$$d = \frac{1 - \frac{f_{mo}}{n}}{1 - \frac{1}{p}}$$



Figuur 2.6: Histogram van `iq`.

waar f_{mo} de frequentie van de modus is (dus de grootste frequentie) en p het aantal verschillende waarden is (of het aantal klassen). Laten we dit berekenen voor `opleiding`. We vragen eerst de frequentieverdeling van `opleiding` op:

```
> table(myData$opleiding)

ped psy soc
 10  18   2
```

Dus $f_{mo} = 18$ en $p = 3$. De maat d is dan gelijk aan

```
> (1 - ( 18/30 )) / (1 - ( 1/3 ))
[1] 0.6
```

De spreidingsmaat d is gemakkelijk te interpreteren. Ze variëert altijd tussen 0 en 1. De waarde 0 betekent dat alle scores identiek zijn (of dat ze allemaal in dezelfde klasse liggen). De spreiding is dus minimaal. De waarde 1 betekent dat elke mogelijke waarde (of klasse) dezelfde frequentie heeft. De spreiding is dus maximaal.

25. Bereken de spreidingsmaat d voor `geslacht` *m.b.v. R.*

2.3.3 Associatiematen

De covariantie De formule van de covariantie is

$$cov_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \tag{2.1}$$

Een positieve waarde duidt een stijgend lineair verband aan. Een negatieve waarde duidt een dalend lineair verband aan. Een waarde gelijk aan 0 betekent dat er geen lineair verband is tussen X en Y. Verder is deze associatiemaat moeilijk te interpreteren omdat de covariantie, net zoals de variantie, afhankelijk is van de meeteenheden. Het variëert niet tussen vaste grenzen.

De covariantie is gevoelig aan outliers en mag niet gebruikt worden met nominale en ordinale variabelen.

De `R` functie om de covariantie te berekenen is `cov`. Voorbeeld:

```
> cov(myData$score, myData$lengte)
[1] 6.068966
```

De correlatiecoëfficiënt De covariantie is moeilijk te interpreteren en wordt vooral als tussenstap in complexe berekeningen gebruikt. Om het lineair verband tussen twee variabelen te meten gebruik je best de correlatiecoëfficiënt van Pearson[2], vaak afgekort als correlatiecoëfficiënt:

$$r_{XY} = \frac{cov_{XY}}{s_X s_Y}. \tag{2.2}$$

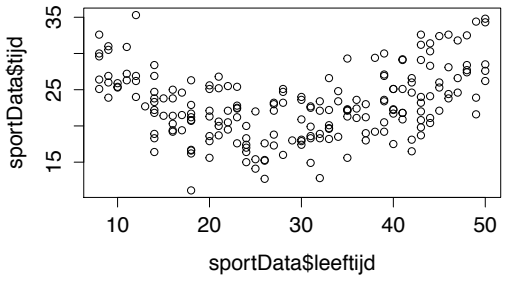
Het is eigenlijk hetzelfde als de covariantie, maar gestandaardiseerd zodanig dat het onafhankelijk van de meeteenheden is en dat het altijd tussen -1 en +1 variëert. De waarde 1 komt overeen met een perfect stijgend lineair verband (alle punten liggen op een stijgende rechte); de waarde -1 komt overeen met een perfect dalend lineair verband (alle punten liggen op een dalende rechte). Een waarde in de buurt van 0 betekent dat er geen lineair verband is tussen de twee variabelen.

De R functie om r_{XY} te berekenen is `cor`. Bv.

```
> cor( myData$gewicht, myData$lengte )
[1] 0.4741137
```

De correlatiecoëfficiënt is gevoelig aan outliers en mag niet gebruikt worden met nominale en ordinale variabelen.

De correlatiecoëfficiënt en de covariantie zijn maten voor het lineair verband. Indien twee variabelen sterk samenhangen, maar niet lineair, dan wordt het verband tussen de twee variabelen niet correct gemeten door de correlatiecoëfficiënt of de covariantie. Voorbeeld: een onderzoeker vraagt 200 personen om 100 meter zo vlug mogelijk te lopen. In het data frame `sportData` (beschikbaar op Ufora) vind je de variabelen `leeftijd`, `gewicht`, `lengte`, `tijd` (tijd om 100 m te lopen, in s), `sport` (aantal uur sport per week), `geslacht` en `type` (basketbal, tennis, voetbal, zwemmen, andere). In Fig. 2.7 zie je het spreidingsdiagram van de variabelen `leeftijd` en `tijd`. Het is duidelijk dat het verband tussen de



Figuur 2.7: Spreidingsdiagram van `leeftijd` en `tijd`.

twee variabelen niet lineair is. De variabele `tijd` is minimaal tussen 20 en 30, maar `tijd` is hoger bij kinderen en vijftigers. Het verband is curvilineair. We berekenen de correlatiecoëfficiënt tussen `leeftijd` en `tijd`:

[2]Karl pearson, 1857–1936

```
> cor( sportData$leeftijd, sportData$tijd )
[1] 0.1673376
```

We bekomen een kleine correlatiecoëfficiënt (0.17). Dit geeft de indruk dat er geen verband is tussen die twee variabelen, maar dat is niet het geval. Er is een vrij sterk verband, toch niet lineair.

26. Bereken de correlatiecoëfficiënt tussen `lengte` *en* `gewicht` *in het data frame* `sportData`*. Gebruik eerst de functie* `cor` *en dan de functie* `cov`*.*

De correlatiecoëfficiënt τ van Kendall Indien je vermoedt (na visuele analyse of omwille van theoretische redenen) dat het verband tussen twee variabelen niet lineair is, maar toch monotoon, dan gebruik je de correlatiecoëfficiënt τ van Kendall[3] (of gewoon τ van Kendall) om de sterkte van het verband te meten. De formule van τ is

$$\tau_{XY} = \frac{\text{aantal concordante paren - aantal discordante paren}}{n(n-1)/2}$$

waar een paar (i, j) concordant is indien

$$x_i > x_j \text{ en } y_i > y_j$$
of
$$x_i < x_j \text{ en } y_i < y_j,$$

en discordant indien

$$x_i > x_j \text{ en } y_i < y_j$$
of
$$x_i < x_j \text{ en } y_i > y_j.$$

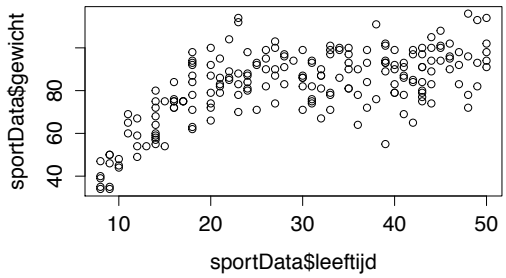
Voorbeeld. We hebben gezien dat het verband tussen `tijd` en `leeftijd` niet lineair is (Fig. 2.7). Maar het is ook niet monotoon want het is eerst dalend en dan stijgend. We mogen de sterkte van het verband tussen `tijd` en `leeftijd` dus niet meten met τ van Kendall. Laten we het verband tussen `leeftijd` en `gewicht` visueel analyseren.

```
> plot( sportData$leeftijd, sportData$gewicht )
```

Je vindt de output in Fig.2.8. Het is op dit spreidingsdiagram duidelijk dat het verband tussen `leeftijd` en `gewicht` monotoon stijgend is. Het is ook gemakkelijk te verklaren: het gewicht stijgt naar gelang kinderen ouder worden; het stijgt nog wanneer ze adolescenten worden en dan blijft het min of meer stabiel tijdens de volwassene jaren, met een lichte tendens om nog zwaarder te worden.

Je kan gemakkelijk τ van Kendall berekenen met de R functie `cor`. Ja, dezelfde functie als de functie die we gebruikten voor r van Pearson! We moeten natuurlijk R zeggen dat we de correlatiecoëfficiënt van Kendall wensen en niet die van Pearson; te dien einde gebruiken we een extra argument tussen de haakjes:

[3] Maurice George Kendall, 1907–1983



Figuur 2.8: Spreidingsdiagram van `leeftijd` en `gewicht`.

```
> cor( sportData$leeftijd, sportData$gewicht, method = "kendall" )
[1] 0.4305121
```

Deze positieve coëfficiënt bevestigt onze visuele analyse: er is een monotoon stijgend verband tussen `leeftijd` en `gewicht`.
R wil niet de correlatiecoëfficiënt van Kendall berekenen indien één van de variabelen niet numeriek is (indien de variabele een factor is). Als we toch de correlatiecoëfficiënt van Kendall willen berekenen, dan moeten we dezelfde truc gebruiken als voor de berekening van de mediaan: de functie `as.numeric`.

27. Bereken τ van Kendall tussen `lengte` *en* `leeftijd` *in het data frame* `sportData`*. Gebruik de functie* `cor`*.*

28. Bereken τ van Kendall tussen `lengte` *en* `motivatie` *in het data frame* `myData`*. Gebruik de functie* `cor`*.*

De regressielijn Indien het verband tussen twee variabelen lineair is, dan kunnen we het representeren d.m.v. een rechte: de regressielijn van Y op X, met vergelijking

$$Y = b_0 + b_1 X.$$

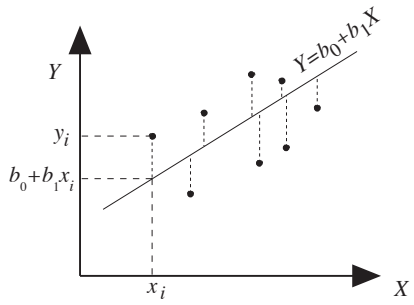
De coëfficiënt b_0 is het intercept van de rechte, i.e., het snijpunt tussen de rechte en de verticale as. De coëfficiënt b_1 is de richtingscoëfficiënt van de rechte. In de statistiek wordt b_1 de regressiecoëfficiënt genoemd. De regressielijn is de best passende rechte. We moeten dus de coëfficiënten b_0 en b_1 zodanig kiezen dat de rechte de puntenwolk inderdaad zo goed mogelijk past. Te dien einde berekenen we de verticale afwijking tussen elk punt (x_i, y_i) in de steekproef en de rechte. Deze afwijking is $y_i - (b_0 + b_1 x_i)$ en wordt residu genoemd (zie Fig. 2.9). We berekenen dan de som van die gekwadrateerde residuen:

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

De best passende rechte is de rechte waarvoor deze som minimaal is. Het is mogelijk te bewijzen dat deze som minimaal is als

$$b_1 = r_{XY} \, \frac{s_Y}{s_X} \quad \text{en} \quad b_0 = \overline{y} - b_1 \overline{x}.$$

Laten we dit in R illustreren met de functie `lm`. De functie `lm` (de naam staat voor *linear model*) heeft één argument nodig: een R formula waarmee we specificeren voor welke variabelen we de regressielijn aanvragen.



Figuur 2.9: Een residu is de verticale afwijking tussen een punt (x_i, y_i) en de regressielijn.

```
> lm( formula = myData$gewicht ~ myData$lengte )

Call:
lm(formula = myData$gewicht ~ myData$lengte)

Coefficients:
  (Intercept)  myData$lengte
      -27.199          0.592
```

De R formula "`myData$gewicht ~ myData$lengte`" bestaat uit twee delen, gescheiden door het symbool "`~`" (tilde). Aan de linkerkant van de tilde vind je de Y-variabele; aan de rechterkant, de X-variabele. Het onderscheid is belangrijk want we meten de afwijkingen parallel aan de as van de Y-variabele.

De output van het commando "`lm(formula = myData$gewicht ~ myData$lengte)`" bestaat uit twee delen. De twee regels

```
Call:
lm(formula = myData$gewicht ~ myData$lengte)
```

gevolgd door drie regels

```
Coefficients:
  (Intercept)  myData$lengte
      -27.199          0.592
```

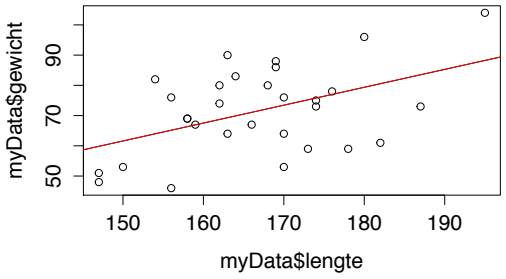
29. Verifiëer of de rechte van Fig. 2.10 door het punt $(\overline{x}, \overline{y})$ gaat.

Het eerste deel is niet echt interessant. In het tweede deel vind je het intercept ($b_0 = -27.199$) en de coëfficiënt van de variabele `lengte` ($b_1 = 0.592$). De vergelijking van de regressielijn van `gewicht` op `lengte` is dus

30. Bereken de vergelijking van de regressielijn van `gewicht` op `lengte` in het data frame `sportData`.

$$\texttt{gewicht} = -27.199 + 0.592\ \texttt{lengte}.$$

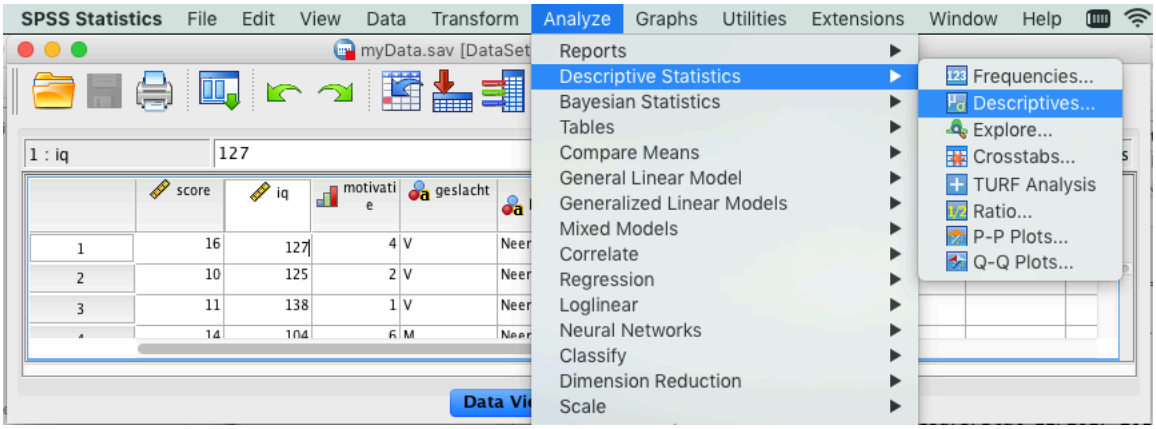
Deze lijn wordt gerepresenteerd op Fig. 2.10.



Figuur 2.10: Regressielijn van `gewicht` op `lengte`.

2.4 SPSS

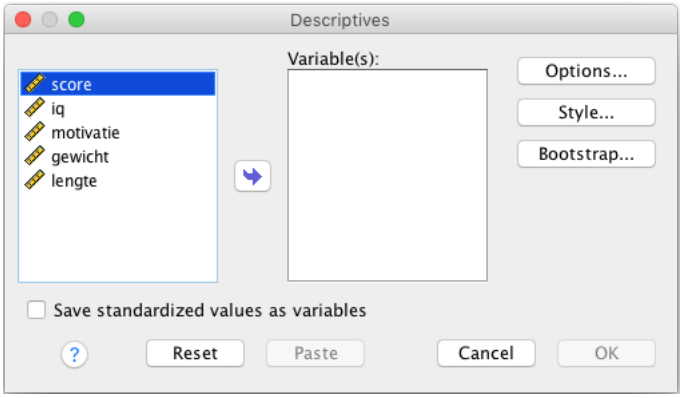
Alhoewel de focus in de cursus duidelijk op `R` en niet op SPSS ligt, gaan we kort zien hoe je reductiematen kan berekenen met SPSS. In het menu "Analyze" selecteer je "Descriptive Statistiscs ▶ Descriptives ..." (zie Fig. 2.11). Een



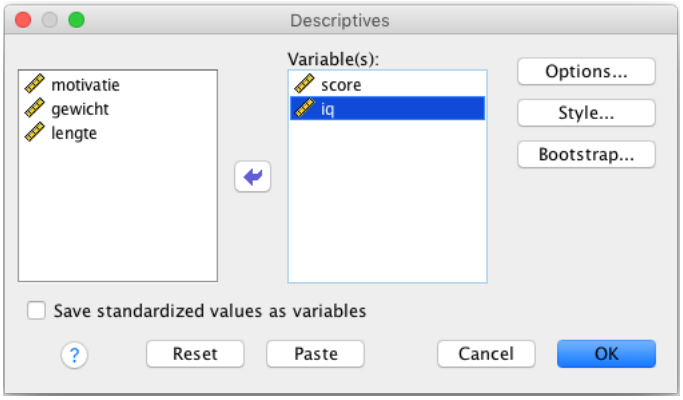
Figuur 2.11: Reductiematen berekenen met SPSS

venster gaat open (Fig. 2.12). In het linkerdeel van het venster zie je de lijst van alle numerieke variabelen. Click op één van die variabelen (bv. score) en dan op het pijltje ➥. Die variabele verschuift naar het rechterdeel van het venster. Doe hetzelfde met elke variabele waarvoor je reductiematen wenst te berekenen (bv. iq). Het venster ziet er zoals Fig. 2.13 uit.

Click nu op "OK" en een nieuw venster gaat open met de resultaten van de berekeningen in een tabel. Deze tabel bevat een rij per geselecteerde variabele en elke rij bevat de kleinste waarde, de grootste waarde, het gemiddelde (Mean) en de standaarddeviatie (Std. Deviation).



Figuur 2.12: Variabelen selecteren om reductiematen te berekenen



Figuur 2.13: Variabelen selecteren om reductiematen te berekenen

2.5 Oplossingen

11) Wat is de frequentieverdeling van `roken`? Gebruik een R commando.

Oplossing:

```
> table( myData$roken )

  Ja Neen
  10   20
```

12) Wat is de relatieve frequentieverdeling van `roken`? Gebruik een R commando.

Oplossing:

```
> table( myData$roken ) / dim( myData )[1]

       Ja          Neen
0.3333333 0.6666667
```

Je kan ook `prop.table` gebruiken:

```
> prop.table( table( myData$roken ) )

       Ja      Neen
0.3333333 0.6666667
```

13) Vraag de bivariate frequentieverdeling aan voor `roken` en `motivatie`.

Oplossing:

```
> table(myData$roken, myData$motivatie)

      1 2 3 4 5 6 7
  Ja  1 1 0 0 6 2 0
  Neen 4 4 2 1 3 3 3
```

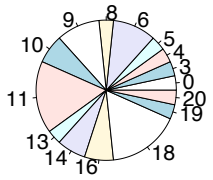
14) Vraag de relatieve bivariate frequentieverdeling aan voor `geslacht` en `opleiding`.

Oplossing:

```
> prop.table(table( myData$geslacht , myData$opleiding ))

          ped        psy        soc
  M 0.23333333 0.23333333 0.00000000
  V 0.10000000 0.36666667 0.06666667
```

15) Teken een cirkeldiagram voor de frequentieverdeling van score m.b.v. R.

Oplossing: Het geschikte commando is `pie( table( myData$score ) )` en de output is



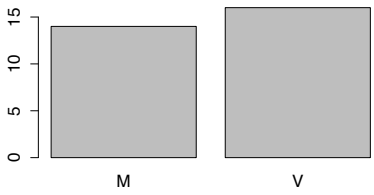
Je had dit cirkeldiagram beter niet getekend. Het is niet echt geschikt voor variabelen van interval, ratio of absoluut meetniveau omdat de volgorde en de afstanden helemaal niet duidelijk zijn. Vergeet niet kritisch na te denken of een statistische techniek adequaat is.

16) Teken een staafdiagram voor de frequentieverdeling van `geslacht` m.b.v. R.

Oplossing: Het geschikte commando is

```
> barplot( table( myData$geslacht ) )
```

en de output is



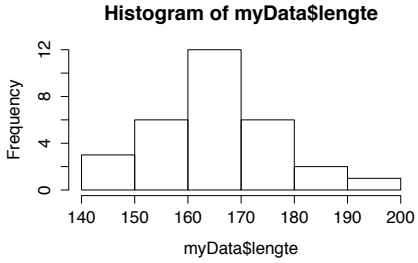
Je had ook een cirkeldiagram kunnen gebruiken voor `geslacht`.

17) Teken een histogram voor de frequentieverdeling van `lengte`, met 5 klassen, m.b.v. R.

Oplossing: Het geschikte commando is

```
> hist( myData$lengte, breaks = 5 )
```

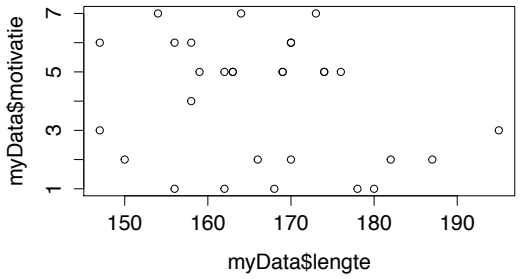
en de output is



Alhoewel we 5 klassen hebben gevraagd, heeft R 6 klassen gebruikt. De reden is duidelijk: met 6 klassen zijn de klassegrenzen ronde getallen. Met 5 klassen was het zeker veel minder aantrekkelijk.

18) Teken een spreidingsdiagram voor `lengte` en `motivatie` m.b.v. R.

Oplossing: Het geschikte commando is `plot( x = myData$lengte, y = myData$motivatie)` en de output is



Oops! De variabele `motivatie` is van ordinaal meetniveau. Dit spreidingsdiagram is *geen* geschikte grafische voorstelling voor variabelen van nominaal en ordinaal meetniveau.

19) Bereken het gemiddelde van `lengte` m.b.v. R.

Oplossing:

```
> mean( myData$lengte )
[1] 166.6667
```

20) Bereken de mediaan van `geslacht` m.b.v. R.

Oplossing: Je mag de mediaan van een nominale variabele niet berekenen: het is zinloos. Als je het toch doet, krijg je een foutmelding:

```
> median( myData$geslacht )

Error in median.default(myData$geslacht) :  need numeric data
```

Let op, R zal niet altijd zo'n melding geven als je iets zinloos doet. Had je "man" en "vrouw" met 0 en 1 gecodeerd, dan had R de mediaan wel berekend en het zou toch zinloos zijn.

21) Bereken de modus van `geslacht` m.b.v. R.

Oplossing:

```
> table( myData$geslacht )

 M  V
14 16
```

De hoogste frequentie is 16 en komt overeen met `V`. De modus is dus 'vrouw'.

22) Bereken de variantie van `lengte` m.b.v. R.

Oplossing:

```
> var(myData$lengte)
[1] 129.8161
```

Na afronding $s^2_{\texttt{lengte}} = 130$.

23) Bereken de standaarddeviatie van `lengte`. Eens met de functie `var` en een met de functie `sd`.

Oplossing:

```
> sqrt(var(myData$lengte))
[1] 11.39369
> sd(myData$lengte)
[1] 11.39369
```

24) Bereken de interkwartiele afstand van `lengte` m.b.v. R.

Oplossing:

```
> IQR( myData$lengte )
[1] 15.5
```

25) Bereken de spreidingsmaat d voor `geslacht` m.b.v. R.

Oplossing:

```
> table(myData$geslacht)

 M  V
14 16
```

Dus $f_{mo} = 16$ en $p = 2$. De maat d is dan gelijk aan

```
> (1 - ( 16/30 ) ) / (1 - ( 1/2 ) )
[1] 0.9333333
```

26) Bereken de correlatiecoëfficiënt tussen `lengte` en `gewicht` in het data frame `sportData`. Gebruik eerst de functie `cor` en dan de functie `cov`.

Oplossing:

```
> cor( sportData$gewicht, sportData$lengte )
[1] 0.566555
> cov(sportData$gewicht, sportData$lengte )/(sd(sportData$gewicht)
   *sd(sportData$lengte))
[1] 0.566555
```

27) Bereken τ van Kendall tussen `lengte` en `leeftijd` in het data frame `sportData`. Gebruik de functie `cor`.

Oplossing:

```
> cor( sportData$lengte, sportData$leeftijd, method="kendall" )
[1] 0.2526885
```

28) Bereken τ van Kendall tussen `lengte` en `motivatie` in het data frame `myData`. Gebruik de functie `cor`.

Oplossing: De variabele `motivatie` is ordinaal. We moeten dus `as.numeric` gebruiken.

```
> cor( myData$lengte, as.numeric( myData$motivatie ), method = "kendall" )
[1] -0.1548777
```

29) Verifiëer of de rechte van Fig. 2.10 door het punt $(\overline{x}, \overline{y})$ gaat.

Oplossing: Je kan het approximatief verifiëren op Fig. 2.10. Je kan het ook exact verifiëren a.d.h.v. de vergelijking van de rechte:

$$\texttt{gewicht} = -27.199 + 0.592\ \texttt{lengte}.$$

Je berekent eerst het gemiddelde van `gewicht` en `lengte`.

```
> mean( myData$gewicht )
[1] 71.46667
> mean( myData$lengte )
[1] 166.6667
```

Dan vul je die twee getallen in in de vergelijking van de rechte en je gaat na of de gelijkheid klopt:

$$71.46667 \stackrel{?}{=} -27.199 + 0.592 \times 166.6667.$$

Ja, het klopt want $-27.199 + 0.592 \times 166.6667$ is gelijk aan $71,4676864$ op verwaarloosbare afrondingsfouten na. De rechte van Fig. 2.10 gaat dus door het punt $(\overline{x}, \overline{y})$.

30) Bereken de vergelijking van de regressielijn van `gewicht` op `lengte` in het data frame `sportData`.

Oplossing:

```
> lm( formula = sportData$gewicht ~ sportData$lengte )

Call:
lm(formula = sportData$gewicht ~ sportData$lengte)

Coefficients:
     (Intercept)  sportData$lengte
        -19.0151            0.5917
```

De vergelijking van de regressielijn van `gewicht` op `lengte` is dus

$$\texttt{gewicht} = -19.0151 + 0.5917\ \texttt{lengte}.$$

Hoofdstuk 3

Kansrekenen

3.1 Toevalsvariabelen en kansverdelingen

Een *toevalsproces* is een proces waarvan de *uitkomst* onvoorspelbaar is. Bijvoorbeeld, als je een persoon bij toeval kiest en zijn I.Q. meet, kan je niet zijn I.Q. voorspellen. Het trekken van een steekproef van elementen uit een populatie is ook een toevalsproces. Bijvoorbeeld, als je de oxytocine niveaus in het speeksel van vijf proefpersonen meet, kan je niet voorspellen wat het resultaat gaat zijn.

Een *gebeurtenis* (voor een toevalsproces) is een mogelijke uitkomst voor dat toevalsproces of een reeks (verzameling) mogelijke uitkomsten.

Een gebeurtenis A realiseert zich (of doet zich voor) als één van de uitkomsten in A zich realiseert.

Voorbeeld : als je een persoon bij toeval kiest en zijn I.Q. meet, "115" is een gebeurtenis. "Meer dan 115" is ook een gebeurtenis; het is de verzameling I.Q.'s groter dan of gelijk aan 115. Deze verzameling wordt vaak door $[115, +\infty[$ aangeduid.

3.1.0.1 De complementaire gebeurtenis

Laat A een gebeurtenis zijn. We noteren A^* de *complementaire* gebeurtenis van A. Het is de gebeurtenis die zich voordoet als en slechts als A zich niet voordoet. Bv., de gebeurtenissen "`iq` > 100" en "`iq` ≤ 100" zijn complementair. De complementaire gebeurtenis van de complementaire gebeurtenis is de oorspronkelijke gebeurtenis.

$$(A^*)^* = A.$$

Men zegt ook dat A en A^* complementair zijn.

3.1.1 Toevalsvariabele

Een *toevalsvariabele* (ook kansveranderlijke) is een variabele waarvan de waarde in een toevalsproces onvoorspelbaar is.

Voorbeeld : bij trekking van een persoon is zijn I.Q. niet voorspelbaar. Daarom is de variabele `iq` een toevalsvariabele. Bij de worp van een dobbelsteen is de uitkomst ook een toevalsvariabele.

De waarde van een toevalsvariabele in een bepaalde herhaling van een toevalsproces wordt een *realisatie* genoemd.

Voorbeeld : in een steekproef van 3 personen zijn de geobserveerde waarden voor de variabele `iq` 115, 98 en 107. Die drie getallen zijn drie realisaties van de toevalsvariabele `iq`.

Zoals de variabelen in de beschrijvende statistiek, kunnen de toevalsvariabelen continu of discreet zijn. Ze kunnen ook van verschillend meetniveau zijn : nominaal, ordinaal, interval, ratio of absoluut.

3.1.2 Kansen

3.1.2.1 Het begrip "kans"

De *kans* van een gebeurtenis A (symbool $P(A)$) bij een toevalsproces wordt gedefiniëerd als de relatieve frequentie van deze gebeurtenis als we het toevalsproces eindeloos zouden herhalen. Formeel,

$$P(A) = \lim_{n\to\infty} \frac{f_A}{n},$$

waar f_A de frequentie van A is.

Bij trekking van een persoon uit een populatie, als je het I.Q. van deze persoon meet, is de kans dat zijn I.Q. boven 110 ligt gelijk aan de relatieve frequentie van de gebeurtenis "I.Q. $>$ 110" als je eindeloos een persoon uit de populatie zou trekken.

Een kans is een relatieve frequentie (met n oneindig). Het is dus een getal tussen 0 en 1.

3.1.2.2 Afhankelijkheid

Twee gebeurtenissen A en B zijn *afhankelijk* als de realisatie van de ene de kans van de andere beïnvloedt. Twee gebeurtenissen zijn *onafhankelijk* als ze niet afhankelijk zijn.

Voorbeeld 1. We beschouwen de gebeurtenissen HS (hooggeschoold) en $V2$ (verdient meer dan 2000€). We weten dat hooggeschoolden vaak hogere inkomen hebben en, omgekeerd, mensen met hoge inkomen zijn vaak hooggeschoold.[1] Je trekt een willekeurige volwassene uit de Vlaamse populatie. Wat is de kans dat deze persoon meer dan 2000€ verdient? M.a.w. wat is $P(V2)$? Je weet het natuurlijk niet. Het is gelijk aan de proportie van Vlamingen die meer dan 2000€ verdienen. Stel nu dat deze persoon hooggeschoold is. Dus de gebeurtenis HS realiseerde zich. Je vermoedt automatisch dat deze persoon meer dan 2000€ verdient (omdat hoge diploma's vaak gepaard gaan met hoge

[1] Er zijn veel uitzonderingen, maar dit is wel correct in doorsnee.

inkomens). De kans dat deze persoon meer dan 2000€ verdient is dus nu groter dan de corresponderende proportie in de Vlaamse populatie. De realisatie van HS heeft dus de kans op $V2$ beïnvloed. Deze twee gebeurtenissen zijn dus afhankelijk.

Voorbeeld 2. Bij trekking van een persoon uit de Vlaamse populatie, noemen we V de gebeurtenis "de proefpersoon is een vrouw" en R de gebeurtenis "de proefpersoon rookt". We weten dat roken frequenter is bij mannen dan bij vrouwen. Je trekt een willekeurige volwassene uit de Vlaamse populatie. Wat is de kans dat deze persoon een vrouw is? Het is duidelijk 0.5 (omdat 50% van de Vlamingen vrouwen zijn). Stel nu dat deze persoon rookt. Dus de gebeurtenis R realiseerde zich. Je vermoedt automatisch dat deze persoon een man is (omdat roken frequenter is bij mannen dan bij vrouwen). De kans dat deze persoon een vrouw is, is dus nu kleiner dan 0.5. De realisatie van R heeft dus de kans op V beïnvloed. Deze twee gebeurtenissen zijn dus afhankelijk.

Voorbeeld 3. Bij trekking van een persoon uit de Vlaamse populatie, noemen we V de gebeurtenis "de proefpersoon is een vrouw" en B de gebeurtenis "de proefpersoon heeft een bril". We weten dat een bril dragen even frequent is bij mannen en bij vrouwen. Je trekt een willekeurige volwassene uit de Vlaamse populatie. Wat is de kans dat deze persoon een vrouw is? Het is nogmaals 0.5. Stel nu dat deze persoon een bril draagt (de gebeurtenis B realiseerde zich). Heeft dit een invloed of je vermoeden dat deze persoon een vrouw is? Neen, omdat brillen even frequent zijn bij vrouwen en bij mannen. De kans dat deze persoon een vrouw is, is dus nu gelijk aan 0.5. De realisatie van B heeft de kans op V niet beïnvloed. Deze twee gebeurtenissen zijn dus onafhankelijk.

Voorbeeld 4. Bij trekking van een persoon uit de Vlaamse populatie, noemen we BV de gebeurtenis "de proefpersoon heeft borstvoeding gekregen" en ASS de gebeurtenis "de proefpersoon heeft autismespectrumstoornis". We weten niet of autismespectrumstoornis al dan niet frequenter is bij personen die borstvoeding hebben gekregen. We kunnen dus niet weten of die twee gebeurtenissen al dan niet afhankelijk zijn. Dit kan wel onderzocht worden.

3.1.3 Kansverdeling

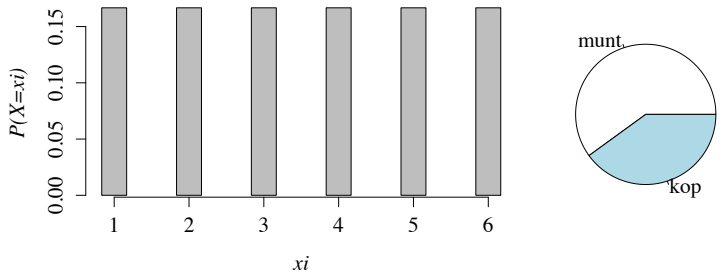
Laat X een discrete toevalsvariabele zijn, met p mogelijke waarden. Laat $x_1, x_2, \ldots, x_p$ de verschillende mogelijke waarden van X zijn, geordend van de kleinste naar de grootste (als de variabele tenminste van ordinaal meetniveau is). De kans op de gebeurtenis "$X = x_i$" wordt $P(X = x_i)$ genoteerd.

De *kansverdeling* van de discrete toevalsvariabele X is een tabel met twee kolommen (of rijen), zoals een frequentieverdeling. Een kolom bevat de waarden $x_1, x_2, \ldots, x_p$ en de andere bevat de kansen $P(X = x_1), P(X = x_2), \ldots, P(X = x_p)$. Voorbeeld: de worp van een dobbelsteen. $X =$ "aantal ogen."

Dezelfde grafische voorstellingen als voor relatieve frequentieverdelingen kunnen met kansverdelingen gebruikt worden (zie Fig. 3.1).

Aantal ogen	kans
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Tabel 3.1: Kansverdeling van de toevalsvariabele “aantal ogen.”



Figuur 3.1: Lijndiagram voor de variabele “aantal ogen” met een zuivere dobbelsteen en cirkeldiagram voor de nominale variabele “munt” of “kop” met een onzuivere munt.

3.1.4 Dichtheidsfunctie

Laat X een continue toevalsvariabele zijn. Omwille van de continuïteit is het niet meer mogelijk aan elke waarde een verschillend symbool toe te kennen. Om een bepaalde waarde van X aan te duiden zullen we gewoon x, x' of x'' gebruiken. Ook soms x_1 of x_2 maar hier representeren x_1 en x_2 niet meer de kleinste en tweede kleinste waarden. De symbolen x_1 en x_2 representeren gewoon twee willekeurige waarden.

Zoals bij discrete toevalsvariabelen “$X = x$” is een gebeurtenis en $P(X = x)$ is zijn kans. Voorbeeld : laat X het gewicht van een willekeurige persoon zijn. X is een toevalsvariabele omdat je zijn waarde niet kan voorspellen. Laat $x = 75.8$ kg. Dan $P(X = x)$ is de kans dat een willekeurige persoon 75.8 kg weegt. Omwille van de continuïteit is deze kans nul. Het is wel mogelijk dat iemand exact 75.8 kg weegt, maar de kans dat je zo’n persoon trekt is oneindig klein en is bijgevolg nul. Dit geldt voor elke continue variabele X en elk getal x:

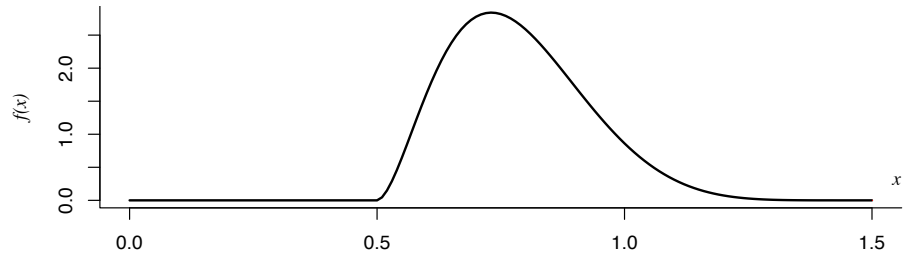
$$P(X = x) = 0 \text{ voor alle } X \text{ en alle } x.$$

Het is dus onmogelijk de kansverdeling van een continue variabele te definiëren.

Om dit probleem te overkomen hebben de wiskundigen een andere techniek ontwikkeld: de dichtheidsfunctie (of densiteitsfunctie). Het is een functie (een

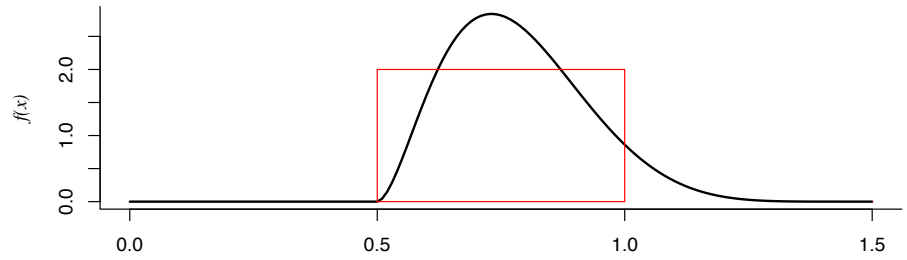
curve) die in zich niet echt interpreteerbaar is, maar ze is zo geconstrueerd dat oppervlakten onder de curve kansen representeren. Het symbool voor de dichtheidsfunctie is vaak f (zoals in deze cursus) maar ook soms p.

In Fig. 3.2 vind je een voorbeeld van een dichtheidsfunctie: de dichtheidsfunctie van de reactietijd bij een bepaald experiment. De totale oppervlakte



Figuur 3.2: Dichtheidsfunctie van de reactietijd, in s.

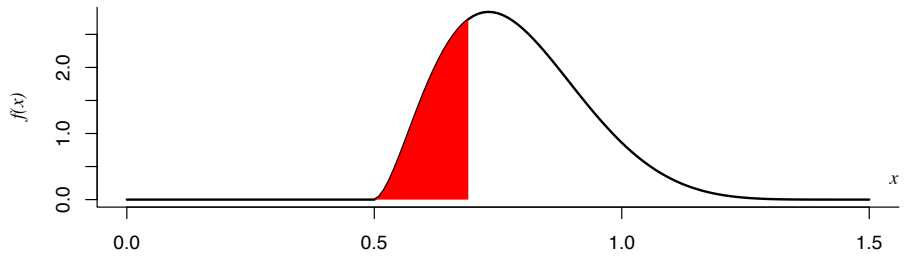
onder deze curve (tussen de curve en de horizontale as) is gelijk aan 1. Om dit te controleren gaan we een rechthoek op dezelfde grafiek tekenen, met ongeveer dezelfde oppervlakte als de totale oppervlakte onder de curve (Fig. 3.3). Wat is de oppervlakte van deze rechthoek? Het is $0.5 \times 2 = 1$.



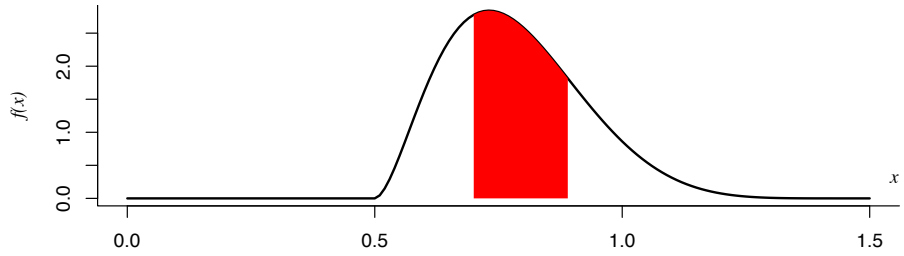
Figuur 3.3: Dichtheidsfunctie van de reactietijd, in s. De oppervlakte van de rode rechthoek is $0.5 \times 2 = 1$.

We tekenen nu de oppervlakte onder de curve tussen 0.5 en 0.7 (Fig. 3.4). Het is mogelijk te bewijzen dat deze oppervlakte gelijk is aan de kans dat de reactietijd tussen 0.5 en 0.7 ligt. De wiskundigen hebben ook een techniek ontwikkeld om deze oppervlakte te berekenen: de integraal. Je hoeft deze techniek niet te begrijpen of te kunnen gebruiken: er zijn softwarepaketten, zoals R, die zulke kansen (of integralen) voor jou berekenen. In dit geval, is de oppervlakte—en dus $P(0.5 < X \leq 0.7)$—gelijk aan 0.29. We zullen later in deze cursus zien hoe je `R` kunt gebruiken om deze kans te bekomen.

We tekenen nu de oppervlakte onder de curve tussen 0.7 en 0.9 (Fig. 3.5). Zoals hierboven is het mogelijk te bewijzen dat deze oppervlakte gelijk is aan $P(0.7 < X \leq 0.9)$. Met een softwarepakket is het mogelijk deze oppervlakte te berekenen: $P(0.7 < X \leq 0.9) = 0.50$.



Figuur 3.4: Dichtheidsfunctie van de reactietijd, in s. De oppervlakte onder de curve tussen 0.5 en 0.7 is gelijk aan 0.29.



Figuur 3.5: Dichtheidsfunctie van de reactietijd, in s. De oppervlakte onder de curve tussen 0.7 en 0.9 is gelijk aan 0.50.

Het is mogelijk te bewijzen dat, in het algemeen, de kans $P(a < X \leq b)$ gelijk is aan de oppervlakte onder de dichtheidsfunctie van X, tussen a en b. Dit geldt voor alle continue toevalsvariabelen.

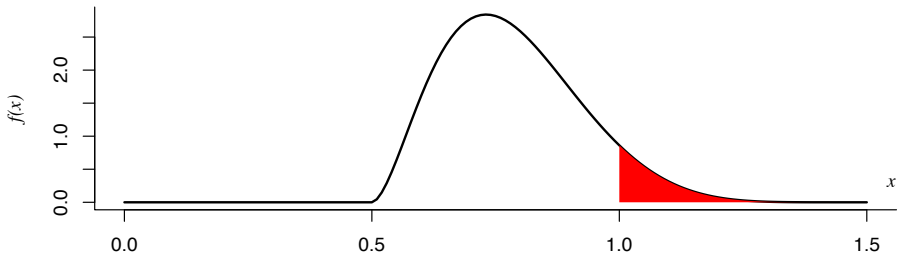
Alhoewel je de integraaltechniek niet beheerst, kan je toch oppervlakten visueel schatten; het tekenen van een dichtheidsfunctie en het arceren van bepaalde oppervlakten kunnen je dus helpen om bepaalde kansen te schatten. Bv., wat is de kans $P(0 < X \leq 0.5)$? Je ziet onmiddellijk op Fig. 3.2 dat de oppervlakte onder de curve tussen 0 en 0.5 gelijk is aan 0. Nog een voorbeeld: wat is de kans $P(1 < X \leq 1.5)$? Op Fig. 3.6 zie je dat deze kans ongeveer een tiende van de totale oppervlakte is. De kans $P(1 < X \leq 1.5)$ is dus ongeveer 10%. Met een softwarepakket kom ik natuurlijk een preciezere waarde: 8%.

31. Wat is de kans dat de reactietijd zich niet tussen 1 en 1.5 s bevindt ?

Met discrete variabelen gebruik je kansverdelingen en bij continue variabelen dichtheidsfuncties. In de“gewone” taal, spreekt men gewoon van verdeling, alhoewel het helemaal niet hetzelfde is.

3.1.5 Bivariate kansverdelingen

Als we in meer dan één toevalsvariabele geïnteresseerd zijn, kunnen we ze afzonderlijk analyseren, zoals in de vorige paragrafen, of tegelijkertijd. In deze cursus zullen we slechts het eenvoudigste geval beschouwen : het geval van twee toevalsvariabelen.



Figuur 3.6: Dichtheidsfunctie van de reactietijd, in s. De oppervlakte onder de curve tussen 1.0 en 1.5 is gelijk aan 0.08.

Hier gaan we dezelfde technieken gebruiken als in bivariate beschrijvende statistiek.

Een bivariate kansverdeling is gewoon een speciale bivariate relatieve frequentieverdeling (met $n \to \infty$). Het kan in de vorm van een tabel voorgesteld worden.

Voorbeeld : bij trekking van een willekeurige volwassene kan men de twee toevalsvariabelen $X =$"opleiding vader" en $Y =$"opleiding" analyseren. Elke variabele kan de waarden 0 (geen diploma), 1 (diploma basis onderwijs), 2 (diploma secundair onderwijs), 3 (diploma hoger onderwijs) nemen. Dus $x_1 = y_1 = 0$, $x_2 = y_2 = 1$, ... Tabel 3.2 stelt de bivariate kansverdeling van de variabelen X en Y voor.

		Y			
		0	1	2	3
	0	0.07	0.19	0.18	0.09
X	1	0.03	0.06	0.11	0.04
	2	0.02	0.05	0.05	0.04
	3	0.01	0.01	0.02	0.03

Tabel 3.2: Kansverdeling van de variabelen $X =$ "opleiding vader" en $Y =$ "opleiding."

Elke cel van de tabel bevat de kans op de overeenkomende gebeurtenis. Bijvoorbeeld, $P(X = 1 \text{ en } Y = 3) = 0.04$. De algemene notatie voor één van de cellen is

$$P(X = x_i \text{ en } Y = y_j) \text{ of } P(X = x_i, Y = y_j).$$

Als de context duidelijk is, schrijven we soms gewoon $P(x_i, y_j)$. Natuurlijk is de som van de kansen van alle gebeurtenissen (van alle cellen) gelijk aan 1. Het aantal mogelijke waarden van X en Y hoeven niet identiek te zijn. We noemen p het aantal mogelijke waarden van X en q het aantal mogelijke waarden van Y. De som van de waarden in een rij geeft ons de kans op de overeenkomende waarde van X.

32. Wat zijn de waarden van p en q in het voorbeeld van Tabel 3.2 ?

$$\sum_{j=1}^{q} P(X = x_i, Y = y_j) = P(X = x_i).$$

Bvb.,

$$P(X = 2) = 0.02 + 0.05 + 0.05 + 0.04 = 0.16.$$

De som van een kolom geeft ons de kans op de overeenkomende waarde van Y.

$$\sum_{i=1}^{p} P(X = x_i, Y = y_j) = P(Y = y_j).$$

Bvb.,

$$P(Y = 3) = 0.09 + 0.04 + 0.04 + 0.03 = 0.20.$$

33. Wat is $\sum_{i=1}^{3} P(X = x_i, Y = y_2)$ *in Tabel 3.2 ?*

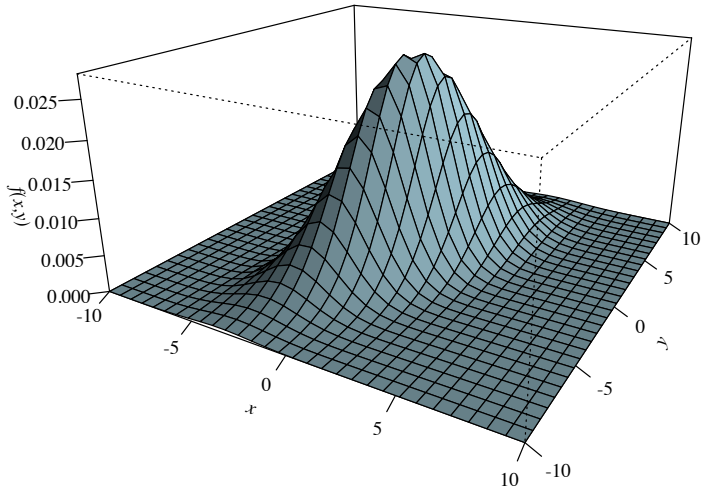
De kansen $P(X = x_i)$ (resp. $P(Y = y_j)$) zijn de marginale kansen van de waarden x_i (resp. y_j).

3.1.6 Bivariate dichtheidsfunctie

Met continue toevalsvariabelen kan men de kansverdeling niet in de vorm van een tabel presenteren want

$$P(X = x \text{ en } Y = y) = 0$$

voor alle x en y. Maar het is mogelijk bivariate dichtheidsfuncties te definiëren. In Fig.3.7 vind je een voorbeeld van een bivariate dichtheidsfunctie.



Figuur 3.7: Een bivariate dichtheidsfunctie

3.1.7 Afhankelijke toevalsvariabelen

We hebben het begrip *afhankelijke gebeurtenissen* al gezien (zie par. 3.1.2.2). Nu gaan we van afhankelijke of onafhankelijke *toevalsvariabelen* spreken. Het is

niet helemaal hetzelfde maar het is gewoon een aanpassing van het begrip van afhankelijke gebeurtenissen. Intuïtief weten we dat de variabelen gewicht (X) en lengte (Y) van een willekeurige persoon niet onafhankelijk zijn. Iemand die groot is zal ook vaak zwaar zijn. Het is geen regel maar het is dikwijls zo. Hoe kunnen we dit preciezer maken ? We gaan de definitie van afhankelijkheid voor variabelen baseren op de definitie van afhankelijkheid voor gebeurtenissen.

3.1.7.1 Discrete variabelen

Twee discrete toevalsvariabelen X en Y zijn *onafhankelijk* als de gebeurtenissen

$$\text{“}X = x_i\text{” en “}Y = y_j\text{”}$$

onafhankelijk zijn, voor alle mogelijke combinaties van i en j. Als ze niet onafhankelijk zijn, dan zegt men dat ze *afhankelijk* zijn. In de praktijk, om na te gaan of twee discrete toevalsvariabelen onafhankelijk zijn, moeten we dus verifiëren of de gelijkheid

$$P(X = x_i \text{ en } Y = y_j) = P(X = x_i)\, P(Y = y_j),$$

geldt voor alle mogelijke combinaties van i en j.

Voorbeeld. Zijn de variabelen X en Y in Tabel 3.2 onafhankelijk ? Om die vraag te beantwoorden moeten we nagaan of $P(X = x_i, Y = y_j) = P(X = x_i)\, P(Y = y_j)$ voor alle combinaties van i en j. Laten we beginnen met $i = 1$ en $j = 1$. Is het waar dat $P(X = x_1, Y = y_1) = P(X = x_1)\, P(Y = y_1)$? We moeten eerst $P(X = x_1)$ en $P(Y = y_1)$ berekenen. $P(X = x_1)$ is de som van de eerste rij. Dus $P(X = x_1) = 0.53$. $P(Y = y_1)$ is de som van de eerste kolom. Dus $P(Y = y_1) = 0.13$. We kunnen $P(X = x_1, Y = y_1)$ rechtstreeks in de tabel lezen : het is 0.07. Nu kunnen we dus verifiëren of $P(X = x_1, Y = y_1) = P(X = x_1)\, P(Y = y_1)$. En we vinden $0.07 \neq 0.53 \times 0.13$. De gelijkheid geldt niet en we kunnen besluiten dat X en Y afhankelijk zijn.

Een ander voorbeeld. In Tabel 3.3 vind je de bivariate kansverdeling van twee variabelen X en Y van interval meetniveau. Zijn ze afhankelijk of onafhankelijk ? We gaan na of $P(X = x_i, Y = y_j) = P(X = x_i)\, P(Y = y_j)$ en we doen dit eerst

		Y		
		1	2	3
	0	0.06	0.02	0.12
X	1	0.09	0.04	0.17
	2	0.15	0.04	0.31

Tabel 3.3: Kansverdeling van twee variabelen X en Y

met $i = 1$ en $j = 1$. We vinden

$$P(X = x_1, Y = y_1) = 0.06 = 0.2 \times 0.3 = P(X = x_1)\, P(Y = y_1).$$

Het is niet genoeg dat "$X = x_1$" en "$Y = y_1$" onafhankelijk zijn om te zeggen dat X en Y onafhankelijk zijn. Dit moet het geval zijn voor alle combinaties van i en j. We gaan dus verder met $i = 1$ en $j = 2$. We vinden

$$P(X = x_1, Y = y_2) = 0.02 = 0.2 \times 0.1 = P(X = x_1)\, P(Y = y_2).$$

Dat klopt nog. Nu $i = 1$ en $j = 3$.

$$P(X = x_1, Y = y_3) = 0.12 = 0.2 \times 0.6 = P(X = x_1)\, P(Y = y_3).$$

$i = 2$ en $j = 1$.

$$P(X = x_2, Y = y_1) = 0.09 = 0.3 \times 0.3 = P(X = x_2)\, P(Y = y_1).$$

$i = 2$ en $j = 2$.

$$P(X = x_2, Y = y_2) = 0.04 \neq 0.3 \times 0.1 = P(X = x_2)\, P(Y = y_2).$$

Hier klopt het niet meer. De gebeurtenissen "$X = x_2$" en "$Y = y_2$" zijn afhankelijk. Derhalve zijn de variabelen X en Y afhankelijk.

3.1.7.2 Continue variabelen

Twee continue toevalsvariabelen X en Y zijn *onafhankelijk* als de gebeurtenissen

$$\text{"}X \leq x\text{" en "}Y \leq y\text{"}$$

onafhankelijk zijn, voor alle mogelijke combinaties van x en y. Als ze niet onafhankelijk zijn, dan zegt men dat ze *afhankelijk* zijn. Bijgevolg, om na te gaan of twee continue toevalsvariabelen onafhankelijk zijn, moet je verifiëren of

$$f_{XY}(x, y) = f(x)f(y)$$

voor alle mogelijke combinaties van x en y. In de praktijk is dit moeilijk te doen omdat het aantal combinaties oneindig is bij continue variabelen. De techniek om dit te doen wordt in deze cursus niet gezien.

34. Zijn de variabelen `gewicht` *en* `lengte` *van het data frame* `myData` *(zie Rubr. 2.3.3, blz. 38) onafhankelijk ?*

3.1.8 Reductietechnieken

Zoals in beschrijvende statistieken, gaan we proberen de verdelingen van de toevalsvariabelen in één of enkele getallen samen te vatten. Deze getallen zijn maten van centrale tendentie of van spreiding. Ze worden meestal door een griekse letter aangeduid. In beschrijvende statistieken gebruikt men meestal gewone letters.

3.1.8.1 Discrete toevalsvariabelen

Kansverdelingen van discrete toevalsvariabelen zijn speciale relatieve frequentieverdelingen (met $n \to \infty$). We kunnen dus alle technieken van Rubr. 2.3 gebruiken. De formules moeten toch licht aangepast worden. We zullen maar twee van de reductietechnieken analyseren : het rekenkundig gemiddelde en de variantie. De anderen (modus, mediaan, interkwartiele afstand, ...) worden niet vaak met kansverdelingen gebruikt en kunnen gemakkelijk aangepast worden.

De verwachting Het gemiddelde in een steekproef is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.1}$$

Het gemiddelde van de kansverdeling van X is het gemiddelde van X in een steekproef met n oneindig. Om die te kunnen berekenen gebruiken we deze formule:

$$E(X) = \sum_{i=1}^{p} P(X = x_i)\, x_i.$$

Het gemiddelde van een toevalsvariabele wordt de *verwachting* of populatiegemiddelde genoemd met symbool $E(X)$. Het symbool E komt van het latijn woord voor verwachting : Expectatio. In plaats van het symbool $E(X)$ gebruikt men vaak μ_X of gewoon μ als de context duidelijk is.
Merk op dat x_i in de formule van de verwachting de i-de kleinste score representeert terwijl x_i in (3.1) de score bij individu i is.

35. Bereken de verwachting van de toevalsvariabele $X =$ "aantal ogen" bij de worp van een dobbelsteen.

36. Je werpt een munt. Als de uitkomst kop is, krijg je 3€; anders krijg je niets. Wat is de verwachting van je winst ?

37. Bij een hazardspel win je 50€ als de dobbelsteen op 6 valt, 10€ op 5, 0€ op 4, -20€ op 3, -30€ op 2 en -30€ op 1. Een negatieve winst is een verlies. Wat is de verwachting van de toevalsvariabele "winst" ?

38. Bereken de variantie van de toevalsvariabele $X =$ "aantal ogen" bij de worp van een dobbelsteen.

39. Wat is de variantie van de toevalsvariabele "winst" van oefening 37 ?

De variantie De variantie (symbool $V(X)$) van een kansverdeling wordt gegeven door

$$V(X) = \sum_{i=1}^{p} P(X = x_i)\,(x_i - E(X))^2.$$

In plaats van het symbool $V(X)$ gebruikt men vaak σ_X^2 of gewoon σ^2 als de context duidelijk is. Dit wordt ook de populatievariantie genoemd.

De vierkantswortel van de variantie wordt de *standaarddeviatie* genoemd. Zijn symbool is σ_X of σ als de context duidelijk is.

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{V(X)}.$$

De variantie en de standaarddeviatie van een toevalsvariabele zijn moeilijk te interpreteren, net zoals de variantie en de standaarddeviatie van een geobserveerde variabele in de beschrijvende statistiek.

3.1.8.2 Continue toevalsvariabelen

Hier is het principe ook hetzelfde als in beschrijvende statistiek maar, omdat het aantal mogelijke waarden van de variabele oneindig is, kunnen we niet meer een som voor alle waarden berekenen. Hier gebruikt men een andere techniek : de integralen. Hieronder vind je de formules van de verwachting en van de variantie van een continue toevalsvariabele. Je hoeft ze niet te kunnen gebruiken.
De verwachting van een continue toevalsvariabele X is

$$E(X) = \int_{-\infty}^{+\infty} f_X(x)\, x\, dx.$$

De variantie van een continue toevalsvariabele X is

$$V(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} f_X(x)\,(x - E(X))^2\, dx.$$

3.1.9 Associatietechnieken

We zullen hier alleen maar het lineaire verband tussen twee variabelen bestuderen.

3.1.9.1 Discrete toevalsvariabelen

We kunnen opnieuw de covariantie en de correlatiecoëfficiënt definiëren (zie Ver. 2.1, blz. 38). De covariantie is

$$COV(X,Y) = \sum_{i=1}^{p}\sum_{j=1}^{q} P(X=x_i, Y=y_j)\big(x_i - E(X)\big)\big(y_j - E(Y)\big).$$

De correlatiecoëfficiënt is

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}.$$

Het zegt ons in welke mate een lineair verband tussen de twee variabelen bestaat. Het teken van $\rho_{X,Y}$ zegt ons ook of het verband stijgend ($\rho_{X,Y} > 0$) of dalend ($\rho_{X,Y} < 0$) is. Zoals in beschrijvende statistiek variëert de correlatiecoëfficiënt tussen -1 (volmaakt negatief verband) en +1 (volmaakt positief verband).

Een waarde van $\rho_{X,Y}$ gelijk nul betekent niet dat er geen verband tussen de twee variabelen bestaat. Een niet lineair verband zou aanwezig kunnen zijn.

40. Maak een visuele schatting van de correlatiecoëfficiënt tussen de variabelen X en Y van onderstaande kansverdeling. Is het positief of negatief?

		Y	
		0	*1*
X	*0*	*0.1*	*0.3*
	1	*0.4*	*0.2*

3.1.9.2 Continue toevalsvariabelen

De definitie van de covariantie wordt aangepast.

$$COV(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,\big(x - E(X)\big)\,\big(y - E(Y)\big)\,dx\,dy.$$

De definitie van de correlatiecoëfficiënt blijft dezelfde als bij discrete toevalsvariabelen. De interpretatie is ook dezelfde.

3.1.10 Enkele nuttige stellingen

3.1.10.1 De kansen van complementaire gebeurtenissen

Laat A en A^* complementaire gebeurtenissen zijn bij een bepaald toevalsproces.

$$P(A) + P(A^*) = P(A \cup A^*) = 1.$$

3.1.10.2 De verwachting van een constante maal een variabele

Laat $Z = aX$ een toevalsvariabele zijn, met a een constante.

$$E(Z) = aE(X).$$

Toepassing van deze stelling: Stel dat X het loon in Euro representeert. De verwachting van X is 1500€. Wat is dan de verwachting van de loon in BEF ? Laten we het loon in BEF door Z aanduiden. Dan $Z = 40X$ en, dankzij de stelling hierboven, $E(Z) = 40E(X) = 60\,000$ BEF. Dit is evident en je had geen stelling nodig om het loon van Euro naar BEF om te zetten, maar we gaan minder evidente stellingen zien.

3.1.10.3 De verwachting van een som

Laat $Z = X + Y$ een toevalsvariabele zijn.

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

M.a.w., de verwachting van een som is de som van de verwachtingen.

Toepassing van deze stelling: je kent de verwachting van het loon van mannen en de verwachting van het loon van vrouwen. Wat is dan de verwachting van het loon van hetero echtparen ? Het loon van een echtpaar (Z) is de som van de lonen van de vrouw (X) en van de man (Y). Dus $Z = X + Y$ en de verwachting van het loon van echtparen is dus de som van de verwachtingen. Dit geldt alhoewel we weten dat lonen van vrouwen en mannen niet onafhankelijk zijn. Mannen met hoge lonen leven vaak met vrouwen met hoge lonen.

3.1.10.4 De verwachting van een verschil

Laat $Z = X - Y$ een toevalsvariabele zijn.

$$E(Z) = E(X) - E(Y).$$

M.a.w., de verwachting van een verschil is het verschil tussen de verwachtingen.

Toepassing van deze stelling: je kent de verwachting van het loon van mannen en de verwachting van het loon van vrouwen. Wat is dan de verwachting van het verschil tussen mannen en vrouwen in hetero echtparen ? Het verschil in een echtpaar (Z) is het verschil tussen de lonen van de vrouw (X) en van de man (Y). Dus $Z = X - Y$ en de verwachting van het verschil is dus het verschil tussen de verwachtingen. Dit geldt alhoewel we weten dat lonen van vrouwen en mannen niet onafhankelijk zijn. Mannen met hoge lonen leven vaak met vrouwen met hoge lonen.

3.1.10.5 De variantie van een som

Laat $Z = X + Y$ een toevalsvariabele zijn.

$$V(Z) = V(X) + V(Y) + 2\,COV(X, Y).$$

M.a.w., de variantie van een som is de som van de varianties plus 2 maal de covariantie.

In het geval van positief gecorreleerde variabelen is dus de variantie van een som groter dan de som van de varianties. Bv. de variantie van het loon

van hetero echtparen is groter dan de som van de varianties van de lonen van vrouwen en van mannen. Het is normaal : mannen met hoge lonen zijn vaak met vrouwen met hoge lonen en mannen met lage lonen zijn vaak met vrouwen met lage lonen. De lonen van mannen en die van vrouwen zijn positief gecorreleerd. Er zijn dus veel echtparen met zeer hoge lonen en veel echtparen met zeer lage lonen. Dit zorgt voor een hoge variantie van de variabele "loon van hetero echtparen".

Numerieke toepassing. We gaan de standaarddeviatie van de som $X + Y$ berekenen voor de variabelen van Tabel 3.3. In rubr. 3.1.9, hebben we al de varianties en de covariantie berekend : $V(X) = 0.61$, $V(Y) = 0.81$ en $COV(X, Y) = 0.01$. De variantie van $X + Y$ is dus $V(X + Y) = 0.61 + 0.81 + 2 \times 0.01 = 1.44$. De standaarddeviatie ($\sigma$) van de som is dus $\sigma_{X+Y} = \sqrt{1.44} = 1.2$.

3.1.10.6 De variantie van een verschil

Laat $Z = X - Y$ een toevalsvariabele zijn.

$$V(Z) = V(X) + V(Y) - 2\,COV(X, Y).$$

M.a.w., de variantie van een aftrekking is de *som* van de varianties *min* 2 maal de covariantie.

Zelfs in het geval van een aftrekking moeten we de varianties optellen.

3.1.10.7 Correlatie en afhankelijkheid

De covariantie van onafhankelijke toevalsvariabelen is altijd nul. Zo ook hun correlatiecoëfficiënt.

Het omgekeerde is niet waar. Wanneer $\rho_{X,Y} = 0$ bestaat er weliswaar geen lineaire samenhang tussen X en Y maar mogelijk wel een niet-lineaire samenhang.

41. Je werpt twee zuivere dobbelstenen. De variabele Z is de som van de twee uitkomsten. Wat is de standaarddeviatie van Z ?

3.2 Bijzondere kansverdelingen

Er zijn een aantal specifieke kansverdelingen die zeer nuttig zijn in de sociale wetenschappen: de binomiale verdeling, de normale verdeling, de χ^2-verdeling, de t-verdeling en de F-verdeling. Ze worden in deze rubriek voorgesteld.

3.2.1 De binomiale verdeling

Met de binomiale verdeling kunnen we de kans modelleren dat een aselecte steekproef van n proefpersonen k personen bevat met een bepaalde kenmerk. Bv. de kans dat een aselecte steekproef van 30 volwassene Vlamingen twee individuen met ASD bevat. Om deze kans te berekenen moeten we uiteraard de prevalentie van ASD kennen, i.e. de proportie π van Vlamingen met ASD. We kunnen ook de kans modelleren dat een aselecte steekproef van 100 jongeren 20 jongeren in het bso bevat. We hebben ook de proportie π van jongeren in het bso nodig.

De binomiale verdeling wordt gegeven door

$$P(X \sim B(n,\pi) = k) = \frac{n!}{k!\ (n-k)!}\ \pi^k (1-\pi)^{(n-k)}. \tag{3.2}$$

De binomiale variabele is discreet en kan de waarden $0, 1, 2, \ldots, n$ aannemen.

Toepassing. De proportie jongeren met leeftijd 16 die in het bso zitten, is 1/3. Je trekt een steekproef van 4 jongeren met leeftijd 16. Wat is de kans dat geen van de vier in het bso zit? De variabele X = aantal jongeren in het bso volgt een binomiale verdeling: $X \sim B(4, 1/3)$. De gevraagde kans is

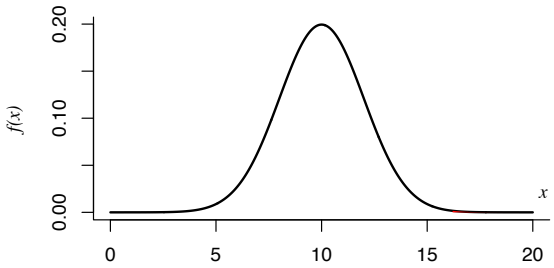
$$\begin{aligned} P(X=0) &= \frac{4!}{0! \times (4-0)!}\ \left(\frac{1}{3}\right)^0 \left(1-\frac{1}{3}\right)^{(4-0)} \\ &= \frac{24}{1 \times 24}\ \left(\frac{1}{3}\right)^0 \left(1-\frac{1}{3}\right)^{(4-0)} \\ &= 1 \times 1 \times \left(\frac{2}{3}\right)^4 \\ &= 1 \times 1 \times \frac{2^4}{3^4} \\ &= \frac{16}{81}. \end{aligned}$$

3.2.2 De normale verdeling

Een normaal verdeelde variabele X met verwachting μ en variantie σ^2 (notatie $X \sim N(\mu, \sigma^2)$) is een continue toevalsvariabele waarvan de dichtheidsfunctie gegeven wordt door

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

Fig. 3.8 stelt de dichtheidsfunctie voor van een normale variabele met verwachting $\mu = 10$ en variantie $\sigma^2 = 4$. Merk op dat deze functie een symmetrische



Figuur 3.8: De dichtheidsfunctie van de normale verdeling met $\mu = 10$ en $\sigma^2 = 4$.

kromme is en dat haar hoogste punt overeenkomt met de verwachting μ. Dit geldt voor alle normale variabelen, los van de waarde van μ en σ.

R biedt een aantal functies om te werken met deze verdeling. De functie pnorm laat je toe om de kans

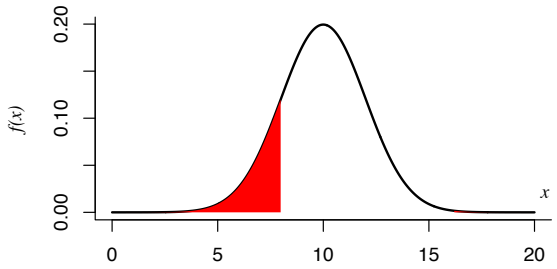
$$P(X \sim N(\mu, \sigma^2) \leq x) \tag{3.3}$$

te bekomen. Dit is de oppervlakte onder de curve, aan de linkerkant van x. De functie pnorm heeft drie argumenten nodig: q is het equivalent van x in (3.3), mean is het equivalent van μ en sd het equivalent van σ.

Voorbeeld: de kans $P(X \sim N(10, 4) \leq 8)$ wordt bekomen d.m.v.

```
> pnorm( q = 8, mean = 10, sd = 2)
[1] 0.1586553
```

Deze kans wordt gerepresenteerd door de rode oppervlakte in Fig. 3.9



Figuur 3.9: De kans $P(X \sim N(10, 4) \leq 8)$ bij een normale verdeling met $\mu = 10$ en $\sigma^2 = 4$.

Om de witte oppervlakte onder de kromme van Fig. 3.9 aan de rechterkant van 8 te berekenen, gebruik je ook de functie pnorm, maar met een extra argument:

```
> pnorm( q = 8, mean = 10, sd = 2, lower.tail=FALSE)
[1] 0.8413447
```

Dankzij het argument lower.tail=FALSE weet R dat je de kans in de rechter staart wil berekenen en niet in de linker staart. De berekende kans is $P(X \sim N(10, 4) \geq 8)$.

We willen nu deze kans $P(8 \leq X \sim N(10, 4) \leq 10)$ berekenen. Het is de rode oppervlakte in Fig. 3.10 Het is gelijk aan de rode oppervlakte van Fig. 3.11 min de rode oppervlakte van Fig. 3.9. We kunnen dit formeel uitdrukken:

42. Bereken de kans $P(X \sim N(10, 4) \leq 10)$, m.b.v. R.

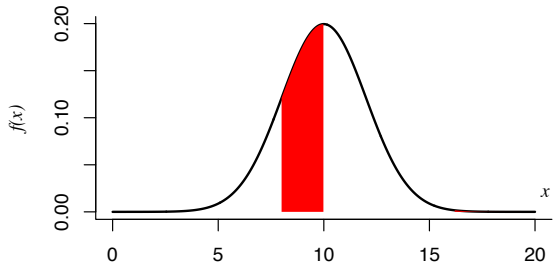
$$P(8 \leq X \sim N(10, 4) \leq 10) = P(X \sim N(10, 4) \leq 10) - P(X \sim N(10, 4) \leq 8).$$

Nu kunnen we dit berekenen m.b.v. R.

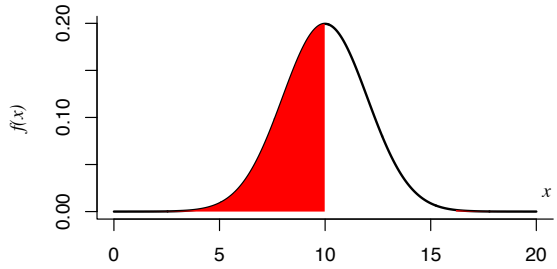
```
> pnorm( q = 10, mean = 10, sd = 2) - pnorm( q = 8, mean = 10, sd = 2)
[1] 0.3413447
```

43. Bereken de kans $P(120 \leq X \sim N(100, 225) \leq 130)$, m.b.v. R.

De functie pnorm geeft de kans onder een bepaalde waarde. De functie qnorm doet het omgekeerde. Ze geeft de waarde met p kans eronder. Voorbeeld:



Figuur 3.10: De kans $P(8 \leq X \sim N(10,4) \leq 10)$ bij een normale verdeling met $\mu = 10$ en $\sigma^2 = 4$.



Figuur 3.11: De kans $P(X \sim N(10,4) \leq 10)$ bij een normale verdeling met $\mu = 10$ en $\sigma^2 = 4$.

```
> qnorm( p = 0.5, mean = 10, sd = 2 )
[1] 10
> qnorm( p = 0.25, mean = 10, sd = 2 )
[1] 8.65102
```

De normale verdeling met verwachting $\mu = 0$ en standaarddeviatie $\sigma = 1$ wordt de standaardnormale verdeling genoemd.

44. Het IQ is normaal verdeeld met $\mu = 100$ en $\sigma = 15$. Welke proportie van de populatie heeft een IQ lager dan 90? Wat is het IQ met 20% van de populatie eronder?

3.2.3 De centrale limietstelling

Stel dat $X_1, \ldots, X_n$, n onafhankelijke toevalsvariabelen zijn, met dezelfde verdeling, met verwachting μ_X en variantie σ_X^2, dan wordt de verdeling van de toevalsvariabele

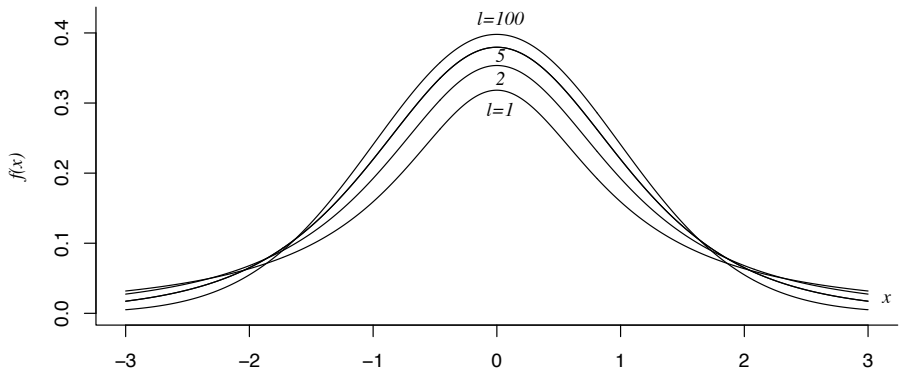
$$\frac{X_1 + \ldots + X_n}{n}$$

naarmate n groter wordt, steeds beter benaderd door de normale verdeling met verwachting μ_X en variantie σ_X^2/n. In de limiet is de benadering perfect.

Merk op dat deze stelling juist is, los van de verdeling van de variabelen $X_1, \ldots, X_n$. Ze hoeven niet normaal te zijn. In de praktijk is de benadering zeer goed voor alle $n \geq 30$, behalve als de verdeling van X zeer scheef is.

3.2.4 De Student verdeling of t-verdeling

De Student[2] verdeling is eigenlijk een oneindige familie van continue kansverdelingen. Elk lid van deze familie wordt gekenmerkt door een positief geheel getal (aantal vrijheidsgraden). De dichtheidsfunctie van de Student verdeling met l vrijheidsgraden (of t_l-verdeling) is een klok en de breedte van deze klok wordt kleiner naar gelang l groter wordt. Fig. 3.12 stelt de dichtheidsfunctie voor van de t_l-verdeling met $l = 1, 2, 5$ en 100. Merk op dat deze functie symmetrisch



Figuur 3.12: De dichtheidsfunctie van de t_l-verdeling met $l = 1, 2, 5$ en 100

is. Het lijkt op de dichtheidsfunctie van de normale verdeling, maar het is niet dezelfde curve. Als l naar oneindig gaat, dan zijn de twee curves wel bijna identiek. Los van de waarde van l is de verwachting altijd nul.

R biedt ook functies om kansen met de t-verdeling te berekenen: de functies `pt` en `qt` werken net zoals de functie `pnorm` en `qnorm`. Hieronder vind je twee voorbeelden.

Wat is de kans $P(Y \sim t_{10} \leq 1.3)$? (zie Fig.3.13)

```
> pt( q = 1.3, df = 10 )
[1] 0.8886171
```

45. Bereken $P(1 \leq Y \sim t_{120})$ m.b.v. R.

Om de kans $P(Y \sim t_{10} \geq 1.3)$ te berekenen gebruik je dezelfde functie met het extra argument `lower.tail=FALSE`.

Als Y t_{10}-verdeeld is, wat is de waarde met 15% kans eronder? (zie Fig.3.14).

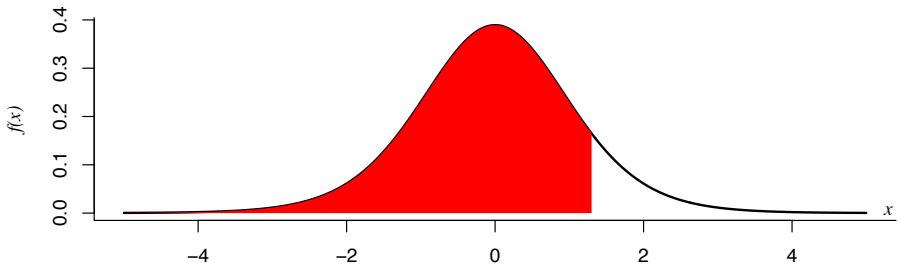
```
> qt( p = 0.15, df = 10 )
[1] -1.093058
```

46. Als Y t_{150}-verdeeld is, wat is de waarde met 60% kans erboven?

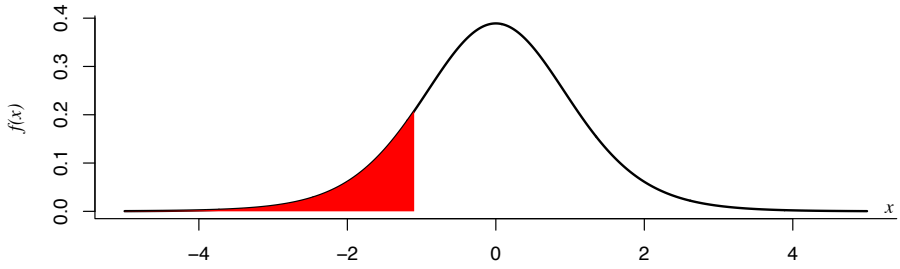
3.2.5 De F-verdeling

De F- verdeling is eigenlijk een oneindige familie van continue kansverdelingen. Elk lid van deze familie wordt gekenmerkt door twee positieve gehele getallen

[2] Student is het pseudoniem van William Sealy Gosset, 1876–1937.



Figuur 3.13: De kans $P(Y \sim t_{10} \leq 1.3)$



Figuur 3.14: Dichtheidsfunctie van t_{10}; de rode oppervlakte is 15%

(aantal vrijheidsgraden). De dichtheidsfunctie van de F-verdeling met l_1 en l_2 vrijheidsgraden (of F_{l_1,l_2}-verdeling) is een asymmetrische klok. Fig. 3.15 stelt de dichtheidsfunctie voor van de F_{l_1,l_2}-verdeling met $l_1 = 2, 10, 100$ en $l_2 = 15$ vrijheidsgraden. Deze verdeling wordt ook de Fisher[3]-Snedecor[4] verdeling genoemd.

R biedt ook functies om kansen met de F-verdeling te berekenen: de functies `pf` en `qf` werken net zoals de functies `pnorm` en `qnorm`. Met deze functies mag je ook het argument `lower.tail=FALSE` gebruiken. Hieronder vind je twee voorbeelden.

Wat is de kans $P(F \sim F_{10,3} \leq 2)$? (zie Fig.3.16)

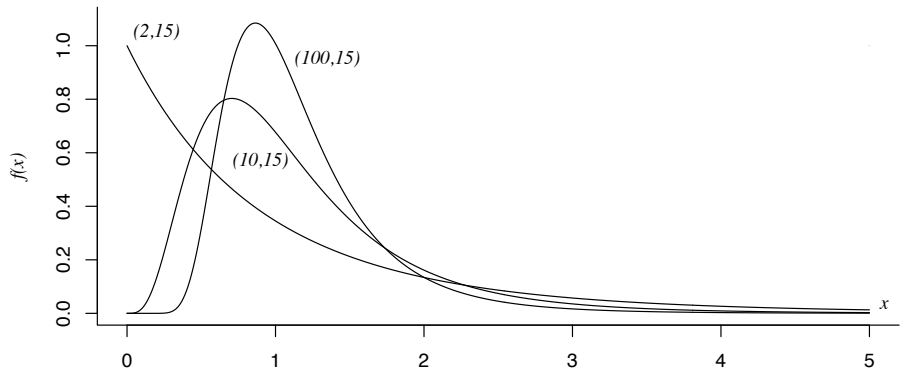
```
> pf( q = 2, df1 = 10, df2 =3 )
[1] 0.6906222
```

Als F $F_{50,10}$-verdeeld is, wat is de waarde met 20% kans eronder? (zie Fig. 3.17).

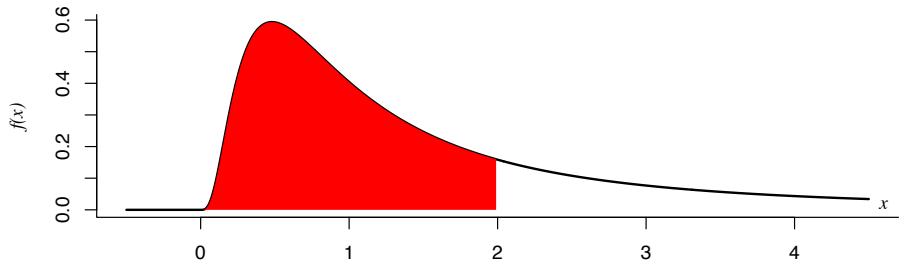
```
> qf( p = 0.2, df1= 50, df2 = 10 )
[1] 0.7052699
```

47. Als X $F_{10,12}$-verdeeld is, wat is de waarde met 20% kans erboven?

[3] Ronald Fisher, 1890–1962.
[4] George W. Snedecor, 1881–1974.



Figuur 3.15: De dichtheidsfunctie van de F_{l_1,l_2}-verdeling met $l_1 = 2, 10, 100$ en $l_2 = 15$ vrijheidsgraden



Figuur 3.16: De kans $P(F \sim F_{10,3} \leq 2)$

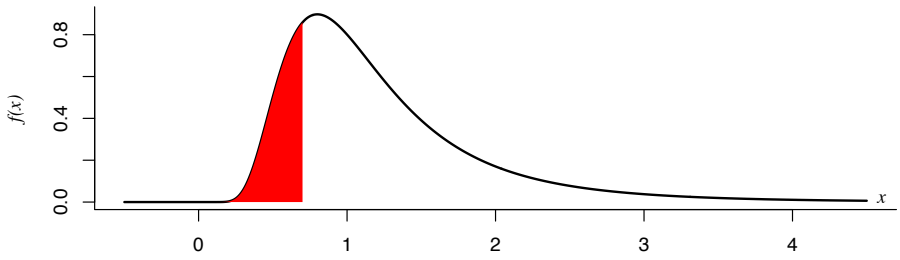
3.3 De steekproevenverdelingen

In de beschrijvende statistiek worden de symbolen $x_1, \ldots, x_n$ gebruikt om de scores van de variabele X in een bepaalde steekproef aan te duiden. Indien we meerdere steekproeven trekken, dan gaat x_1 variëren, op een onvoorspelbare manier. Hetzelfde geldt uiteraard voor $x_2, \ldots, x_n$. De score van de variabele X bij individu 1 is dus een toevalsvariabele en wordt aangeduid door X_1. Bij een specifieke steekproef is x_1 een realisatie van de toevalsvariabele X_1. Op dezelfde manier definiëren we $X_2, \ldots, X_n$.

Indien we het gemiddelde van de variabele X in meerdere steekproeven berekenen, gaat ze ook variëren. In het kansrekenen gebruiken we dus ook een speciale notatie ($\overline{X}$) voor de toevalsvariabele "gemiddelde van X", terwijl $\overline{x}$ het gemiddelde in een specifieke steekproef representeert. De formule van $\overline{X}$ is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Op dezelfde manier kunnen we de toevalsvariabele "variantie van X" de-



Figuur 3.17: Dichtheidsfunctie van $F_{50,10}$; de rode oppervlakte is 20%

finiëren, met symbool S_X^2:

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

of

$$S_X^2 = \frac{\mathrm{SS}_X}{n-1}.$$

Merk op dat we dezelfde notatie SS_X gebruiken voor $\sum_{i=1}^{n}(X_i - \overline{X})^2$ en $\sum_{i=1}^{n}(x_i - \overline{x})^2$. De context maakt het duidelijk wat we bedoelen met SS_X.

We kunnen hetzelfde doen met de mediaan, de modus, de interkwartiele afstand, enz. Al die nieuwe toevalsvariabelen die een combinatie zijn van de toevalsvariabelen $X_1, \ldots, X_n$ zijn, worden steekproefgrootheden of statistieken genoemd. Steekproefgrootheden zijn toevalsvariabelen en hebben dus een kansverdeling. Deze kansverdeling wordt een steekproevenverdeling genoemd. De steekproevenverdeling van een steekproefgrootheid wordt altijd geanalyseerd onder de hypothese dat de steekproef het resultaat is van n lukrake trekkingen met teruglegging (men spreekt ook van een aselecte steekproef).

3.4 De steekproevenverdeling van $\overline{X}$

De verwachting van $\overline{X}$ is altijd dezelfde als die van X:

$$E(\overline{X}) = E(X) = \mu_X$$

maar de variantie van $\overline{X}$ is kleiner dan die van X:

$$V(\overline{X}) = \frac{1}{n} V(X) = \frac{\sigma_X^2}{n}.$$

Stel dat $X_1, \ldots, X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met een normale verdeling $N(\mu_X, \sigma_X^2)$, dan zal $\overline{X}$ ook normaal verdeeld zijn:

$$\overline{X} \sim N(\mu_X, \sigma_X^2/n).$$

Stel dat $X_1, \ldots, X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met een onbekende verdeling met verwachting μ_X en variantie σ_X^2, dan zal $\overline{X}$ bij benadering normaal verdeeld zijn indien $n > 30$ en indien de verdeling van X niet te scheef is:

$$\overline{X} \sim N(\mu_X, \sigma_X^2/n).$$

48. Het IQ is normaal verdeeld met $\mu = 100$ en $\sigma = 15$. Als je steekproeven van 15 individuen trekt, wat is de kans dat het steekproef-gemiddelde kleiner dan 90 is?

3.5 Oplossingen

31) Wat is de kans dat de reactietijd zich niet tussen 1 en 1.5 s bevindt ?

Oplossing: "Niet tussen 1 en 1.5 s" is de complementaire gebeurtenis van "tussen 1 en 1.5 s." Het antwoord is dus

$$1 - P(1 < X \leq 1.5) \approx 1 - 0.08 = 0.92.$$

32) Wat zijn de waarden van p en q in het voorbeeld van Tabel 3.2 ?

Oplossing: $p = q = 4$.

33) Wat is $\sum_{i=1}^{3} P(X = x_i, Y = y_2)$ in Tabel 3.2 ?

Oplossing:

$$\begin{aligned}\sum_{i=1}^{3} P(X = x_i, Y = y_2) &= P(X = x_1, Y = y_2) + P(X = x_2, Y = y_2) + P(X = x_3, Y = y_2) \\ &= P(X = 0, Y = 1) + P(X = 1, Y = 1) + P(X = 2, Y = 1) \\ &= 0.19 + 0.06 + 0.05 = 0.30.\end{aligned}$$

34) Zijn de variabelen `gewicht` en `lengte` van het data frame `myData` (zie Rubr. 2.3.3, blz. 38) onafhankelijk ?

Oplossing: De variabelen `gewicht` en `lengte` zijn geobserveerde variabelen; ze zijn realisaties van de toevalsvariabelen gewicht en lengte. We zijn geneigd te denken dat de twee toevalsvariabelen afhankelijk zijn omdat, in onze steekproef, de variabelen sterk gecorreleerd zijn ($r = 0.47$). Maar dat is slechts een steekproef. Het kan zijn dat de twee toevalsvariabelen onafhankelijk zijn maar dat we een niet representatieve steekproef hebben getrokken. Het is dus onmogelijk iets te concluderen. Later zullen we technieken zien om dit soort veralgemening te doen.

35) Bereken de verwachting van de toevalsvariabele $X =$"aantal ogen" bij de worp van een dobbelsteen.

Oplossing:

$$\begin{aligned}E(X) &= \sum_{i=1}^{6} P(X = x_i)\, x_i \\ &= P(X = x_1)\, x_1 + P(X = x_2)\, x_2 + \ldots + P(X = x_6)\, x_6 \\ &= \frac{1}{6}\,1 + \frac{1}{6}\,2 + \ldots + \frac{1}{6}\,6 = 3.5.\end{aligned}$$

36) Je werpt een munt. Als de uitkomst kop is, krijg je 3€; anders krijg je niets. Wat is de verwachting van je winst ?

Oplossing: We gebruiken het symbool X voor de winst. De twee mogelijke waarden van de winst zijn 0 en 3.

$$\begin{aligned} E(X) &= \sum_{i=1}^{2} P(X = x_i)\, x_i \\ &= P(X = x_1)\, x_1 + P(X = x_2)\, x_2 \\ &= \frac{1}{2}\, 0 + \frac{1}{2}\, 3 = 1.5. \end{aligned}$$

37) Bij een hazardspel win je 50€ als de dobbelsteen op 6 valt, 10€ op 5, 0€ op 4, -20€ op 3, -30€ op 2 en -30€ op 1. Een negatieve winst is een verlies. Wat is de verwachting van de toevalsvariabele "winst" ?

Oplossing: Eerst moet je de mogelijke waarden van de variabele kennen. Ze zijn $x_1 = -30$, $x_2 = -20$, $x_3 = 0$, $x_4 = 10$ en $x_5 = 50$ ($p = 5$). Dan moet je hun kans berekenen. $P(X = -30) = 2/6$ omdat de winst -30 in twee gevallen kan voorkomen. $P(X = -20) = 1/6$, $P(X = 0) = 1/6$, $P(X = 10) = 1/6$, $P(X = 50) = 1/6$. Eindelijk,

$$E(X) = \frac{2}{6}\,(-30) + \frac{1}{6}\,(-20) + \frac{1}{6}\,0 + \frac{1}{6}\,10 + \frac{1}{6}\,50 = -3.33.$$

Speel je vaak ($n \to \infty$) aan dit spel, dan verlies je in doorsnee elke keer 3.33€. Het is dus geen billijk spel.

38) Bereken de variantie van de toevalsvariabele $X =$"aantal ogen" bij de worp van een dobbelsteen.

Oplossing:

$$\begin{aligned} V(X) &= \sum_{i=1}^{6} P(X = x_i)\,(x_i - 3.5)^2 \\ &= \frac{1}{6}\,(1 - 3.5)^2 + \frac{1}{6}\,(2 - 3.5)^2 + \ldots + \frac{1}{6}\,(6 - 3.5)^2 = 2.92. \end{aligned}$$

39) Wat is de variantie van de toevalsvariabele "winst" van oefening 37 ?

Oplossing:

$$V(X) = \frac{2}{6}\,(-30+3.33)^2 + \frac{1}{6}\,(-20+3.33)^2 + \frac{1}{6}\,(0+3.33)^2 + \frac{1}{6}\,(10+3.33)^2 + \frac{1}{6}\,(50+3.33)^2 = 788.9.$$

40) Maak een visuele schatting van de correlatiecoëfficiënt tussen de variabelen X en Y van onderstaande kansverdeling. Is het positief of negatief?

		Y	
		0	1
X	0	0.1	0.3
	1	0.4	0.2

Oplossing: De combinatie $X = 0$ en $Y = 0$ komt zelden voor (kans $= 0.1$). De combinatie $X = 1$ en $Y = 1$ komt zelden voor (kans $= 0.15$). Maar de andere combinaties komen vaker voor: ($X = 0$ en $Y = 1$) of ($X = 1$ en $Y = 0$). Maw heb je vaak een kleine X-waarde met een grote Y-waarde of omgekeerd. De correlatiecoëfficiënt is dus negatief. Als je de correlatiecoëfficiënt berekent, kom je $\rho_{XY} = -0.41$ uit.

41) Je werpt twee zuivere dobbelstenen. De variabele Z is de som van de twee uitkomsten. Wat is de standaarddeviatie van Z?

Oplossing: De twee aparte uitkomsten noemen we X en Y. Dus $Z = X + Y$. De variabelen X en Y zijn onafhankelijk, want er is geen verband tussen de twee dobbelstenen. De covariantie tussen X en Y is dus nul. De variantie van Z is de som van de twee aparte varianties plus 2 maal de covariantie. Omdat de covariantie nul is, vinden we dat de variantie van Z simpelweg de som van de twee aparte varianties is. Zo, $V(Z) = V(X) + V(Y)$. We hebben $V(X)$ al berekend (oef. 38). Het is 2.92. $V(Y)$ is ook gelijk aan 2.92. Uiteindelijk, $V(Z) = 2 \times 2.92 = 5.84$ en $\sigma_Z = 2.42$.

42) Bereken de kans $P(X \sim N(10, 4) \leq 10)$, m.b.v. R.

Oplossing:

```
> pnorm( q = 10, mean = 10, sd = 2 )
[1] 0.5
```

Het is geen verrassing dat $P(X \sim N(10, 4) \leq 10) = 0.5$ want de dichtheidsfunctie van de normale verdeling $N(10, 4)$ is symmetrisch om het punt 10.

43) Bereken de kans $P(120 \leq X \sim N(100, 225) \leq 130)$, m.b.v. R.

Oplossing:

```
> pnorm( q = 130, mean = 100, sd = 15) - pnorm( q = 120, mean = 100, sd = 15 )
[1] 0.06846109
```

44) Het IQ is normaal verdeeld met $\mu = 100$ en $\sigma = 15$. Welke proportie van de populatie heeft een IQ lager dan 90? Wat is het IQ met 20% van de populatie eronder?

Oplossing: De twee vragen zijn
$P(X \sim N(100, 15^2) \leq 90) = ?$
$P(X \sim N(100, 15^2) \leq ?) = 0.20.$
De antwoorden zijn

```
> pnorm( q = 90, mean = 100, sd = 15 )
[1] 0.2524925
> qnorm( p = 0.20, mean = 100, sd = 15 )
[1] 87.37568
```

45) Bereken $P(1 \leq Y \sim t_{120})$ m.b.v. R.

Oplossing:

```
> pt( q = 1, df = 120 , lower.tail = FALSE)
[1] 0.1596614
```

46) Als Y t_{150}-verdeeld is, wat is de waarde met 60% kans erboven?

Oplossing:

```
> qt( p = 0.6, df = 150 , lower.tail = FALSE)
[1] -0.2537969
```

47) Als X $F_{10,12}$-verdeeld is, wat is de waarde met 20% kans erboven?

Oplossing:

```
> qf( p = 0.2, df1 = 10, df2 = 12, lower.tail = FALSE )
[1] 1.663042
```

Je kan ook deze commando gebruiken:

```
> qf( p = 0.8, df1 = 10, df2 = 12 )
[1] 1.663042
```

48) Het IQ is normaal verdeeld met $\mu = 100$ en $\sigma = 15$. Als je steekproeven van 15 individuen trekt, wat is de kans dat het steekproef-gemiddelde kleiner dan 90 is?

Oplossing: We gebruiken het symbool X voor het IQ. We weten dat $X \sim N(100, 15^2)$. Bijgevolg $\overline{X} \sim N(100, 15^2/15) = N(100, 15)$. De gevraagde kans wordt dan met de functie `pnorm` berekend.

```
> pnorm(q=90, mean = 100, sd = sqrt(15))
[1] 0.004911637
```

De gevraagde kans is ongeveer 0.5%.

Hoofdstuk 4

Puntschatting

Het komt vaak voor dat we de verdeling van een variabele in een populatie niet kennen, of niet volledig. We wensen dus één of meerdere parameters te schatten, op basis van een steekproef (bvb. de parameters μ en σ van een normale variabele). Om een parameter θ te schatten gebruiken we een steekproefgrootheid (of statistiek). In het algemeen, noemen we ze een *schatter* (symbool Q). Een schatter heeft dus een steekproevenverdeling. Op basis van een steekproef berekenen we de steekproefgrootheid Q en we bekomen een getal (de realisatie van de steekproefgrootheid). Dit getal wordt de *schatting* genoemd. Zijn symbool is $\hat{\theta}$.

Een schatter is een toevalsvariabele. Telkens als we een steekproef trekken weten we niet wat de waarde van de schatter zal zijn. Een schatting is geen toevalsvariabele. Een schatting is de waarde (de realisatie) van de schatter in een bepaalde steekproef.

4.1 Eigenschappen van een goede schatter

Om goede schattingen te bekomen willen we schatters gebruiken die zo vaak mogelijk een goede schatting geven. Goed betekent hier "niet te verschillend van de te schatten parameter θ". Q is een goede schatter van θ indien

- ze zuiver (unbiased) is, i.e., de verwachting van de schatter is gelijk aan de te schatten parameter:

  $$E(Q) = \theta.$$

  Met andere woorden, we weten dat de schatter zelden een perfecte schatting zal geven. Soms zal de schatting te groot zijn. Soms te klein. Maar in doorsnee willen we dat de schatter gelijk is aan de parameter.

- de variantie van de schatter, $V(Q)$, kleiner wordt naarmate de steekproefgrootte toeneemt. Dit drukt uit dat de schatter nauwkeuriger zal zijn

wanneer de steekproef groter wordt. Als we verschillende schatters hebben voor een bepaalde populatieparameter, dan zeggen we dat de schatter met de kleinste variantie het efficiëntst is.

Met andere woorden, we weten dat de schatter zelden een perfecte schatting zal geven. Soms zal de schatting te groot zijn. Soms te klein. Maar we willen dat de afwijkingen zo klein mogelijk zijn.

De twee bekendste methoden om een parameter te schatten zijn de *grootste aannemelijkheid* (maximum likelihood) en de *kleinste kwadraten* (least squares). Geen van die twee methoden is perfect. De grootste aannemelijkheid methode levert altijd efficiënte schatters maar ze zijn niet altijd zuiver.

4.2 Standaardfout

De standaarddeviatie van een schatter wordt de standaardfout genoemd (Engels: standard error, SE). Dit is een maat voor de precisie van de schatter. Hoe kleiner de standaardfout van een schatter, hoe beter zijn precisie.

4.3 Enkele schatters

Hieronder vind je enkele belangrijke schatters.

4.3.1 De verwachting

De grootste aannemelijkheid schatter van de verwachting van een variabele X is de steekproefgrootheid $\overline{X}$. Hij is zuiver en efficiënt. We zullen dus geen andere schatter bespreken. De corresponderende schatting (realisatie van de schatter) is $\hat{\mu}_X = \overline{x}$.

49. Gebruik de gegevens van het data frame `sportData` *om de verwachting van de toevalsvariabelevariabele* `leeftijd` *te schatten, m.b.v. R.*

De standaardfout van de schatter $\overline{X}$ is $\sigma_X/\sqrt{n}$ (zie Rubr. 3.4). Dit noteren we ook $\mathrm{SE}_{\overline{X}} = \sigma_X/\sqrt{n}$. Dus hoe groter de steekproef, hoe kleiner de standaardfout en hoe beter de precisie van de schatter.

4.3.2 De variantie

De grootste aannemelijkheid schatter van de variantie van een variabele X is de steekproefgrootheid

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{\mathrm{SS}_X}{n}.$$

Deze schatter is niet zuiver en wordt niet gebruikt. Een goede schatter van de variantie σ_X^2 van een variabele X is de steekproefgrootheid

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{\mathrm{SS}_X}{n-1}.$$

Hij is zuiver en we zullen dus altijd deze schatter gebruiken. De realisatie van deze schatter in een steekproef is

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

en wordt gebruikt als schatting van σ_X^2. Dus $s_X^2 = \hat{\sigma}_X^2$.

Voorbeeld. Je wil de gegevens van het data frame `myData` gebruiken om de populatievariantie van de variabele `iq` te schatten.

```
> var( myData$iq )
[1] 246.5471
```

Dus $\hat{\sigma}^2_{\texttt{iq}} = 246.6$.

50. Gebruik de gegevens van het data frame `sportData` *om de populatievariantie van de toevalsvariabelevariabele* `sport` *te schatten, m.b.v. R.*

4.3.3 De covariantie

Een goede schatter van de covariantie COV_{XY} is de steekproefgrootheid

$$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}).$$

Hij is zuiver. Zijn realisatie in een steekproef wordt als schatting van COV_{XY} gebruikt:

$$\widehat{COV}_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Voorbeeld. Je wil de gegevens van het data frame `myData` gebruiken om de populatiecovariantie tussen de variabelen `gewicht` en `lengte` te schatten.

```
> cov( myData$gewicht, myData$lengte )
[1] 76.85057
```

Dus $\widehat{COV}_{\texttt{gewicht,lengte}} = 76.9$.

51. Gebruik de gegevens van het data frame `sportData` *om de populatiecovariantie tussen de toevalsvariabelen* `sport` *en* `tijd` *te schatten, m.b.v. R.*

4.3.4 De correlatiecoëfficiënt

De grootste aannemelijkheid schatter van de correlatiecoëfficiënt ρ_{XY} is de overeenkomende correlatiecoëfficiënt in de steekproef. Deze correlatiecoëfficiënt in een bepaalde steekproef is r_{XY}. Laat R_{XY} de toevalsvariabele zijn die in elke steekproef gelijk is aan de correlatiecoëfficiënt r_{XY}. R_{XY} is de grootste aannemelijkheid schatter van ρ_{XY}. Hij is zuiver en efficiënt.

52. Gebruik de gegevens van het data frame `sportData` *om de correlatiecoëfficiënt ρ tussen de toevalsvariabelen* `sport` *en* `tijd` *te schatten, m.b.v. R.*

4.4 Oplossingen

49) Gebruik de gegevens van het data frame `sportData` om de verwachting van de toevalsvariabelevariabele `leeftijd` te schatten, m.b.v. R.

Oplossing: De schatting $\hat{\mu}$ is gewoon $\overline{x}$. Laten we dus het gemiddelde van `leeftijd` berekenen.

```
> mean( sportData$leeftijd )
[1] 29.29
```

De schatting is dus $\hat{\mu}_{\texttt{leeftijd}} = 29.3$.

50) Gebruik de gegevens van het data frame `sportData` om de populatievariantie van de toevalsvariabelevariabele `sport` te schatten, m.b.v. R.

Oplossing:

```
> var( sportData$sport )
[1] 1.13316
```

De schatting is dus $\hat{\sigma}^2_{\texttt{sport}} = 1.13$.

51) Gebruik de gegevens van het data frame `sportData` om de populatiecovariantie tussen de toevalsvariabelen `sport` en `tijd` te schatten, m.b.v. R.

Oplossing:

```
> cov( sportData$tijd, sportData$sport )
[1] 2.039384
```

De schatting is dus $\widehat{COV}_{\texttt{tijd,sport}} = 2.04$.

52) Gebruik de gegevens van het data frame `sportData` om de correlatiecoëfficiënt ρ tussen de toevalsvariabelen `sport` en `tijd` te schatten, m.b.v. R.

Oplossing:

```
> cor( sportData$tijd, sportData$sport )
[1] 0.4080658
```

De schatting is dus $\hat{\rho}_{\texttt{tijd,sport}} = 0.41$.

Hoofdstuk 5

Intervalschatting - betrouwbaarheidsintervallen

Omwille van het toeval, omwille van de variabiliteit die inherent is aan de trekking van elke aselecte steekproef is een puntschatting bijna altijd fout. Hopelijk niet erg fout, maar toch fout. We gebruiken dus vaak intervalschattingen i.p.v. puntschattingen: we trekken een steekproef, op basis daarvan berekenen we een betrouwbaarheidsinterval en we beschouwen (of we beweren) dan dat de te schatten parameter binnen het betrouwbaarheidsinterval ligt. Dat kan ook fout zijn, maar, indien we het betrouwbaarheidsinterval goed kiezen en indien we deze techniek regelmatig gebruiken, dan zal de te schatten parameter wel meestal binnen het betrouwbaarheidsinterval liggen.

De technieken om een betrouwbaarheidsinterval te berekenen variëren naargelang de parameter die je wil schatten en naar gelang eventuele bijkomende assumpties. Hieronder zien we in detail hoe je een betrouwbaarheidsinterval voor de verwachting van een toevalsvariabele X kunt berekenen.

5.1 Betrouwbaarheidsinterval voor μ_X

We veronderstellen dat σ_X onbekend is, zoals in bijna elke realistische situatie. Er bestaat een andere techniek voor het geval dat σ_X wel bekend is. We zien die techniek niet.

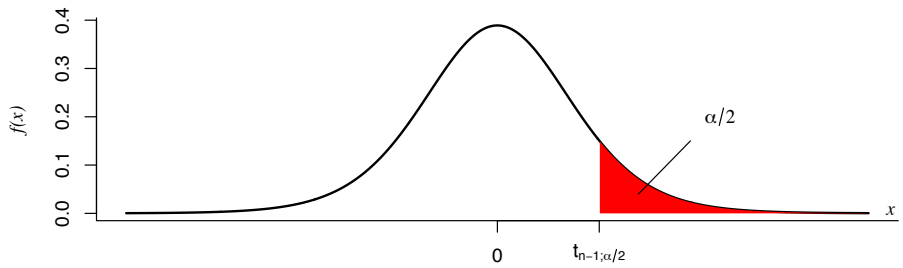
5.1.1 De verdeling van X is normaal

Indien de toevalsvariabele X normaal verdeeld is, met verwachting μ_X en standaarddeviatie σ_X, dan is het mogelijk te bewijzen dat de steekproefgrootheid

$$T = \frac{\overline{X} - \mu_X}{S_X/\sqrt{n}}$$

Studentverdeeld is; in het bijzonder, $T \sim t_{n-1}$.

We duiden met $t_{n-1;\alpha}$ de waarde aan van de t_{n-1}-variabele zodat de oppervlakte rechts gelijk is aan α. Dit wordt vaak een kritieke waarde genoemd. De waarde van de t_{n-1}-variabele zodat de oppervlakte rechts gelijk is aan $\alpha/2$ is dus $t_{n-1;\alpha/2}$ (Fig. 5.1). Dus



Figuur 5.1: De kritieke waarde $t_{n-1;\alpha/2}$

$$P(T > t_{n-1;\alpha/2}) = \alpha/2.$$

53. Bereken $t_{10;0.05}$ m.b.v. R.

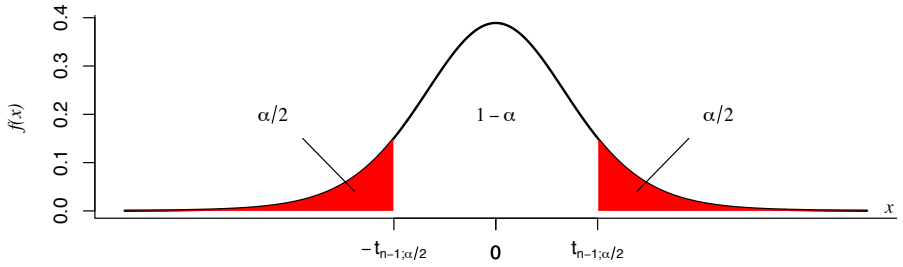
Omwille van de symmetrie van de dichtheidsfunctie van de Student verdeling,

$$P(T < -t_{n-1;\alpha/2}) = \alpha/2.$$

Bijgevolg (Fig. 5.2),

$$P(-t_{n-1;\alpha/2} < T < t_{n-1;\alpha/2}) = 1 - \alpha.$$

We kunnen dit herschrijven als



Figuur 5.2: De kritieke waarden $-t_{n-1;\alpha/2}$ en $t_{n-1;\alpha/2}$

$$P\left(-t_{n-1;\alpha/2} < \frac{\overline{X} - \mu_X}{S_X/\sqrt{n}} < t_{n-1;\alpha/2}\right) = 1 - \alpha$$

of

$$P\left(\overline{X} - t_{n-1;\alpha/2}\,\frac{S_X}{\sqrt{n}} < \mu_X < \overline{X} + t_{n-1;\alpha/2}\,\frac{S_X}{\sqrt{n}}\right) = 1 - \alpha.$$

We gebruiken dus onderstaande formule voor het betrouwbaarheidsinterval voor μ_X

$$\left[\overline{x} - t_{n-1;\alpha/2}\ \frac{s_X}{\sqrt{n}}, \overline{x} + t_{n-1;\alpha/2}\ \frac{s_X}{\sqrt{n}}\right].$$

Bij een bepaald onderzoek en een bepaalde steekproef kunnen we niet weten of het betrouwbaarheidsinterval de verwachting μ_X bevat, maar indien je deze formule gebruikt telkens als je een betrouwbaarheidsinterval nodig hebt, dan zal je intervalschatting meestal juist zijn; de proportie van onderzoeken waar je intervalschatting correct zal zijn, zal $1 - \alpha$ zijn.

We wensen natuurlijk dat onze betrouwbaarheidsintervallen zo vaak mogelijk correct zijn. Onderzoekers gebruiken dus meestal een kleine waarde voor α. In de sociale wetenschappen is $\alpha = 0.05$ de norm.

Voorwaarde: X moet tenminste van intervalniveau zijn en X moet normaal verdeeld zijn.

5.1.2 De verdeling van X is niet normaal of onbekend

Als X niet normaalverdeeld is, dan is de verdeling van de statistiek T van vorige rubriek niet Student en we mogen de formule van vorige rubriek niet gebruiken. Maar, indien de steekproefgrootte groter dan 30 is en indien de verdeling van X niet te scheef is, dan wordt de verdeling van

$$T = \frac{\overline{X} - \mu_X}{S_X/\sqrt{n}}$$

goed benaderd door de Student-verdeling met $n-1$ vrijheidsgraden. De formule voor het betrouwbaarheidsinterval voor de verwachting is dus, zoals hierboven,

$$\left[\overline{x} - t_{n-1;\alpha/2}\ \frac{s_X}{\sqrt{n}}, \overline{x} + t_{n-1;\alpha/2}\ \frac{s_X}{\sqrt{n}}\right].$$

Dit wordt ook als volgt genoteerd:

$$\left[\overline{x} \pm t_{n-1;\alpha/2}\ \frac{s_X}{\sqrt{n}}\right]$$

of (zie Rubr. 4.2)

$$\left[\overline{x} \pm t_{n-1;\alpha/2}\ \mathrm{SE}_{\overline{X}}\right]. \tag{5.1}$$

In woorden, het betrouwbaarheidsinterval voor de verwachting is gelijk aan de schatting van de verwachting plus of minus een kritieke waarde maal de standaardfout van de schatter van de verwachting.

Voorwaarde: X moet tenminste van intervalniveau zijn. De steekproef moet groot zijn ($n \geq 30$).

Voorbeeld. Op basis van de gegevens in `myData` willen we een betrouwbaarheidsinterval berekenen voor de verwachting van `iq`, met een betrouwbaarheid van 95%. We weten niet of de verdeling van `iq` normaal is, maar we mogen bovenstaande benadering gebruiken omdat $n = 30$.

```
> n <- dim( myData )[1]
> kr.waarde <- qt( p = 0.025, df = n-1, lower.tail = FALSE )
> mean(myData$iq) - kr.waarde * sd( myData$iq ) / sqrt(n)
[1] 111.4035
> mean(myData$iq) + kr.waarde * sd( myData$iq ) / sqrt(n)
[1] 123.1298
```

Het betrouwbaarheidsinterval is dus $[111.4, 123.1]$. We beschouwen nu dat $\mu_{\texttt{iq}}$ binnen dit interval ligt. We noemen dit een betrouwbaarheidsinterval met betrouwbaarheid 95%, maar eigenlijk is dit misleidend. De kans 95% heeft niets te maken met dit specifieke interval; wel met de techniek om dit interval op te stellen. Deze techniek levert correcte intervallen in 95% van de gevallen. Maar in dit precieze geval, zegt de techniek niets. Je mag dus niet zeggen dat $\mu_{\texttt{iq}}$ binnen het interval $[111.4, 123.1]$ ligt met 95% kans. Je hebt de steekproef al getrokken en het toeval speelt dus geen rol meer. Van kansen spreken is bijgevolg irrelevant.

54. Bereken een betrouwbaarheidsinterval voor $\mu_{\texttt{tijd}}$ op basis van `sportData`*, met een betrouwbaarheid van 95% en ook met een betrouwbaarheid van 90%.*

5.2 Andere betrouwbaarheidsintervallen

Dit hoofdstuk heeft als doel de essentie van betrouwbaarheidsintervallen te presenteren, op basis van een specifiek geval: de schatting van de verwachting. Het is helemaal niet uitputtend. Verder in deze cursus gaan we nog veel andere betrouwbaarheidsintervallen zien: voor het verschil tussen twee verwachtingen, voor de correlatiecoëfficiënt, voor de regressiecoëfficiënt, voor proporties, enz.

Voor veel (maar niet alle) betrouwbaarheidsintervallen is de formule analoog aan (5.1): het is de schatting van de onbekende parameter θ plus of minus een kritieke waarde van de t-verdeling maal de standaardfout van de schatter. Met andere woorden is het betrouwbaarheidsinterval voor parameter θ gelijk aan

$$\left[\hat{\theta} \pm t_{l;\alpha/2}\ \mathrm{SE}_Q\right], \tag{5.2}$$

waar l het aantal vrijheidsgraden van de t-verdeelde schatter Q.

5.3 Oplossingen

53) Bereken $t_{10;0.05}$ m.b.v. R.

Oplossing:

```
> qt( p = 0.05, df = 10, lower.tail = FALSE )
[1] 1.812461
```

54) Bereken een betrouwbaarheidsinterval voor $\mu_{\texttt{tijd}}$ op basis van `sportData`, met een betrouwbaarheid van 95%.

Oplossing: We weten niet of de verdeling van `tijd` normaal is, maar we mogen de benadering gebruiken omdat $n > 30$. We beginnen met het betrouwbaarheidsinterval met een betrouwbaarheid van 95%.

```
> n <- dim( sportData )[1]
> mean( sportData$tijd ) - qt( p = 0.025, df = n - 1, lower.tail = FALSE )
* sd(sportData$tijd)/sqrt(n)
[1] 22.27036
> mean( sportData$tijd ) + qt( p = 0.025, df = n - 1, lower.tail = FALSE)
* sd( sportData$tijd ) / sqrt( n )
[1] 23.57964
```

We besluiten dus dat $\mu_{\texttt{tijd}} \in [22.3, 23.6]$. We berekenen u het betrouwbaarheidsinterval met een betrouwbaarheid van 90%.

```
> n <- dim( sportData )[1]
> mean( sportData$tijd ) - qt( p = 0.05, df = n - 1, lower.tail = FALSE )
* sd(sportData$tijd)/sqrt(n)
[1] 22.37639
> mean( sportData$tijd ) + qt( p = 0.05, df = n - 1, lower.tail = FALSE)
* sd( sportData$tijd ) / sqrt( n )
[1] 23.47361
```

We besluiten dus dat $\mu_{\texttt{tijd}} \in [22.4, 23.5]$.

Dit betrouwbaarheidsinterval is smaller dan de vorige omdat de betrouwbaarheid nu lager is (90% i.p.v. 95%). Het betrouwbaarheidsinterval is preciezer, maar de betrouwbaarheid is lager.

Hoofdstuk 6

De statistische toetsen

6.1 Zijn de studenten van de FPPW slimmer ?

Je wil nagaan of de doorsnee student van de FPPW slimmer is dan de doorsnee Vlaming. Je kent de verwachting van het IQ van Vlamingen: het is 100. Maar je kent de verwachting $\mu_{\texttt{iq}}$ van het IQ van FPPW studenten niet. Je trekt dus een steekproef van 30 FPPW studenten en je meet hun IQ. Je vindt de gegevens in het data frame `myData`. Je berekent het gemiddelde IQ in de steekproef:

```
> mean(myData$iq)
[1] 117.2667
```

Het gemiddelde in de steekproef van FPPW studenten is groter dan de verwachting in de Vlaamse populatie: $\overline{\texttt{iq}} > 100$. Op eerste zicht geeft dit de indruk dat FPPW studenten inderdaad slimmer zijn dan de doorsnee Vlaming. M.a.w. ben je geneigd om te denken dat $\mu_{\texttt{iq}} > 100$. Maar je hebt een kleine steekproef getrokken en het is dus perfect plausibel dat bovenstaande ongelijkheid een gevolg van het toeval is eerder dan van een werkelijk verschil tussen FPPW studenten en gewone Vlamingen. Met andere woorden, je hebt misschien toevallig veel slimme studenten getrokken; zou je een andere steekproef trekken, dan zou het verschil misschien veel kleiner of zelfs omgekeerd zijn. We moeten dus voorzichtig zijn als we een bevinding op het niveau van een steekproef willen veralgemenen naar een populatie.

Indien het steekproefgemiddelde $\overline{\texttt{iq}}$ gelijk aan 130 zou zijn, i.p.v. 117, dan zou je niet twijfelen: zo'n groot verschil (t.o.v. 100) kan bijna niet het resultaat van het toeval zijn. Je zou dus besluiten dat FPPW studenten gemiddeld gezien slimmer zijn dan Vlamingen in het algemeen ($\mu_{\texttt{iq}} > 100$).

Indien het steekproefgemiddelde $\overline{\texttt{iq}}$ gelijk aan 101 zou zijn, i.p.v. 117, dan zou je ook niet twijfelen: zo'n klein verschil (t.o.v. 100) kan zeker het resultaat van het toeval zijn. Je zou dus besluiten dat FPPW studenten niet slimmer zijn ($\mu_{\texttt{iq}} = 100$).

Met een gemiddelde gelijk aan 117 zit je dus in een grijze zone. Je kan moeilijk beslissen. Je weet niet wat de kans is om een steekproefgemiddelde

gelijk aan 117 te observeren, indien $\mu_{\texttt{iq}} = 100$. Is deze kans klein, dan zal je zeker de hypothese "$\mu_{\texttt{iq}} = 100$" verwerpen. Zo nee, dan zal je de hypothese "$\mu_{\texttt{iq}} = 100$" aanvaarden. Met behulp van de wetten van het kansrekenen is het mogelijk om deze kans te berekenen. Vanaf hier gebruiken we het symbool X (toevalsvariabele) of x (geobserveerde variabele) voor het IQ.

6.1.0.1 Bijkomende hypothese: $\sigma_X = 15$

We maken nu twee hypotheses. De eerste hypothese is dat we de variantie van de toevalsvariabele X kennen: het is 15^2, zoals in de Vlaamse populatie. Deze hypothese heeft als doel de berekeningen te vereenvoudigen. Binnenkort zullen we zien dat we deze hypothese niet nodig hebben. De tweede hypothese is belangrijker en maakt intrinsiek deel uit van de procedure om na te gaan of FPPW studenten slimmer zijn dan de doorsnee Vlaming. We kennen μ_X niet maar we gaan voorlopig veronderstellen dat $\mu_X = 100$ zoals in de Vlaamse populatie. Indien onze hypothese correct is (i.e., $\mu_X = 100$), dan kennen we de verdeling van $\overline{X}$ (Rubr. 3.4):

$$\overline{X} \sim N(\mu_X, \sigma_X^2/n) = N(100, 15^2/30).$$

We weten dus welke waarden van $\overline{X}$ frequent voorkomen en de welke onwaarschijnlijk zijn (onder de hypothese $\mu_X = 100$). We kunnen de kans berekenen dat $\overline{X} \geq 117$, onder de hypothese $\mu_X = 100$. Dat is

$$P(N(100, 15^2/30) \geq 117).$$

Laten we deze kans berekenen:

```
> pnorm(mean=100, sd=sqrt(15^2/30), q=117,  lower.tail= FALSE)
[1] 2.691323e-10
```

Wat betekent deze kans? Het is de kans (indien $\mu_X = 100$) om een steekproef te trekken waar het gemiddelde groter dan of gelijk aan 117 is. Dus, als FPPW studenten niet slimmer zijn en als we veel steekproeven trekken, dan gaan slechts 3 steekproeven op tien miljard een gemiddelde groter dan 117 hebben. M.a.w., zo'n een groot gemiddelde is helemaal niet plausibel indien $\mu_X = 100$. We besluiten dus dat $\mu_X > 100$. We hebben dus een techniek gevonden om de bevinding in de steekproef ($\overline{x} > 100$) te veralgemenen naar de populatie ($\mu_X > 100$).

De kans die we net berekend hebben, heet de overschrijdingskans: de kans dat het gemiddelde in onze steekproef (117) overschreden wordt indien we veel steekproeven trekken. Deze kans wordt ook p-waarde of p-value genoemd.

6.1.0.2 Zonder bijkomende hypothese betreffende σ_X

We kennen de variantie van het IQ in de populatie van alle Vlamingen (het is 15^2), maar niet in de specifieke subpopulatie van FPPW studenten: het hoeft niet 15^2 te zijn. We kunnen de steekproefgrootheid $\overline{X}$ zoals hierboven

niet gebruiken en we gebruiken dus een andere steekproefgrootheid, met een Student verdeling:

$$T = \frac{\overline{X} - \mu_X}{S_X/\sqrt{n}} \sim t_{n-1}.$$

We kennen μ_X ook niet en we gaan nog voorlopig veronderstellen dat $\mu_X = 100$ zoals in de Vlaamse populatie. Derhalve

$$T = \frac{\overline{X} - 100}{S_X/\sqrt{n}} \sim t_{n-1}.$$

Laten we de waarde van T in onze steekproef berekenen:

$$t = \frac{\overline{x} - 100}{s_X/\sqrt{n}}$$

```
> (mean(myData$iq)-100)/sqrt(var(myData$iq)/30)
[1] 6.023087
```

Dus $t = 6.02$. Als dit nul of bijna nul zou zijn, dan zou je zonder twijfel besluiten dat FPPW studenten niet slimmer zijn want $t \approx 0$ impliceert $\overline{x} \approx 100$ en dit biedt geen evidentie dat FPPW studenten slimmer zijn. Is $t = 6.02$ groot genoeg om te besluiten dat FPPW studenten wel slimmer zijn? Is dit een plausibele waarde indien onze hypothese correct is (i.e., $\mu_X = 100$) of is dit een onwaarschijnlijke waarde? Indien onze hypothese correct is (i.e., $\mu_X = 100$), dan kennen we de verdeling van T. We weten dus welke waarden van T frequent voorkomen en de welke onwaarschijnlijk zijn. We berekenen de kans dat $T \geq 6.02$ (als $\mu_X = 100$).

```
> n <- 30
> pt(q= 6.023087, df=n-1, lower.tail= FALSE)
[1] 7.475305e-07
```

Wat betekent deze kans? Het is de kans (indien $\mu_X = 100$) om een steekproef te trekken waar

$$t = \frac{\overline{x} - 100}{s_X/\sqrt{n}} \geq 6.02.$$

Of de kans (indien $\mu_X = 100$) om een steekproef te trekken waar

$$\overline{x} \geq 100 + 6.02 \times s_X/\sqrt{n} = 117.2667.$$

Het is dus opnieuw de overschrijdingskans, de kans om een steekproef te trekken met een gemiddelde groter dan 117. Dus, als FPPW studenten niet slimmer zijn en als we veel steekproeven trekken, dan gaan slechts 7 steekproeven op tien miljoen een gemiddelde groter dan 117 hebben. M.a.w., zo'n een groot gemiddelde is helemaal niet plausibel indien $\mu_X = 100$. We besluiten dus dat $\mu_X > 100$.

Merk op dat de kans die we net berekend hebben (zonder de hypothese $\sigma = 15$) niet dezelfde is als de kans die we vroeger berekend hebben (met de hypothese $\sigma = 15$). Maar ze zijn allebei zeer klein en leiden allebei tot de verwerping van de hypothese $\mu_X = 100$. Het is niet altijd zo.

6.2 To be or not to be

In bovenstaand voorbeeld moeten we kiezen tussen twee alternatieven, twee hypothesen : FPPW studenten zijn slimmer dan de doorsnee Vlaming *of* niet. Dit soort toestand is zeer gewoon. We willen weten of een theorie geldt of niet. Zijn de variabelen X en Y gecorreleerd of niet. Is de proportie π groter in de populatie A dan in populatie B of niet, ...

Het is zelden mogelijk dit soort vragen met zekerheid te beantwoorden. Als de verschijnselen die we observeren niet zo veranderlijk zouden zijn, dan zou het gemakkelijk zijn. Bijvoorbeeld, als het IQ van alle Vlamingen (behalve FPPW studenten) 100 zou zijn en 117 bij alle studenten in onze steekproef, zouden we met zekerheid kunnen zeggen dat FPPW studenten slimmer zijn dan gewone Vlamingen. Maar het is niet zo. Bijna alle fenomenen die we observeren zijn zeer veranderlijk.

Inductieve statistiek speelt hier een belangrijke rol. Het helpt ons een beslissing te nemen. Wegens de veranderlijkheid weten we nooit of onze observaties het resultaat van het toeval of van iets anders zijn. Inductieve statistiek geeft ons de mogelijkheid om de kans van het "toeval" alternatief te berekenen. In het voorbeeld van het IQ van FPPW studenten, kunnen we de kans berekenen dat het hoge gemiddelde IQ in de steekproef een gevolg van het toeval is.

Maar de inductieve statistiek gaat niet verder. Het zegt ons niet wat het correcte alternatief is. Het zegt ons niet wat we moeten beslissen. We kennen de waarheid niet. De inductieve statistiek kent ze ook niet. Wanneer we tussen twee alternatieven moeten kiezen, kunnen we de rol van het toeval kwantificeren en onze beslissing nemen, dankzij de inductieve statistiek, met kennis van zaken.

We zullen nooit met zekerheid weten of FPPW studenten even slim zijn. We kunnen nog tientallen steekproeven trekken. Ze zullen misschien de hypothese van een verschil confirmeren. Maar we kunnen niet zeker zijn dat, als we nog andere steekproeven trekken, de conclusie niet omgekeerd zal zijn.

6.3 De toetsingsprocedure

De redenering die we in het voorbeeld van het IQ hebben gevolgd kan gebruikt worden om veel andere soorten hypothesen te toetsen, maar de precieze berekeningen en de gebruikte steekproefgrootheden moeten telkens aangepast worden naar de context. In verband met de verscheidene contexten zullen we dus over verschillende toetsen beschikken om tussen twee hypotheses te kiezen.

De theorie van de statistische toetsen formaliseert de verschillende *toetsen* en de bijbehorende begrippen. Hieronder formaliseren we de verschillende stappen van een *toetsingsprocedure*.

6.3.1 Theoretische hypothese

Eerst heb je een theoretische hypothese. Bijvoorbeeld, FPPW studenten zijn slimmer dan gewone Vlamingen; mannen en vrouwen lopen niet even snel; er

is een verband tussen opleiding en inkomen; mannen roken meer dan vrouwen; enz.

6.3.2 Statistische hypothese H_a of alternatieve hypothese

Dan vertaal je de theoretische hypothese in de taal van de kansrekening. Je bekomt een *alternatieve hypothese* (symbool H_a) met betrekking tot een parameter van de populatie. Bijvoorbeeld,

- X is de toevalsvariabele "IQ" bij FPPW studenten.

$$H_a : \mu_X > 100.$$

- X is de toevalsvariabele "tijd om 100 m te lopen" bij mannen; Y is de toevalsvariabele "tijd om 100 m te lopen" bij vrouwen.

$$H_a : \mu_X \neq \mu_Y.$$

- X is de toevalsvariabele "opleiding." Y is de toevalsvariabele "inkomen."

$$H_a : \rho_{X,Y} \neq 0.$$

- De parameter π_m is de proportie van mannen die roken. De parameter π_v is de proportie van vrouwen die roken.

$$H_a : \pi_m > \pi_v.$$

Merk op dat de vertaling niet altijd evident is. In het voorbeeld van de tijd om 100m te lopen zouden we ook de vertaling

$$MD(X) \neq MD(Y)$$

kunnen gebruiken, waarbij $MD(X)$ voor de mediaan van de toevalsvariabele X staat.

Merk ook het verschil op tussen de *eenzijdige* en *tweezijdige* toetsen. In een tweezijdige toets is de alternatieve hypothese onder de vorm

$$\ldots \neq \ldots$$

Bijvoorbeeld, $\mu_X \neq \mu_Y$. Men verwacht een verschil tussen de twee parameters maar men weet niet in welke richting. Beide zijn mogelijk.

In een eenzijdige toets is de alternatieve hypothese onder de vorm

$$\ldots > \ldots \text{ of } \ldots < \ldots$$

Bijvoorbeeld, $\mu_X > 100$. Men verwacht een verschil tussen de twee parameters in een bepaalde richting.

6.3.3 Nulhypothese H_0

De *nulhypothese* is de tweede hypothese, het tweede alternatief. Ze moet strijdig zijn met de alternatieve hypothese. Als H_0 juist is, dan moet H_a verkeerd zijn. Bijvoorbeeld,

- *IQ.* $H_0 : \mu_X = 100.$
- *Lopen.* $H_0 : \mu_X = \mu_Y.$
- *Opleiding/inkomen.* $H_0 : \rho_{X,Y} = 0.$
- *Rokers.* $H_0 : \pi_m = \pi_v.$

We gaan de nulhypothese gebruiken om te proberen de alternatieve hypothese te verwerpen, om te tonen dat de observaties gewoon het effect van het toeval zijn. Bijgevolg, onder H_0 moet het mogelijk zijn om kansen te berekenen. Hieruit volgt dat H_0 niet onder de vorm $\ldots < \ldots$ of $\ldots \neq \ldots$ kan voorkomen. Bijvoorbeeld, als de nulhypothese "$\mu_X < 100$" is, dan weten we niet welke waarde van μ_X we in de berekeningen moeten gebruiken. Daartegenover staat dat als de nulhypothese "$\mu_X = 100$" is, dan is de waarde van μ_X wel bepaald : ze is 100.

6.3.4 Eerste beslissing

Het gebeurt soms dat de steekproef helemaal geen steun biedt voor de alternative hypothese (bv. het gemiddelde I.Q. in de steekproef van FPPW studenten is kleiner dan 100). Dan heb je eigenlijk geen statistische toets nodig: je mag rechtsreeks de nulhypothese aanvaarden.

Je gaat wel verder met de toetsingsprocedure indien de steekproef de alternatieve hypothese in zekere mate steunt.

55. Je vermoedt dat basketbalspelers langer zijn dan de anderen (gemiddeld gezien). Wordt deze hypothese gesteund door het data frame `sportData`*?*

6.3.5 Toetsingsgrootheid G

We kiezen nu een steekproefgrootheid (of statistiek) die gerelateerd is aan de onbekende parameter van de hypothese. We noemen ze *toetsingsgrootheid*. Onder de nulhypothese mag de toetsingsgrootheid geen onbekende parameter bevatten en zijn steekproevenverdeling moet bekend zijn. In het voorbeeld van het IQ gebruiken we de toetsingsgrootheid

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

Onder H_0 is

$$\frac{\overline{X} - 100}{S_X/\sqrt{n}}$$

Student-verdeeld met $n - 1$ vrijheidsgraden.

6.3.6 Overschrijdingskans of p-waarde

Onze steekproef steunt de alternatieve hypothese in zekere mate (zo neen, dan heb je de procedure al stopgezet bij stap 6.3.4. We berekenen de kans om een steekproef te trekken die onze alternatieve hypothese even sterk of zelfs sterker ondersteunt dan onze steekproef. Hierbij moeten we een onderscheid tussen een- en tweezijdige toetsen maken.

6.3.6.1 Eenzijdige toets

Welke waarde van G zou leiden tot een verwerping van de nulhypothese? Een hoge waarde (zoals in het IQ voorbeeld) of een lage waarde? Indien hoge waarden evidentie bieden tegen de nulhypothese, dan berekenen we de overschrijdingskans

$$P(\, G \geq g \,). \tag{6.1}$$

Indien lage waarden evidentie bieden tegen de nulhypothese, dan berekenen we de overschrijdingskans

$$P(\, G \leq g \,). \tag{6.2}$$

6.3.6.2 Tweezijdige toets

Hier bieden hoge *en* lage waarden evidentie tegen de nulhypothese. Indien $g < E(G)$ (je hebt dus een lage waarde geobserveerd), dan bereken je de overschrijdingskans

$$2P(\, G \leq g \,). \tag{6.3}$$

Indien $g > E(G)$ (je hebt dus een hoge waarde geobserveerd), dan bereken je de overschrijdingskans

$$2P(\, G \geq g \,). \tag{6.4}$$

6.3.7 Beslissing

Indien de p-waarde klein is, dan is onze steekproef moeilijk compatibel met de nulhypothese: het is niet waarschijnlijk om zo een steekproef te trekken indien de nulhypothese correct is. We verwerpen dan de nulhypothese.

Wat is een kleine kans? Onderzoekers gebruiken meestal een drempel van 5%. Soms wordt ook 10% of 1% gebruikt. In de humane wetenschappen is 5% de norm (zoals in het IQ-voorbeeld). Deze drempel wordt aangeduid door het symbool α en wordt de onbetrouwbaarheidsdrempel of het significantieniveau of gewoon de significantie genoemd.

Het is mogelijk aan te tonen dat de toetsingsprocedure leidt tot onterechte verwerpingen van de nulhypothese met een kans gelijk aan α.

6.4 De toetsingsprocedure in actie

Om de toetsingsprocedure duidelijk te maken gaan we opnieuw het IQ-voorbeeld analyseren.

Theoretische hypothese De theoretische hypothese in dit voorbeeld is dat de FPPW studenten slimmer zijn dan de doorsnee Vlaming.

Alternatieve hypothese We noemen X het IQ van de FPPW studenten. Er zijn verscheidene mogelijke alternatieve hypothesen. We kunnen beweren dat de modus van de toevalsvariabele X verschillend is van de modus van het IQ in Vlaanderen. We kunnen ook de medianen of de gemiddelden vergelijken.

Om deze keuze te doen moeten we voorzichtig nadenken. Wat zijn de eigenschappen van de verschillende maten ? Welke van die maten is het best geschikt in ons geval. Is de gevoeligheid aan outliers een probleem of niet ? Wat is het meetniveau van onze variabele ?

In dit voorbeeld hebben we een variabele van intervalniveau. Het vergelijken van gemiddelden is dus zinvol. We kiezen dus de volgende alternatieve hypothese.

$$H_A : \ \mu_X > 100.$$

De toets is dus een éénzijdige toets.

Nulhypothese De nulhypothese is $H_0 : \ \mu_X = 100$.

Eerste beslissing Het gemiddelde in de steekproef is 117. Dit is groter dan 100 en steunt dus de alternatieve hypothese. We moeten verder gaan met de toetsingsprocedure.

Toetsingsgrootheid

$$T = \frac{\overline{X} - \mu_X}{S_X/\sqrt{n}}$$

is een steekproefgrootheid die afhankelijk is van de hypothesen. Zijn steekproevenverdeling is bekend : ze is Student-verdeeld. Onder H_0 is er geen onbekende parameter: de verwachting is $\mu_X = 100$. Deze steekproefgrootheid kan dus onze toetsingsgrootheid zijn.

p-waarde Waarden van de toetsingsgrootheid T die evidentie bieden tegen de nulhypothese zijn hoge waarden van T: ze komen overeen met waarden van $\overline{X}$ die groter zijn dan 100. De p-waarde is dus

$$P\left(T \sim t_{n-1} \geq \frac{\overline{x} - 100}{s_X/\sqrt{n}}\right).$$

Deze kans wordt berekend met

```
> n <- 30
> pt(q= 6.023087, df=n-1, lower.tail= FALSE)
[1] 7.475305e-07
```

Beslissing De p-waarde is kleiner dan $\alpha = 0.05$. Bijgevolg verwerpen we de nulhypothese.

6.5 De keuze van de toetsingsgrootheid

De keuze van een toetsingsgrootheid is niet gemakkelijk. Soms is het moeilijk een geschikte toetsingsgrootheid te vinden; soms zijn er meerdere en dan is de keuze ook moeilijk. Omdat de keuze moeilijk is, gaan we in de praktijk nooit de toetsingsgrootheid zelf vinden. Er zijn boeken waar, voor elke toestand, voor elke toets, de geschiktste toetsingsgrootheid voorgesteld wordt. In deze cursus worden hieronder enkele toetsingsgrootheden voorgesteld.

6.5.1 Het toetsen van een hypothese betreffende μ

Dit is het geval dat in het IQ-voorbeeld behandeld wordt. Men wil toetsen of de verwachting van een toevalsvariabele verschillend is van een bepaalde waarde (deze bepaalde waarde is vaak de verwachting van dezelfde variabele in een andere populatie). We onderscheiden twee gevallen.

6.5.1.1 σ is bekend

De toetsingsgrootheid is $\overline{X}$. Zijn verdeling is de standaardnormale verdeling $N(\mu_X, \sigma_X^2/n)$. Deze toets wordt de z-toets voor één steekproef genoemd. In de praktijk wordt het bijna nooit gebruikt omdat σ bijna nooit bekend is.

Voorwaarden: X moet tenminste van intervalniveau zijn. X moet normaal verdeeld zijn of de steekproef moet groot zijn.

6.5.1.2 σ is onbekend

De toetsingsgrootheid is

$$\frac{\overline{X} - \mu_X}{S_X/\sqrt{n}}$$

Zijn verdeling is de Student-verdeling met $n - 1$ vrijheidsgraden. Deze toets wordt de t-toets voor één steekproef genoemd.

Voorwaarden: X moet tenminste van intervalniveau zijn. X moet normaal verdeeld zijn of de steekproef moet groot zijn.

`R` biedt een speciale functie om de t-toets gemakkelijk uit te voeren: `t.test`.

Vb. Zijn FPPW studenten even dik als de doorsnee Vlaming? De verwachting van de Body Mass Index (BMI)[1] van Vlamingen is bekend:[2] het is 25.3 [Charafeddine et al., 2011a]. De verwachting van de BMI van FPPW studenten is natuurlijk onbekend, maar we beschikken over gegevens m.b.t. een steekproef van 30 studenten en we gaan deze gegevens gebruiken om te proberen de onderzoeksvraag (Zijn FPPW studenten even dik als de doorsnee Vlaming?) te beantwoorden. We stellen eerst een nulhypothese op: $\mu_X = 25.3$, waar X de BMI van FPPW studenten representeert. De alternatieve hypothese is dan $H_a : \mu_X \neq 25.3$.

Laten we de gemiddelde BMI in de steekproef berekenen, dus $\overline{x}$.

```
> bmi <- myData$gewicht / (myData$lengte/100)^2
> mean(bmi)
[1] 25.74648
```

We vinden $\overline{x} = 25.7$. Het is niet gelijk aan 25.3 en dit steunt dus de alternatieve hypothese in zekere mate. Maar dit is misschien toevallig: misschien is de nulhypothese correct en hebben we toevallig een steekproef getrokken met veel dikke mensen. Om een beslissing te maken gebruiken we een t-toets voor één steekproef (dit mag want X is van ratio meetniveau en $n = 30$).

```
> t.test(x = bmi, mu = 25.3, alternative = "two.sided")

One Sample t-test

data:  bmi
t = 0.5307, df = 29, p-value = 0.5997
alternative hypothesis: true mean is not equal to 25.3
95 percent confidence interval:
 24.02585 27.46711
sample estimates:
mean of x
 25.74648
```

We bespreken eerst het commando (de eerste regel). Het argument `x` duidt de data aan en `mu`, de nulhypothese. Met het argument `alternative` kunnen we de alternatieve hypothese specificeren: `"greater"` of `"less"` voor éénzijdige toetsen; `"two.sided"` voor tweezijdige toetsen.

We bespreken nu enkele elementen van de output. `t = 0.5307` geeft de waarde van de toetsingsgrootheid T. `df = 29` is het aantal vrijheidsgraden $(n - 1)$. Je vindt dan de overschrijdingskans: `p-value = 0.5997`. De volgende regel geeft de alternatieve hypothese:
`alternative hypothesis: true mean is not equal to 25.3`.

[1] De BMI is gelijk aan het lichaamsgewicht (in kg) gedeeld door de lichaamslengte (in m) in het kwadraat.

[2] Eigenlijk is de verwachting niet bekend, maar de schatting van de verwachting is gebaseerd op zodanig veel data dat de schatting zeer precies en betrouwbaar is en we kunnen ze beschouwen als de echte verwachting.

De volgende twee regels geven het betrouwbaarheidsinterval voor μ_X: $[24.02585, 27.46711]$. De laatste drie regels hebben betrekking tot de schatting van μ_X: $\hat{\mu}_X = 25.74648$.

Laten we de p-waarde gebruiken om een beslissing te maken. De p-waarde is 59%. Wat betekent dit? Stel dat de nulhypothese correct is (FPPW studenten zijn even dik) en stel dat je veel steekproeven van 30 FPPW studenten trekt. Dan gaan 59% van die steekproeven de alternatieve hypothese steunen even veel of nog sterker dan onze steekproef (in `myData`). Met andere woorden is onze steekproef perfect compatibel met de nulhypothese en we gaan dus de nulhypothese aanvaarden.

Laten we nu het betrouwbaarheidsinterval bespreken. Op basis van het betrouwbaarheidsinterval besluiten we dat μ_X binnen het interval $[24.0, 27.5]$ ligt. Dit biedt ons een tweede manier om de nulhypothese te toetsen: de vooropgestelde waarde 25.3 ligt binnen het betrouwbaarheidsinterval dat `R` berekend heeft op basis van de steekproef. De vooropgestelde waarde 25.3 is dus compatibel met de gegevens en we hebben dus geen reden om de nulhypothese te verwerpen.

Het is mogelijk te bewijzen dat de beslissing die je neemt op basis van het betrouwbaarheidsinterval altijd identiek is aan de beslissing die je neemt op basis van de p-waarde. We beschikken dus over twee equivalente technieken om een hypothese te toetsen.

We bespreken nu een extra argument van de functie `t.test`. Naast de drie argumenten (`x`, `mu`, `alternative`) die we gezien hebben, mogen we nog het argument `conf.level` gebruiken, om een betrouwbaarheid verschillend van 95% te gebruiken. Vb.

```
> t.test(x = bmi, mu = 25.3, alternative = "two.sided",
  conf.level = 0.9)

One Sample t-test

data:  bmiFPPW
t = 0.5307, df = 29, p-value = 0.5997
alternative hypothesis: true mean is not equal to 25.3
90 percent confidence interval:
 24.31702 27.17593
sample estimates:
mean of x
 25.74648
```

Het enige verschil in de output is het betrouwbaarheidsinterval. Het is nu $[24.31702, 27.17593]$; het is smaller omdat het minder betrouwbaar is.

Als we het argument `conf.level` niet gebruiken dan gaat `R` ervan uit dat we een betrouwbaarheid van 0.95 wensen te gebruiken. De waarde 0.95 wordt de 'default value' genoemd.

Vb. Zijn FPPW studenten slimmer dan de doorsnee Vlaming? Om deze vraag te beantwoorden gebruiken we opnieuw de functie `t.test` (dit mag want `iq` is van interval meetniveau en is normaal verdeeld).

```
> t.test(x = myData$iq,  mu = 100, alternative = "greater" )

One Sample t-test

data:  myData$iq
t = 6.0231, df = 29, p-value = 7.475e-07
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 112.3957      Inf
sample estimates:
mean of x
 117.2667
```

De output is zoals vroeger toen we de hypothese m.b.t. de BMI hebben getoetst. We bespreken toch de twee regels die het betrouwbaarheidsinterval voor $\mu_{\texttt{iq}}$ rapporteren. Het is een eenzijdig betrouwbaarheidsinterval omdat de alternatieve hypothese eenzijdig is: $[112.3957, \infty[$. Dit soort betrouwbaarheidsinterval wordt niet vaak gebruikt.

Met de informatie in de output kan je een beslissing nemen, op twee verschillende en equivalente manieren: (a) de p-waarde is kleiner dan 0.05 en je moet dus de nulhypothese verwerpen; (b) de veronderstelde waarde 100 ligt niet binnen het betrouwbaarheidsinterval en je moet dus de nulhypothese verwerpen.

6.5.2 Het toetsen van een hypothese betreffende twee verwachtingen

Je wil nagaan of mannen gemiddeld gezien meer verdienen dan vrouwen. Je trekt dus twee aselecte steekproeven: één van vrouwen en één van mannen. Je berekent het gemiddelde in beide steekproeven en je komt $\overline{x}_v < \overline{x}_m$ uit. Dit wijst aan dat mannen inderdaad meer verdienen dan vrouwen, maar dit zou toevallig kunnen zijn, omdat je steekproef misschien niet representatief is. Je wil dus de alternatieve hypothese $\mu_m > \mu_v$ statistisch toetsen. De nulhypothese is $\mu_m = \mu_v$ of $\mu_m - \mu_v = 0$.

We onderscheiden twee werkwijzen, in functie van de manier waarop de steekproeven getrokken worden.

- Je trekt een steekproef van n hetero echtparen en je beschikt dus over n scores voor de vrouwen en n scores voor de *overeenkomende* mannen.

- Je trekt een steekproef van n_v vrouwen en, los van deze, een andere steekproef van n_m mannen (n_v en n_m kunnen maar hoeven niet identiek te zijn).

De twee situaties zijn sterk verschillend en dienen met verschillende technieken geanalyseerd te worden. In het eerste geval zijn de lonen van de mannen en die van de vrouwen gecorreleerd (mannen met hoge lonen leven vaak met vrouwen met hoge lonen); in het tweede geval zijn ze niet gecorreleerd: er is geen verband tussen de n_v vrouwen en de n_m mannen. De eerste situatie wordt "afhankelijke steekproeven" genoemd (Engels: paired samples). De tweede situatie wordt "onafhankelijke steekproeven" genoemd (Engels: independent samples).

De vergelijking van twee verwachtingen, met afhankelijke steekproeven, is zeer frequent in de humane wetenschappen. Hieronder vind je een paar voorbeelden. Een variabele wordt twee keer gemeten bij elke individu in een steekproef: ééns voor een manipulatie, ééns na de manipulatie. De manipulatie kan een training zijn, een therapie, het toedienen van een geneesmiddel, het kijken naar beeldstimuli, een les, enz. Of de steekproef bestaat uit paren (dyads in het engels) en een variabele wordt gemeten bij elke persoon van elk paar. De paren kunnen bestaan uit kind-moeder, lesgever-leerling, werknemer-werkgever, man-vrouw, patiënt-therapeut, enz.

In het vervolg spreken we van de variabelen X_1 en X_2, met verwachtingen μ_1 en μ_2.

6.5.2.1 Onafhankelijke steekproeven

We onderscheiden nu drie gevallen in functie van min of meer restrictieve hypotheses m.b.t. de varianties. We beginnen met de meeste restrictieve hypothese.

σ_1 en σ_2 zijn bekend De toetsingsgrootheid is dan

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Deze toets wordt de z-toets voor twee steekproeven genoemd. Deze toets wordt bijna nooit gebruikt omdat de varianties bijna nooit bekend zijn. Het wordt niet verder gezien.

σ_1 en σ_2 zijn gelijk maar onbekend De toetsingsgrootheid is

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}. \tag{6.5}$$

Deze toets wordt de t-toets voor twee onafhankelijke steekproeven genoemd. Deze toets wordt tegenwoordig zelden gebruikt omdat we nooit echt weten of de varianties identiek zijn. Er bestaat wel een andere toets om de gelijkheid van de varianties te toetsen, maar het is toch afgeraden om de t-toets voor twee onafhankelijke steekproeven te gebruiken.

Geen hypothese m.b.t. σ_1 en σ_2 De toetsingsgrootheid is

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_l,$$

met

$$l = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}.$$

Deze toets wordt de Welch[3] t-toets voor twee onafhankelijke steekproeven genoemd. Hij wordt ook vaak t-toets tout court genoemd.

Voorwaarden. X moet tenminste van intervalniveau zijn. X moet normaal verdeeld zijn in beide populaties of beide steekproeven moeten groot zijn.

Vb. Lopen mannen en vrouwen even snel? Je wil dit nagaan a.d.h.v. de gegevens in het data frame `sportData`. De alternatieve hypothese is $\mu_{\texttt{tijd},m} \neq \mu_{\texttt{tijd},v}$. De nulhypothese is $\mu_{\texttt{tijd},m} = \mu_{\texttt{tijd},v}$. We gaan eerst twee vectoren aanmaken met de scores van de variabele `tijd` voor de vrouwen en voor de mannen.

```
> tijdV <- sportData$tijd[sportData$geslacht == "V"]
> tijdM <- sportData$tijd[sportData$geslacht == "M"]
```

We berekenen de twee gemiddelden:

```
> mean(tijdV)
[1] 23.20472
> mean(tijdM)
[1] 22.60957
```

Dit wijst aan dat mannen sneller lopen, maar dit kan toevallig zijn. We gaan dit statistisch toetsen m.b.v. de Welch t-toets (dit mag want de tijd is van ratio meetniveau, $n_1 > 30$ en $n_2 > 30$). We voeren de Welch t-toets uit met de `R` functie `t.test`. Het is dezelfde functie als vroeger, maar de argumenten zijn nu anders.

```
> t.test(x=tijdM, y=tijdV, alternative = "two.sided")

Welch Two Sample t-test

data:  tijdM and tijdV
t = -0.89953, df = 197.89, p-value = 0.3695
alternative hypothesis: true difference in means is not equal to 0
```

[3] B. L. Welch, 1911–1989

```
95 percent confidence interval:
 -1.8998569  0.7095718
sample estimates:
mean of x mean of y
 22.60957  23.20472
```

We bespreken eerst het commando (eerste twee regels hierboven). De twee eerste argumenten van de functie `t.test` wijzen de gegevens aan (`x=tijdM, y=tijdV`). Het volgende argument (`alternative = "two.sided"`) ken je al.

We bespreken nu enkele elementen uit de output. Je vindt de waarde van de toestingsgrootheid (6.5): `t = -0.89953`. De p-value is 0.3695. Dus, indien de nulhypothese correct is (vrouwen en mannen lopen even snel) en indien je veel steekproeven trekt, dan zal je in 37% van de steekproeven een toetsingsgrootheid bekomen die groter dan of gelijk is aan 0.89953, in absolute waarde[4]. Dit betekent dat, onder H_0, de bekomen waarde 0.89953 helemaal niet uitzonderlijk is ($0.37 > 0.05 = \alpha$). We hebben dus geen reden om de nulhypothese te verwerpen. De output geeft nog het 95%-betrouwbaarheidsinterval voor het verschil tussen de twee verwachtingen: van -1.90 tot 0.71. Dit interval bevat het getal 0. Het is dus plausibel dat de twee verwachtingen niet van elkaar verschillen. We besluiten dus dat vrouwen even snel als mannen lopen.

Vb. Zijn mannen dikker dan vrouwen? Je wil dit nagaan a.d.h.v. de gegevens in het data frame `sportData`. De alternatieve hypothese is $\mu_{\texttt{bmi},m} > \mu_{\texttt{bmi},v}$. De nulhypothese is $\mu_{\texttt{bmi},m} = \mu_{\texttt{bmi},v}$. We gaan eerst de BMI berekenen en twee vectoren aanmaken met de scores van de variabele `bmi` voor de vrouwen en voor de mannen.

```
> bmi <- sportData$gewicht/(sportData$lengte/100)^2
> bmiV <- bmi[sportData$geslacht == "V"]
> bmiM <- bmi[sportData$geslacht == "M"]
```

We berekenen de twee gemiddelden:

```
> mean(bmiV)
[1] 28.33743
> mean(bmiM)
[1] 28.01112
```

De gemiddelde BMI van mannen ligt dus lager dan de gemiddelde BMI van vrouwen in de steekproef. De gegevens steunen de alternatieve hypothese dus helemaal niet. Derhalve hoeven we geen statistische toets te gebruiken en we aanvaarden de nulhypothese.

[4] De absolute waarde van een getal is dat getal zonder zijn teken. Bv. de absolute waarde van -3 is 3; de absolute waarde van 5 is 5; de absolute waarde van $-1/3$ is $1/3$; de absolute waarde van 0.7 is 0.7.

Vb. Zijn mannelijke FPPW studenten even lang als vrouwelijke FPPW studenten? Je wil dit nagaan a.d.h.v. de gegevens in het data frame `myData`. De alternatieve hypothese is $\mu_{\texttt{lengte},m} \neq \mu_{\texttt{lengte},v}$. De nulhypothese is $\mu_{\texttt{lengte},m} = \mu_{\texttt{lengte},v}$. We maken twee vectoren aan met de scores van de variabele `lengte` voor de vrouwen en voor de mannen.

```
> lengteV <- myData$lengte[myData$geslacht == "V"]
> lengteM <- myData$lengte[myData$geslacht == "M"]
```

We berekenen de twee gemiddelden:

```
> mean(lengteM)
[1] 164.5714
> mean(lengteV)
[1] 168.5
```

De gemiddelde lengte van mannen ligt dus lager dan de gemiddelde lengte van vrouwen in de steekproef. De gegevens steunen dus de alternatieve hypothese en we gaan de Welch t-toets gebruiken om de hypothese te toetsen ... maar dit mag *niet* want $n_1 < 30$ en $n_2 < 30$ en we weten niet of de lengte normaal verdeeld is. We mogen dus niet verder met de toetsingsprocedure. We hebben te weinig gegevens (de steekproef is te klein) om te kunnen besluiten.

56. Doen mannen even veel aan sport als vrouwen? Toets deze hypothese a.d.h.v. een t-toets, met de gegevens van het data frame `sportData`.

6.5.2.2 Afhankelijke steekproeven

Het probleem met twee afhankelijke steekproeven wordt herleid tot een probleem met één steekproef. We definiëren de toevalsvariabele D als het verschil tussen de scores van de man en van de vrouw of het verschil tussen de scores voor en na de manipulatie: $D = X_1 - X_2$. De geobserveerde verschillen in de steekproef zijn $d_1, \ldots, d_n$. Indien de nulhypothese correct is, t.t.z. indien $\mu_1 = \mu_2$, dan geldt $\mu_D = 0$ (zie Rubr. 3.1.10.4). En omgekeerd. In plaats van "$\mu_1 = \mu_2$" te toetsen gaan we dus "$\mu_D = 0$" toetsen. Deze hypothese heeft betrekking op één verwachting en we beschikken over één steekproef van d-waarden (d staat voor difference). We gebruiken dus de standaard t-toets voor één steekproef.

57. Ga na of de verwachting van het IQ dezelfde is bij mannelijke en vrouwelijke FPPW studenten, a.d.h.v. `myData`.

Voorwaarden. X moet tenminste van intervalniveau zijn. $X_1 - X_2$ moet normaal verdeeld zijn of de steekproef moet groot zijn.

Vb. Maak je meer rijfouten met een pijnstiller? Je wil nagaan of een autobestuurder onder invloed van pijnstillers meer rijfouten maakt. Veertig proefpersonen worden getrokken. Je doet een "dubbel blind" experiment d.w.z. noch de persoon, noch de proefleider weten of er een pijnstiller of een placebo wordt toegediend. De test wordt tweemaal afgenomen. Sommige personen krijgen eerst de pijnstiller, de anderen eerst de placebo.

Het data frame `rijfoutenData` bevat de gegevens. Laten we de data bekijken.

```
> rijfoutenData
   rijfoutenMet rijfoutenZonder
1            26              22
2            24              19
3            22              20
4            33              32
...         ...             ...
39           36              35
40           22              17
```

Er zijn twee variabelen en 40 individuen. Bij elke individu worden twee variabelen geobserveerd: Het aantal rijfouten met de pijnstiller (`rijfoutenMet`) en het aantal rijfouten zonder de pijnstiller (`rijfoutenZonder`). We berekenen het gemiddelde van beide variabelen:

```
> mean(rijfoutenData$rijfoutenMet)
[1] 30
> mean(rijfoutenData$rijfoutenZonder)
[1] 28
```

Dit wijst aan dat autobestuurders meer fouten maken met een pijnstiller dan zonder (in overeenstemming met de alternatieve hypothese). We gaan toetsen of dit gewoon door het toeval verklaard kan worden. De alternatieve hypothese is $\mu_M > \mu_Z$. De nulhypothese is $\mu_M = \mu_Z$. We maken een vector aan, met de verschillen:

```
> d <- rijfoutenData$rijfoutenMet-rijfoutenData$rijfoutenZonder
```

58. Bereken het gemiddelde $\overline{d}$ van de vector `d`*, m.b.v. R. Kan je het resultaat bekomen d.m.v. een eenvoudige redenering?*

We voeren nu een standaard t-toets (`d` is van ratio meetniveau; is `d` normaal verdeeld?):

```
> t.test(x=d, mu = 0, alternative = "greater")

One Sample t-test

data:  d
t = 3.8545, df = 39, p-value = 0.000211
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 1.125761      Inf
sample estimates:
mean of x
        2
```

Bij de eerste regel gebruiken we het argument `alternative = "greater"` want `d` is gedefinieerd als `rijfoutenMet` min `rijfoutenZonder`. Onze alternatieve hypothese voorspelt dat de verwachting van D groter dan 0 is. Dus `greater`. Hadden we `d` gedefinieerd als `rijfoutenZonder` min `rijfoutenMet`, dan moesten we het argument `alternative = "less"` gebruiken.

In de output zie je dat de p-waarde 0.000211 bedraagt en dat het betrouwbaarheidsinterval voor μ_D de waarde nul niet bevat. We moeten dus de nulhypothese verwerpen.

Er is een andere manier, nog eenvoudiger, om deze toets uit te voeren. We hoeven niet de verschillen zelf te berekenen. We gebruiken de functie `t.test` met de argumenten `x = rijfoutenData$rijfoutenMet` en `y = rijfoutenData$rijfoutenZonder` en ook `paired=TRUE`:

```
> t.test(x = rijfoutenData$rijfoutenMet,
y = rijfoutenData$rijfoutenZonder, alternative = "greater",
paired = TRUE)

Paired t-test

data:  rijfoutenData$rijfoutenMet and rijfoutenData$rijfoutenZonder
t = 3.8545, df = 39, p-value = 0.000211
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.125761      Inf
sample estimates:
mean of the differences
                      2
```

De output is vanzelfsprekend. Merk op dat de p-waarde en het betrouwbaarheidsinterval dezelfde zijn als hierboven.

59. Het data frame `FB` *bevat gegevens m.b.t. 172 Facebook gebruikers: geslacht, aantal vrienden, aantal likes in 2016 en 2017. Ga na of vrouwen in het algemeen meer vrienden hebben dan mannen.*

6.6 Het toetsen van een hypothese betreffende een proportie

Men wil toetsen of een proportie in een populatie verschillend is van een bepaalde waarde (deze bepaalde waarde is vaak een proportie in een andere populatie).

Voorbeeld : de proportie van alcoholisten in de populatie tussen 30 en 40 jaar is 8%. Volgens een theorie is deze proportie groter bij werklozen. Je wil dus toetsen of de proportie van alcoholisten bij werklozen tussen 30 en 40 jaar groter is dan 8%. De nulhypothese luidt als '$H_0 : \pi = 0.08$' en de alternatieve hypothese als 'H_a: $\pi > 0.08$'. In een steekproef van 10 werklozen vindt je 1 alcoholist. De geobserveerde proportie is dus $1/10 = 0.10 > 0.08$. Dit steunt de alternatieve hypothese. Maar is dit sterk genoeg om de nulhypothese te verwerpen ? Kan dit niet toevallig zijn ?

60. Ga na of Facebook gebruikers in het algemeen meer likes hebben gedaan in 2017 dan in 2016.

De overschrijdingskans is in dit geval de kans dat je één of meer alcoholisten vindt in een steekproef van tien indien de proportie in de populatie 8% is : $P(B(10, 0.08) \geq 1)$. Met andere woorden, de overschrijdingskans is de kans dat jij je data (1 alcoholist) observeert of data die de alternatieve hypothese nog sterker steunen (meer dan 1 alcoholist). Deze kans kan gemakkelijk berekend

worden.

$$\begin{aligned} P(B(10,0.08) \geq 1) &= 1 - P(B(10,0.08) < 1) \\ &= 1 - P(B(10,0.08) = 0) \\ &= 1 - \frac{10!}{0!\,10!} \times 0.08^0 \times 0.92^{10} \\ &= 1 - 0.43 = 0.57. \end{aligned}$$

De overschrijdingskans is dus niet gering en veel groter dan 5%. Met andere woorden is het zeer aannemelijk dat we 1 alcoholist (of meer) vinden in een steekproef van 10, gewoon bij toeval. Derhalve aanvaarden we de nulhypothese.

De toets die we net gezien hebben wordt de exacte binomiale toets genoemd.

Voorwaarden. Geen specifieke assumptie is noodzakelijk om deze toets te mogen gebruiken.

Het softwarepakket `R` bevat een functie om deze toets uit te voeren:

```
> binom.test(x=1, n=10, p=0.08, alternative = "greater")

Exact binomial test

data:  1 and 10
number of successes = 1, number of trials = 10, p-value = 0.5656
alternative hypothesis: true probability of success is greater than 0.08
95 percent confidence interval:
 0.005116197 1.000000000
sample estimates:
probability of success
                   0.1
```

Merk op dat de p-waarde dezelfde is als wat we hoger hebben gevonden. Let op, het argument `p` van de functie `binom.test` verwijst naar de theoretische proportie (dus π onder de nulhypothese) en niet naar de p-waarde.

61. Twintig procent van de in de laatste 10 jaar afgestudeerde ingenieurs zijn vrouwen. In een bedrijf werden 40 ingenieurs aangeworven in de laatste 10 jaar. Slechts 2 van die ingenieurs zijn vrouwen (5%). Dit geeft de indruk dat deze bedrijf genderdiscriminatie doet. Toets de hypothese dat de bedrijf discrimineert tussen mannen en vrouwen. Eerst met een rekenmachine, dan met R.

6.7 De normaliteitsassumptie

Bij een aantal toetsen hebben we gezien dat de variabele normaalverdeeld moet zijn (normaliteitsassumptie). Dit garandeert dat de toetsingsgrootheid (of Statistiek) een bepaalde theoretische verdeling heeft (vaak Student) en dat de p-waarde die we berekenen correct is.

Indien de steekproef zeer groot is ($n \geq 100$), is deze assumptie niet belangrijk omdat de Centrale Limietstelling garandeert dat de toetsingsgrootheid toch bij benadering de geschikte theoretische verdeling volgt. Indien de steekproef

gewoon groot is ($30 \leq n < 100$), dan is deze assumptie niet echt belangrijk, behalve als de verdeling van X helemaal niet normaal is. En indien de steekproef klein is ($n < 30$), dan is de assumptie echt belangrijk.

We hebben dus een techniek nodig om na te gaan of de verdeling van een toevalsvariabele al dan niet normaal is. Er bestaan statistische toetsen[5] om dit te doen, maar ze werken niet zeer goed wanneer de steekproef klein is: ze hebben een lage power.[6] Ze werken dus niet zeer goed wanneer we ze nodig hebben. Daarom worden ze niet vaak gebruikt en men gebruikt liever een visuele techniek: de normale quantile-quantile plot.

6.7.1 De normale quantile-quantile plot

Deze plot maakt deel uit van de bredere familie van de quantile-quantile plots (meestal afgekort als qq-plot of QQ-plot). Een qq-plot is een grafiek waarbij de kwantielen van de steekproef en die van een theoretische verdeling geplot worden. In het geval van een normale qq-plot worden de kwantielen van de steekproef en die van de standaardnormale verdeling geplot.

Indien de verdeling in de steekproef zeer dicht bij een normale verdeling ligt, dan zullen alle punten van de normale qq-plot bijna op de diagonaal van de grafiek liggen. Naar gelang de verdeling in de steekproef verder ligt van een normale verdeling, gaan de punten van de normale qq-plot verder van de diagonaal van de grafiek liggen. We illustreren dit met een paar voorbeelden.

De kolom `x` in de data frame `qq.steekproef` bevat 200 random getallen die door een softwarepakket gegenereerd werden en die uit een normale verdeling getrokken werden. Hieronder vind je de eerste regels van deze data frame.

```
> head(qq.steekproef)
          x           y         z          w
1  87.57693  0.42834632 6.0785429 -0.1082866
2 114.16227  1.15916344 2.7814090  1.7822712
3  99.09212 -0.76236454 5.2623967  1.9919844
4  95.14886 -0.84775312 1.8983243 -1.7150549
5  81.84619 -0.06543868 1.8835752  1.2099772
6 114.47231 -0.99469718 0.4661606 -3.2831064
```

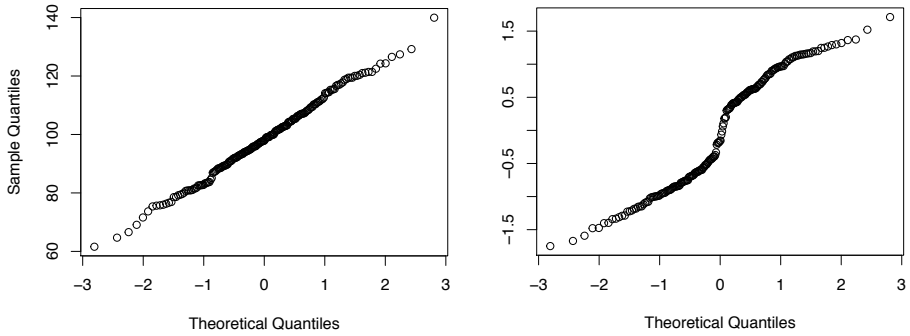
Met de functie `qqnorm` vragen we de normale qq-plot van `x` op:

```
> qqnorm(qq.steekproef$x)
```

De output wordt in Fig. 6.1 links weergegeven. Op deze grafiek zie je duidelijk dat alle 200 punten zeer dichtbij de diagonaal liggen. Dit wijst aan dat het zeer plausibel is dat de steekproef uit een normale verdeling getrokken werd.

De kolom `y` in dezelfde data frame bevat 200 random getallen die door een softwarepakket getrokken zijn uit een *niet*-normale verdeling. Met de functie `qqnorm` vragen we de normale qq-plot van `y` op:

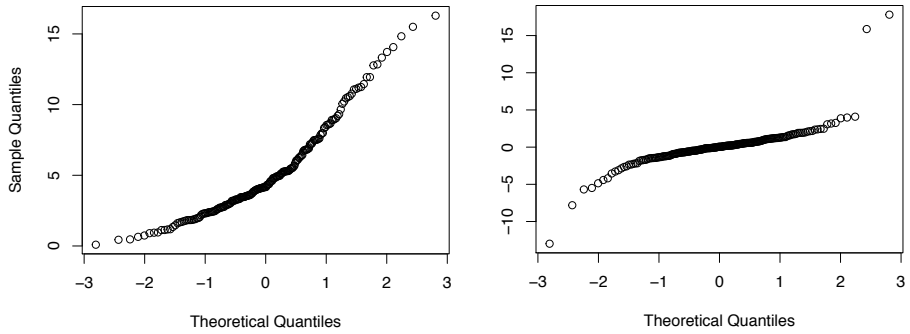
[5] De Kolmogorov-Smirnov toets en de Shapiro-Wilk toets.
[6] Dit concept wordt in het volgende hoofdstuk gezien.



Figuur 6.1: De normale qq-plot voor `qq.steekproef$x` (links) en `qq.steekproef$y` (rechts).

```
> qqnorm(qq.steekproef$y)
```

De output wordt in Fig. 6.1 rechts weergegeven. Op deze grafiek zie je duidelijk



Figuur 6.2: De normale qq-plot voor `qq.steekproef$z` (links) en `qq.steekproef$w` (rechts).

dat de 200 punten niet echt dichtbij de diagonaal liggen en dat er een patroon zit in de afwijkingen. Dit wijst aan dat dat de steekproef waarschijnlijk niet uit een normale verdeling getrokken werd.

De kolom `z` in dezelfde data frame bevat 200 random getallen die door een softwarepakket getrokken zijn uit een *niet*-normale verdeling. Met de functie `qqnorm` vragen we de normale qq-plot van `z` op:

```
> qqnorm(qq.steekproef$z)
```

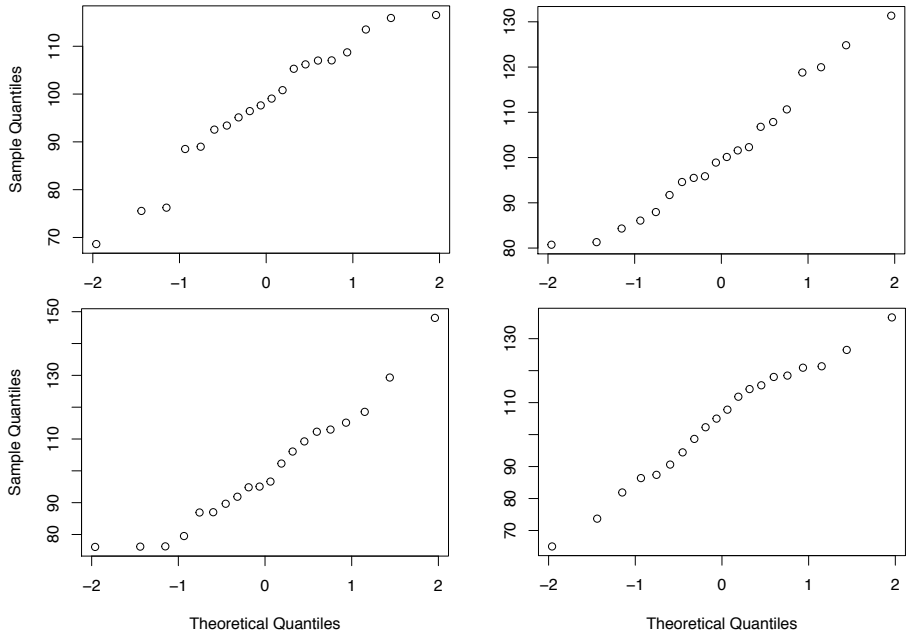
De output wordt in Fig. 6.2 links weergegeven. Op deze grafiek zie je duidelijk dat de 200 punten niet echt dichtbij de diagonaal liggen en dat er ook een patroon zit in de afwijkingen. Dit wijst aan dat dat de steekproef waarschijnlijk niet uit een normale verdeling getrokken werd.

De kolom `w` in dezelfde data frame bevat 200 random getallen die door een softwarepakket getrokken zijn uit een *niet*-normale verdeling. Met de functie `qqnorm` vragen we de normale qq-plot van `w` op:

```
> qqnorm(qq.steekproef$w)
```

De output wordt in Fig. 6.2 rechts weergegeven. Op deze grafiek zie je duidelijk dat de 200 punten niet echt dichtbij de diagonaal liggen en dat er ook een patroon zit in de afwijkingen. Dit wijst aan dat dat de steekproef waarschijnlijk niet uit een normale verdeling getrokken werd.

De vier voorbeelden die we net gezien hebben zijn gemakkelijk om visueel te analyseren omdat de steekproef groot is ($n = 200$). Het is dan simpel om te beslissen wanneer de normaliteitsassumptie voldaan is. Als n klein is, dan wordt de visuele analyse moeilijker. Dit wordt geïllustreerd in Fig. 6.3 met 4 normale qq-plots van vier steekproeven ($n = 20$) die allemaal uit een normale verdeling getrokken zijn. Het is hier minder evident om te besluiten. De punten

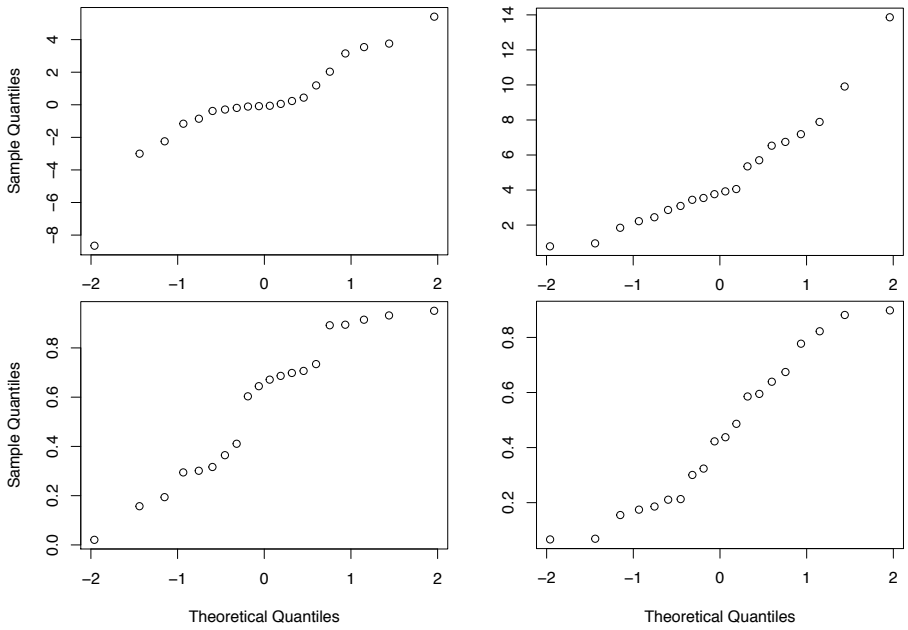


Figuur 6.3: Vier normale qq-plots voor vier steekproeven die uit een normale verdeling getrokken zijn.

liggen niet echt op de diagonaal maar ook niet ver van de diagonaal. Bovendien is er geen patroon te zien in de afwijkingen. We kunnen dus voorzichtig besluiten dat de vier steekproeven uit een normale verdeling getrokken zijn. Dit besluit is misschien fout, maar, als het fout is, dan zijn de steekproeven waarschijnlijk getrokken uit een verdeling die bij benadering normaal verdeeld is. En de p-waarde die we dan zouden berekenen onder de normaliteitsassumptie zou bij benadering correct zijn. Indien ze ver van de significantiedrempel van 5% ligt, dan kunnen we het besluit van de statistische toets vertrouwen (zonder te vergeten dat een statistische toets geen correcte beslissing kan garanderen). Indien de p-waarde dichtbij de significantiedrempel van 5% ligt, dan kunnen we

het besluit van de statistische toets niet vertrouwen en we gebruiken de toets gewoon niet.

In Fig. 6.4 vind je nog 4 normale qq-plots van vier steekproeven ($n = 20$) die allemaal uit een *niet*-normale verdeling getrokken zijn. Bij sommigen (linker kolom) is het verschil t.o.v. Fig. 6.3 duidelijk. Voor de anderen niet.



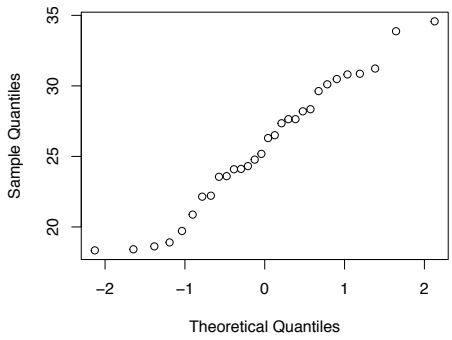
Figuur 6.4: Vier normale qq-plots voor vier steekproeven die uit een *niet*-normale verdeling getrokken zijn.

6.7.2 Toepassing van de normale qq-plot

BMI van FPPW studenten. In Rubr. 6.5.1 hebben we getoetst of FPPW studenten even dik zijn als de doorsnee Vlaming. Te dien einde hebben we een t-toets gebruikt. Om deze toets te mogen gebruiken, moeten we checken of de BMI normaal verdeeld is bij FPPW studenten. We gaan dus de normale qq-plot tekenen voor `bmi`.

```
> qqnorm(bmi)
```

De output van dit commando wordt in Fig. 6.5 weergegeven. De punten liggen niet super ver van de diagonaal en we kunnen dus besluiten dat de verdeling van `bmi` normaal of bijna normaal is. De p-waarde (0.5997) die we toen hadden berekend, was dus min of meer correct. Omdat deze waarde ver van de 0.05 drempel ligt, kunnen we de toets als geldig beschouwen (zonder te vergeten dat een statistische toets geen correcte beslissing kan garanderen).



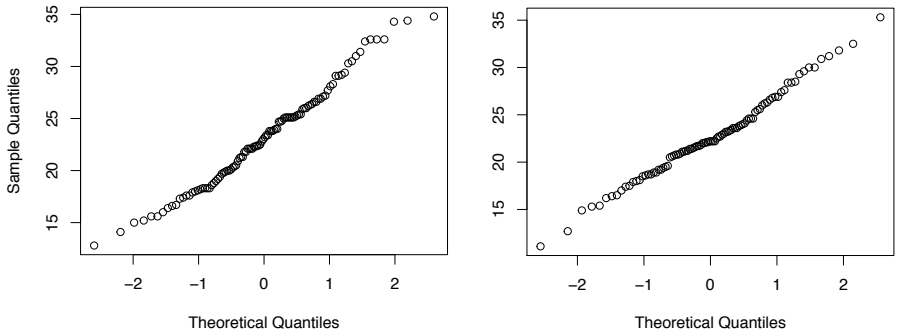
Figuur 6.5: De normale qq-plot voor `bmi`.

Lopen mannen even snel als vrouwen? In Rubr. 6.5.2 hebben we ook getoetst of mannen en vrouwen even snel lopen. Te dien einde hebben we een t-toets voor onafhankelijke steekproeven gebruikt. Om deze toets te mogen gebruiken, moeten we checken of de de variabele `tijd` normaal verdeeld is bij mannen en bij vrouwen. We gaan dus de normale qq-plot tekenen voor `tijdM` bij de mannen en `tijdV` bij de vrouwen.

62. Teken de normale qq-plot van variabele `myData$iq`*, m.b.v. R. Wat besluit je? Is de t-toets op p.96 valide?*

```
> qqnorm(tijdV)
> qqnorm(tijdM)
```

De output van deze commando's wordt in Fig. 6.6 weergegeven. De punten lig-



Figuur 6.6: De normale qq-plot voor `tijdV` (links) en `tijdM` (rechts).

gen niet ver van de diagonaal. Bovendien is de steekproef groot ($n = 200$) en de Centrale Limiet Stelling garandeert dan dat de toetsingsgrootheid bij benadering normaal verdeeld is. We kunnen dus de toets als geldig beschouwen (zonder te vergeten dat een statistische toets geen correcte beslissing kan garanderen).

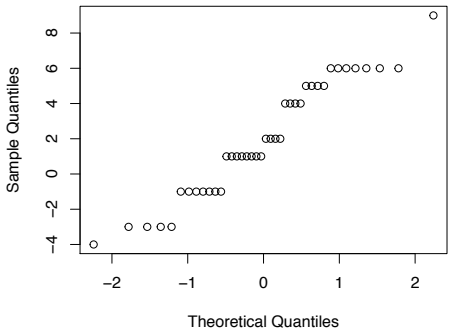
63. Op p. 101 hebben we deze hypothese getoetst: zijn mannelijke FPPW studenten even lang als vrouwelijke FPPW studenten? Was de normaliteitsassumptie in orde?

Rijfouten met een pijnstiller. In Rubr. 6.5.2 hebben we getoetst of autobestuurders meer rijfouten maken indien ze onder invloed van een pijnstiller

rijden. Ten dien einde hebben we een t-toets voor afhankelijke steekproeven gebruikt. Om deze toets te mogen gebruiken, moeten we checken of de de variabele `d` normaal verdeeld is. We gaan dus de normale qq-plot tekenen voor `d` .

```
> qqnorm(d)
```

De output van dit commando wordt in Fig. 6.7 weergegeven. De punten lig-



Figuur 6.7: De normale qq-plot voor `d`.

gen duidelijk niet op de diagonaal en er is een patroon in de afwijkingen. De steekproef werd dus duidelijk niet getrokken uit een normale verdeling. Dit hadden we kunnen voorspellen want het aantal fouten een discrete variabele is. Desalniettemin zijn de afwijkingen niet groot en de steekproef is niet klein ($n = 40$). We kunnen dus beschouwen dat de p-waarde die we berekend hebben (0.000211) min of meer correct is. Bovendien ligt deze p-waarde ver van de significantiedrempel van 5% en we mogen de t-toets als valide beschouwen.

6.8 De significantie

Het resultaat van een statistische toets wordt vaak op volgende manier gerapporteerd. Er is een significant verschil tussen het aantal rijfouten met en zonder pijnstiller. Of, de gemiddelde tijd bij vrouwen (23.2") is niet significant groter dan de gemiddelde tijd bij mannen (22.6"). Dit wordt vaak verkeerd geïnterpreteerd want het gebruik van het woord "significant" is misleidend in deze context.

Laten we het eerste voorbeeld beschouwen: "er is een significant verschil tussen het aantal rijfouten met en zonder pijnstiller". Deze uitdrukking betekent dat het geobserveerde verschil (tussen $\bar{x}_M$ en $\bar{x}_Z$) groot genoeg is om te kunnen besluiten dat het niet toevallig is; om te kunnen besluiten dat de populatieparameters μ_M en μ_Z niet gelijk aan elkaar zijn. Dit betekent geenszins dat het verschil tussen μ_M en μ_Z groot is of dat het verschil tussen μ_M en μ_Z niet verwaarloosbaar is. Het betekent enkel dat het verschil niet nul is, alhoewel

het misschien zeer klein is; en misschien zelfs irrelevant in de context van het onderzoek.

Tweede voorbeeld: "de gemiddelde tijd bij vrouwen (23.2") is niet significant groter dan de gemiddelde tijd bij mannen (22.6")" Dit betekent dat het geobserveerde verschil (tussen $\bar{x}_V$ en $\bar{x}_M$) *niet* groot genoeg is om te kunnen besluiten dat het niet toevallig is; om te kunnen besluiten dat de populatieparameters μ_V en μ_M gelijk zijn. Dit betekent geenszins dat het verschil tussen μ_V en μ_M klein of verwaarloosbaar is. Het betekent dat het verschil tussen μ_V en μ_M nul is; precies nul.

In de gewone taal, als we spreken van iets significant, dan bedoelen we dat het relevant of waardevol is. In de statistiek niet. Om verwarringen te vermijden spreken veel auteurs van statistisch significante verschillen. Ze bedoelen hiermee dat het geobserveerde verschil groot genoeg is om te besluiten (a.d.h.v. en statistische toets) dat de corresponderende populatieparameters verschillend van elkaar zijn (ook als het verschil vewarloosbaar is).

6.9 De fouten

De trekking van een steekproef is een toevalsproces en je weet nooit wat de uitkomst ervan gaat zijn. Je weet dus nooit of een steekproef representatief is. De beslissing die je maakt bij het einde van een toetsingsprocedure kan dus altijd fout zijn. Bij een toets maken we gebruik van de wetten van het kansrekenen om te garanderen dat we *meestal* geen fout maken. Laten we nu meer aandacht besteden aan de fouten en aan het risico op fouten.

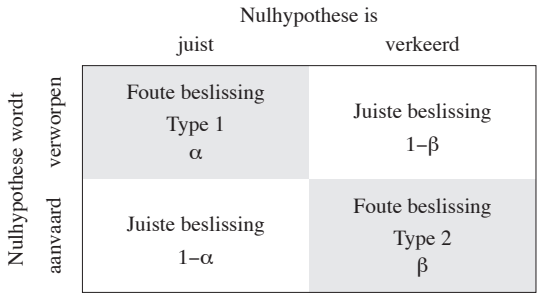
6.9.1 De twee soorten fouten

Stel dat de nulhypothese H_0 juist is en dat je H_0 verwerpt. Dit is fout. Deze fout wordt de *fout van de eerste soort* genoemd. Er moet dus nog een ander soort fout zijn. Ja en zijn naam is ... *fout van de tweede soort*. Wat is deze fout van de tweede soort? Het is de fout die je maakt als je de alternatieve hypothese verwerpt ook indien ze juist is. Het is het ten onrechte verwerpen van de alternatieve hypothese.

De kans op een fout van de eerste soort is de significantie α van de toets. Het is de kans, onder H_0, dat je toevallig een steekproef trekt die leidt tot een verwerping van H_0 alhoewel H_0 juist.

De kans om een fout van de tweede soort te maken is *niet* α. De kans om een fout van de tweede soort te maken wordt aangeduid door het symbool β. De kans om *geen* fout van de tweede soort te maken wordt het *onderscheidingsvermogen* of de *power* genoemd en is gelijk aan $1 - \beta$. Fig 6.8 stelt de vier mogelijke gevallen en hun kansen voor.

Er bestaat natuurlijk een verband tussen α en β. Als we een zeer lage α hanteren (bv. 0.001), dan gaat de p-waarde bijna nooit kleiner dan α zijn en we gaan bijna nooit de nulhypothese verwerpen. We gaan die dus vaak aanvaarden



Figuur 6.8: De vier mogelijke gevallen en hun kansen.

alhoewel ze misschien fout is. Met andere woorden gaan we vaak een fout van de tweede soort maken. De kans β is dus groot.

Integendeel als we een grote α kiezen, is dan het acceptatie-interval smal en we gaan de nulhypothese vaak verwerpen. We gaan dus de alternatieve hypothese vaker aanvaarden dan in het eerste geval. Bijgevolg zal β kleiner zijn.

In het kort, hoe kleiner α, hoe groter β en omgekeerd. In het volgende hoofdstuk gaan we zien hoe we β kunnen berekenen.

64. In Rubr. 6.6 hebben we beslist dat de proportie van alcoholisten in de werkloze populatie niet verschilt van de corresponderende proportie bij de algemene populatie. Deze beslissing was misschien fout. Zo ja, welke soort fout was dat?

6.10 Oplossingen

55) Je vermoedt dat basketbalspelers langer zijn dan de anderen (gemiddeld gezien). Wordt deze hypothese gesteund door het data frame `sportData`?

Oplossing: Laten we twee vectoren aanmaken met de lengtes van de basketbalspelers en van de anderen. Het `R` symbool voor "niet gelijk aan" is "`!=`".

```
> basket <- sportData$lengte[sportData$type == "basketbal"]
> anderen <- sportData$lengte[sportData$type != "basketbal"]
```

Nu berekenen we de gemiddelden in de twee groepen.

```
> mean(basket)
[1] 167.6129
> mean(anderen)
[1] 171.3728
```

De gemiddelde lengte van de basketbalspelers in onze steekproef is kleiner dan de gemiddelde lengte van de anderen. De gegevens steunen onze hypothese niet en we hebben dus geen statistische toets nodig om te besluiten dat basketbalspelers niet langer zijn dan mensen die aan een andere sport doen. Let op, deze conclusie geldt alleen in de doelpopulatie, i.e. de populatie waaruit de steekproef getrokken is (blijkbaar geen populatie van topsporters).

56) Doen mannen even veel aan sport als vrouwen? Toets deze hypothese a.d.h.v. een t-toets, met de gegevens van het data frame `sportData`.

Oplossing: De nulhypothese is dat $\mu_{\texttt{sport,v}} = \mu_{\texttt{sport,m}}$. De alternatieve hypothese is dat $\mu_{\texttt{sport,v}} \neq \mu_{\texttt{sport,m}}$. We maken twee vectoren aan met de gegevens voor de vrouwen en de mannen.

```
> sportV <- sportData$sport[sportData$geslacht == "V"]
> sportM <- sportData$sport[sportData$geslacht == "M"]
```

We berekenen de twee gemiddelden:

```
> mean(sportM)
[1] 2.691489
> mean(sportV)
[1] 3.033019
```

In onze steekproef doen dus mannen minder aan sport dan vrouwen. We berekenen nu de p-waarde.

```
> t.test(x=sportV, y=sportM, alternative = "two.sided")

Welch Two Sample t-test

data:  sportV and sportM
```

```
t = 2.3111, df = 197.63, p-value = 0.02186
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.05010717 0.63295184
sample estimates:
mean of x mean of y
 3.033019  2.691489
```

De p-waarde is 2% en we besluiten dan mannen niet even veel aan sport doen als vrouwen.

57) Ga na of de verwachting van het IQ dezelfde is bij mannelijke en vrouwelijke FPPW studenten, a.d.h.v. `myData`.

Oplossing: De steekproeven zijn te klein om de Welch t-toets te mogen gebruiken.

58) Bereken het gemiddelde $\overline{d}$ van de vector `d`, m.b.v. R. Kan je het resultaat bekomen d.m.v. een eenvoudige redenering?

Oplossing:

```
> mean(d)
[1] 2
```

Het gemiddelde verschil is 2. Dit is gelijk aan $30 - 28$.

59) Het data frame `FB` bevat gegevens m.b.t. 172 Facebook gebruikers: geslacht, aantal vrienden, aantal likes in 2016 en 2017. Ga na of vrouwen in het algemeen meer vrienden hebben dan mannen.

Oplossing: De nulhypothese is dat $\mu_{\texttt{friends,v}} = \mu_{\texttt{friends,m}}$. De alternatieve hypothese is dat $\mu_{\texttt{friends,v}} > \mu_{\texttt{friends,m}}$. We maken twee vectoren aan met de gegevens voor de vrouwen en de mannen.

```
> fV <- FB$friends[FB$geslacht == "V"]
> fM <- FB$friends[FB$geslacht == "M"]
```

We berekenen de twee gemiddelden:

```
> mean(fV)
[1] 133.3152
> mean(fM)
[1] 123.45
```

De gegevens steunen de alternatieve hypothese. We berekenen nu de p-waarde. Let op, de twee steekproeven zijn onafhankelijk en we gebruiken dus het argument "`paired = TRUE`" niet.

```
> t.test(x=fV, y=fM, alternative = "greater")

Welch Two Sample t-test

data:  fV and fM
t = 1.683, df = 142.66, p-value = 0.04728
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1607387       Inf
sample estimates:
mean of x mean of y
 133.3152  123.4500
```

De p-waarde is 4.7% en we besluiten dan vrouwen meer Facebook vrienden hebben dan mannen.

60) Ga na of Facebook gebruikers in het algemeen meer likes hebben gedaan in 2017 dan in 2016.

Oplossing: De nulhypothese is dat $\mu_{\texttt{like2016}} = \mu_{\texttt{like2017}}$. De alternatieve hypothese is dat $\mu_{\texttt{like2016}} < \mu_{\texttt{like2017}}$. We berekenen de twee gemiddelden:

```
> mean(FB$like2016)
[1] 297.8023
> mean(FB$like2017)
[1] 299.343
```

De gegevens steunen de alternatieve hypothese. We berekenen nu de p-waarde. Let op, de twee steekproeven zijn afhankelijk (twee scores per gebruiker) en we gebruiken nu wel het argument "`paired = TRUE`".

```
> t.test(x=FB$like2016, y=FB$like2017, alternative = "less", paired = TRUE)

Paired t-test

data:  FB$like2016 and FB$like2017
t = -0.13167, df = 171, p-value = 0.4477
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 17.81134
sample estimates:
mean of the differences
              -1.540698
```

De p-waarde is 45% en we besluiten dan het aantal likes niet gestegen is.

61) Twintig procent van de in de laatste 10 jaar afgestudeerde ingenieurs zijn vrouwen. In een bedrijf werden 40 ingenieurs aangeworven in de laatste 10 jaar. Slechts 2 van die ingenieurs zijn vrouwen (5%). Dit geeft de indruk

dat deze bedrijf genderdiscriminatie doet. Toets de hypothese dat de bedrijf discrimineert tussen mannen en vrouwen. Eerst met een rekenmachine, dan met R.

Oplossing: Proportie van aangeworven vrouwen $= \pi$. $H_0 : \pi = 0.2$. $H_a : \pi < 0.2$. Aantal aangeworven vrouwen $\sim B(40, 0.2)$. Overschrijdingskans:

$$\begin{aligned} P[B(40,0.2) \leq 2] &= P[B(40,0.2) = 2] + P[B(40,0.2) = 1] + P[B(40,0.2) = 0] \\ &= \frac{40!}{2!\,38!} \times 0.2^2 \times 0.8^{38} + \frac{40!}{1!\,39!} \times 0.2^1 \times 0.8^{39} + \frac{40!}{0!\,40!} \times 0.2^0 \times 0.8^{40} \\ &= 780 \times 0.040 \times 0.00021 + 40 \times 0.200 \times 0.00017 + 1 \times 1 \times 0.00013 \\ &= 0.0065 + 0.0014 + 0.00013 = 0.008. \end{aligned}$$

De overschrijdingskans is veel kleiner dan 0.05 en we verwerpen de nulhypothese.

Nu met R:

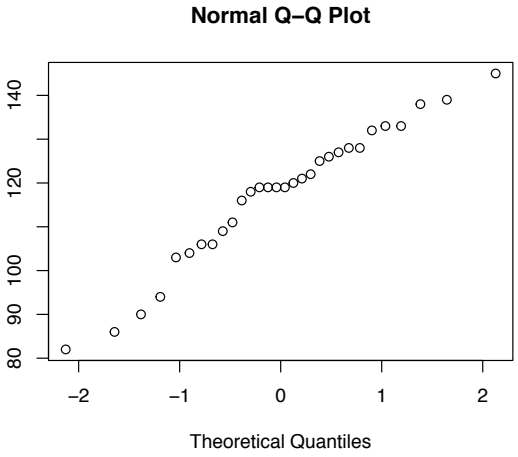
```
> binom.test(x=2, n=40, p=0.2, alternative = "less")

Exact binomial test

data:  2 and 40
number of successes = 2, number of trials = 40, p-value = 0.007942
alternative hypothesis: true probability of success is less than 0.2
95 percent confidence interval:
 0.000000 0.149152
sample estimates:
probability of success
                  0.05
```

62) Teken de normale qq-plot van variabele `myData$iq`, m.b.v. `R`. Wat besluit je? Is de t-toets op p.96 valide?

Oplossing:

```
> qqnorm(myData$iq)
```

Output:



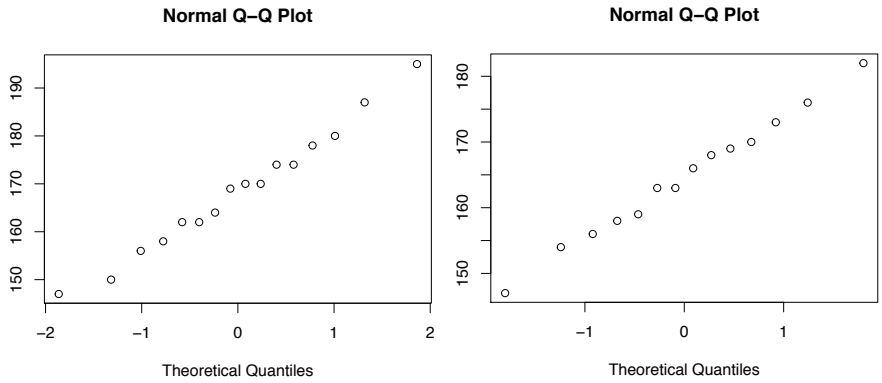
De punten liggen min of meer op de diagonaal. De normaliteitsassumptie is in orde.

63) Op p. 101 hebben we deze hypothese getoetst: zijn mannelijke FPPW studenten even lang als vrouwelijke FPPW studenten. Was de normaliteitsassumptie in orde?

Oplossing:

```
> qqnorm(lengteV)
> qqnorm(lengteM)
```

Output:



De normaliteitsassumptie is in orde bij de vrouwen (links) en de mannen (rechts).

64) In Rubr. 6.6 hebben we beslist dat de proportie van alcoholisten in de werkloze populatie niet verschilt van de corresponderende proportie bij de algemene populatie. Deze beslissing was misschien fout. Zo ja, welke soort fout was dat?

Oplossing: We hebben de nulhypothese aanvaard (onderste rij van Fig. 6.8). Als dit een foutieve beslissing was, dan was de nulhypothese verkeerd (rechterkolom van Fig. 6.8). We hebben dus een fout van de tweede soort gemaakt.

Hoofdstuk 7

De power

Bij het vorige hoofdstuk hebben we gezien dat er twee soorten fouten zijn bij het uitvoeren van een statistische toets.

- Je maakt een fout van de eerste soort als je de nulhypothese verwerpt terwijl ze correct is. De kans op deze fout (indien H_0 correct is) is α. Je kiest zelf deze kans vooraleer je de toets uitvoert.
- Je maakt een fout van de tweede soort als je de alternatieve hypothese verwerpt terwijl ze correct is. De kans op deze fout (indien H_a correct is) is β. De kans om deze fout niet te maken (de kans om H_a te aanvaarden indien H_a correct is) is $1 - \beta$ en wordt het onderscheidingsvermogen of power genoemd. In het volgende paragraaf zien we aan de hand van een voorbeeld hoe deze kans berekend kan worden.

7.1 De power bij het toetsen van een hypothese betreffende een proportie

In paragraaf 6.6 hebben we gezien dat de proportie van alcoholisten in de populatie tussen 30 en 40 jaar gelijk is aan 8%. Volgens een theorie is deze proportie groter bij werklozen. We hebben getoetst of de proportie van alcoholisten bij werklozen tussen 30 en 40 jaar groter is dan 8%. De nulhypothese was 'H_0 : $\pi = 0.08$' en de alternatieve hypothese was 'H_a: $\pi > 0.08$'. Ten dien einde hebben we een steekproef van 10 werklozen getrokken waarin er 1 alcoholist was. De geobserveerde proportie van alcoholisten was $p = 1/10 = 0.1$ en was dus groter dan de proportie van de nulhypothese, i.e., 0.08. Dit steunde de alternatieve hypothese, maar we hebben gezien dat de geobserveerde proportie (0.1) niet genoeg afweek van 0.08 om de nulhypothese te verwerpen. Inderdaad, de p-waarde was $P(B(10, 0.08) \geq 1) = 0.57$.

We zijn zeker dat we geen fout van de eerste soort hebben gemaakt (de nulhypothese ten onrechte verwerpen), maar we hebben misschien een fout van de tweede soort gemaakt. We zullen het natuurlijk nooit weten, maar we kunnen

ons afvragen of de procedure die we gebruikt hebben in orde was qua fout van de tweede soort. Laten we dus de power van de toets berekenen. De power is de kans dat we de nulhypothese verwerpen indien ze fout is, dus indien π groter dan 0.08 is. Maar hoe groter? 0.081, 0.1, 0.5, ...? Het is intuïtief duidelijk dat de kans om de nulhypothese te verwerpen zeer klein is indien $\pi = 0.081$. Het verschil tussen 0.081 en 0.08 is zodanig klein dat het bijna ondetecteerbaar is. Het is ook intuïtief duidelijk dat de kans om de nulhypothese te verwerpen bijna 100% is indien $\pi = 0.5$. Om de power te berekenen gaan we dus een specifieke waarde van π veronderstellen en niet zomaar $\pi > 0.08$. We stellen dus een specifieke alternatieve hypothese op. Bv. $H_a : \pi = 0.15$.

We kunnen dan de R functie `powerBinom` gebruiken om de power te berekenen. Deze functie is niet beschikbaar bij de standaard installatie van R. Om die functie te kunnen gebruiken, moet je eerst het package `exactci` installeren en opladen. Een package is een bundel van functies die niet standaard geïmplementeerd zijn in R en die wel geïnstalleerd en opgeladen kunnen worden. Om het package `exactci` te installeren (van een Internet server downloaden), moet je het tabblad "Packages" selecteren in het deelvenster rechtsonder. Dan klik je op "Install", een venster verschijnt, je vult de naam van het package in (exactci) en je klikt op de knop "Install". Dit geeft een aantal meldingen, linksonder, in het rood, die je mag negeren. Vanaf dit moment is het package opgeslaan op je harde schijf, maar het is nog niet klaar voor gebruik. Je moet het package opladen in het geheugen van `R`. Onder het tabblad "Packages" zie je een lijst van alle geïnstalleerde packages. Om één van die packages op te laden moet je gewoon het vakje naast het gewenste package aanvinken. Nu ben je klaar om de functies van het package `exactci` te gebruiken. Op het moment dat je het programma `R` zal stoppen, zal het package gewist worden van het geheugen. Dus, de volgende keer dat je `R` zal opstarten zal je opnieuw het package `exactci` moeten opladen door het juist vakje aan te vinken, maar je zal het package niet meer moeten installeren.[1]

De functie `powerBinom` heeft vijf argumenten nodig: de steekproefgrootte `n`, de proportie onder de nulhypothese `p0`, de proportie onder de alternatieve hypothese `p1`, de significantie[2] `sig.level` $= \alpha$ en de soort hypothese (`alternative = "two.sided"` of `"one.sided"`). We zijn nu klaar om de functie `powerBinom` te gebruiken. Laten we de power van de binomiale toets berekenen, bij ons alcoholisten-voorbeeld, met $n = 10$, $\alpha = 0.05$ en onder de specifieke alternatieve hypothese $\pi = 0.15$.

```
> powerBinom(n = 10, p0 = 0.08, p1 = 0.15, sig.level = 0.05,
   alternative  = "one.sided")

     power and sample size for single binomial response
```

[1] Als je `R` op Athena gebruikt dan hoef je geen package te installeren. Alle packages zijn geïnstalleerd op Athena. Maar je moet wel de package `exactci` opladen.

[2] Voor een onbekende reden gebruikt de functie `powerBinom` het argument `sig.level` in plaats van `conf.level`. Hier moet je dus 0.05 aangeven i.p.v. 0.95.

```
           n = 10
          p0 = 0.08
          p1 = 0.15
       power = 0.1798035
 alternative = one.sided
   sig.level = 0.05

NOTE: use rejections in correct direction only
```

We komen een power van 18% uit. Dit is natuurlijk veel te klein: de kans om de juiste beslissing te nemen, indien H_a juist is, is slechts 18%. Dit is niet acceptabel.

De power en het significantieniveau We hebben al gezien dat α en β negatief gecorreleerd zijn. We gaan dit illustreren aan de hand van ons voorbeeld. Stel dat we een significantieniveau van 20% gebruiken, i.p.v. 5%. Wat is dan de power? We gebruiken nog dezelfde functie, maar met het argument `sig.level = 0.20`.[3]

```
> powerBinom(n = 10, p0 = 0.08, p1 = 0.15, sig.level = 0.20,
  alternative  = "one.sided")

     power and sample size for single binomial response

           n = 10
          p0 = 0.08
          p1 = 0.15
       power = 0.4557002
 alternative = one.sided
   sig.level = 0.2

NOTE: use rejections in correct direction only
```

We komen een power van 46% uit. Dit is beter, maar nog niet acceptabel. Merk op dat de power hoger zou zijn indien je zou beslissen in functie van de uitkomst van een muntstuk: de power zou 50% zijn. Onze power van 46% is dus duidelijk veel te laag.

De power en de specifieke alternatieve hypothese We hebben al gezien dat de power waarschijnlijk afhankelijk is van de echte waarde van π. We kennen deze waarde niet en, bij vorige paragraaf, hebben we verondersteld dat de echte waarde van π gelijk aan 0.15 is. Dit was onze specifieke alternatieve hypothese. Laten we de impact van de specifieke alternatieve hypothese nagaan.

[3] Eigenlijk is het argument `sig.level` facultatief. Je hoeft het te gebruiken indien je wenst af te wijken van de gebruikelijke 0.05. Als je het niet vermeldt dan gaat R er van uit dat je een significantie gelijk aan 0.05 wenst te hanteren. Men zegt dat 0.05 de 'default value' van het argument `sig.level` is. In het vervolg laten we dit argument meestal weg.

We berekenen opnieuw de power, maar onder een andere specifieke alternatieve hypothese: $\pi = 0.081$.

```
> powerBinom(n= 10, p0= 0.08, p1= 0.081, alternative= "one.sided")

     power and sample size for single binomial response

              n = 10
             p0 = 0.08
             p1 = 0.081
          power = 0.04137187
    alternative = one.sided
      sig.level = 0.05

NOTE: use rejections in correct direction only
```

De power daalt tot 4%. Het is nog lager dan vroeger. Dit is eigenlijk geen verrassing. Indien de echte proportie (8.1%) zo dicht bij de proportie (8%) van de nulhypothese ligt, is het zeer onwaarschijnlijk dat we dit kunnen detecteren.

Laten we nog ééns de power berekenen, maar onder de specifieke alternatieve hypothese: $\pi = 0.35$.

```
> powerBinom(n= 10, p0= 0.08, p1= 0.35, alternative = "one.sided")

     power and sample size for single binomial response

              n = 10
             p0 = 0.08
             p1 = 0.35
          power = 0.7383926
    alternative = one.sided
      sig.level = 0.05

NOTE: use rejections in correct direction only
```

De power stijgt tot 74%. Het is veel hoger dan vroeger. Dit is ook geen verrassing. Indien de echte proportie (35%) zo sterk afwijkt van de proportie (8%) van de nulhypothese ligt, is het zeer waarschijnlijk dat we dit kunnen detecteren.

De power en de steekproefgrootte We kunnen verwachten dat hoe groter de steekproef, hoe kleiner de kans om een fout te maken. Laten we dit verifiëren. We berekenen opnieuw de power van de binomiale toets, bij ons alcoholistenvoorbeeld, met $\alpha = 0.05$ en onder de specifieke alternatieve hypothese $\pi = 0.15$. Deze keer gebruiken we een grotere steekproef: $n = 150$.

```
> powerBinom(n= 150, p0= 0.08, p1= 0.15, alternative= "one.sided")
```

```
     power and sample size for single binomial response

              n = 150
             p0 = 0.08
             p1 = 0.15
          power = 0.8188063
    alternative = one.sided
      sig.level = 0.05

NOTE: use rejections in correct direction only
```

We komen een power van 82% uit. De power is inderdaad gestegen omwille van de grotere steekproefgrootte.

65. Bereken de power van de binomiale toets, bij het alcoholisten-voorbeeld, met $n = 40$, $\alpha = 0.05$ en onder de specifieke alternatieve hypothese $\pi = 0.15$.

Stel nu dat we de minimale steekproefgrootte willen bepalen om een power van 90% te garanderen (met $\alpha = 0.05$ en onder de specifieke alternatieve hypothese $\pi = 0.15$). Om dit te berekenen, gebruiken we nogeens de functie `powerBinom`, maar we laten het argument `n` vallen en we gebruiken het extra argument `power`.

```
> powerBinom(power= 0.90, p0= 0.08, p1= 0.15, alternative= "one.sided")

     power and sample size for single binomial response

              n = 177
             p0 = 0.08
             p1 = 0.15
          power = 0.9017898
    alternative = one.sided
      sig.level = 0.05

NOTE: use rejections in correct direction only
```

We vinden dat de minimale steekproefgrootte om een power van 90% te garanderen gelijk is aan 177.

66. Bereken de minimale steekproefgrootte bij het alcoholisten-voorbeeld om een power van 95% te garanderen (met $\alpha = 0.05$ en onder de specifieke alternatieve hypothese $\pi = 0.20$).

7.2 De power bij het toetsen van een hypothese betreffende een verwachting

In Rubr. 6.5.1.2 hebben we de BMI van FPPW studenten geanalyseerd en we hebben besloten dat de verwachting van de BMI van FPPW studenten niet verschillend is van de BMI van de doorsnee Vlaming (25.3). We zijn opnieuw zeker dat we geen fout van de eerste soort hebben gemaakt (de nulhypothese ten onrechte verwerpen), maar we hebben misschien een fout van de tweede soort gemaakt. We zullen het natuurlijk nooit weten, maar we kunnen ons afvragen of de procedure die we gebruikten in orde was qua fout van de tweede soort. Laten we dus de power van de toets berekenen. De power is de kans dat we de

nulhypothese verwerpen indien ze fout is, dus indien μ_X verschillend van 25.3 is. Maar hoe verschillend? 25.4, 26, 30, ...? Het is intuïtief duidelijk dat de kans om de nulhypothese te verwerpen zeer klein is indien $\mu_X = 25.4$. Het verschil tussen 25.4 en 25.3 is zodanig klein dat het bijna ondetecteerbaar is. Het is ook intuïtief duidelijk dat de kans om de nulhypothese te verwerpen bijna 100% is indien $\mu_X = 30$.

Om de power te berekenen moeten we een specifieke alternatieve hypothese opstellen. Welk verschil (in absolute waarde) qua BMI willen we kunnen detecteren met onze toets, in de context van dit onderzoek? Een verschil van 0.1? Dit correspondeert ongeveer met een verschil van 300g. Dit is belachelijk klein en niet relevant. Willen we een BMI-verschil van 1 (ongeveer 3kg) detecteren? Misschien wel. Een BMI-verschil van 2 (ongeveer 6kg)? Ja, zeker! Vanaf 2 of 3 kg lijkt het relevant, vanuit een volksgezondheidsperspectief, te weten dat er een verschil is. En men kan dan eventueel nagaan wat de oorzaken van het verschil zijn (dit is geen statistiek meer), enz. Onder 1 kg is het verschil gewoon een detail.[4] We gaan dus de power berekenen onder de specifieke alternatieve hypothese dat de absolute waarde van het BMI-verschil gelijk aan 1 is. Dit correspondeert met $\mu_X = 26.3$ of $\mu_X = 24.3$.

Om de berekeningen uit te voeren gebruiken we de functie `power.t.test`. Deze functie heeft zes argumenten nodig: de steekproefgrootte `n` $= n$, de schatting van de standaarddeviatie `sd` $= s_X$, de significantie `sig.level` $= \alpha$[5], het aantal steekproeven (`type = "one.sample"`), de soort hypothese (`alternative = "two.sided"`) en `delta`: de absolute waarde van het verschil tussen de verwachtingen onder H_0 en H_a. Het lijkt hier redelijk om 1 te hanteren voor `delta`. We zijn nu klaar om de functie `power.t.test` te gebruiken:

```
> bmi <- myData$gewicht / (myData$lengte/100)^2
> sd(bmi)
[1] 4.60793
> power.t.test(n= 30, delta= 1, sd= 4.61, alternative = "two.sided",
sig.level = 0.05, type = "one.sample")

     One-sample t test power calculation

              n = 30
          delta = 1
             sd = 4.61
      sig.level = 0.05
          power = 0.208701
    alternative = two.sided
```

[4] Merk op dat dit afhankelijk van de context is. In een onderzoek rond het gewicht van vroeggeborenen is het zeer belangrijk een verschil van 100 g te kunnen detecteren tussen baby's die behandeling A of B krijgen.

[5] Zoals `powerBinom` gebruikt de functie `power.t.test` het argument `sig.level` in plaats van `conf.level`. Hier moet je dus 0.05 aangeven i.p.v. 0.95. Dit argument is facultatief, zoals bij de functie `powerBinom`. Zijn 'default value' is 0.05.

De power is 21%. M.a.w. indien we deze toets gebruiken in dezelfde context (met dezelfde steekproefgrootte, dezelfde significantie, enz.) met veel steekproeven en indien het verschil tussen de verwachtingen onder H_0 en H_a 1 bedraagt (in absolute waarde), dan zullen we het verschil detecteren (t.t.z. de nulhypothese verwerpen) in 21% van de gevallen. Dit is slechter dan wat we zouden uitkomen met een muntstuk! Dit is niet verwonderend: met een kleine steekproef ($n = 30$) mag je niet verwachten dat de power hoog is (behalve als je een groot verschil wenst te detecteren).

De power en het verschil tussen de verwachtingen Stel dat het echte verschil 2 zou zijn (ongeveer 6kg). Wat zou dan de power zijn?

```
> power.t.test(n= 30, delta= 2, sd= 4.61, alternative= "two.sided",
type = "one.sample")

     One-sample t test power calculation

              n = 30
          delta = 2
             sd = 4.61
      sig.level = 0.05
          power = 0.6319688
    alternative = two.sided
```

De power zou 63% zijn. Dit is nog te laag.

De power en de steekproefgrootte. Stel dat we een grotere steekproef trekken. bv. 100. Wat zou de power zijn?

```
> power.t.test(n= 100, delta= 1, sd= 4.61, alternative= "two.sided",
type = "one.sample")

     One-sample t test power calculation

              n = 100
          delta = 1
             sd = 4.61
      sig.level = 0.05
          power = 0.5746241
    alternative = two.sided
```

De power blijft laag. Welke steekproegrootte zou een deftige power garanderen? We kunnen dit berekenen met dezelfde R functie `power.t.test`. We laten het argument `n` weg en we voegen het argument `power` toe.

```
> power.t.test(power = 0.90, delta = 1, sd = 4.61, sig.level = 0.05,
  alternative  = "two.sided", type = "one.sample")
```

```
     One-sample t test power calculation

              n = 225.2343
          delta = 1
             sd = 4.61
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

Dus om een power van 90% te garanderen moet onze steekproefgrootte minstens 226 zijn.

67. In Rubr. 6.5.1, p. 94, hebben we een t-toets gebruikt om de verwachting van het IQ van FPPW studenten te vergelijken met het IQ van de doorsnee Vlaming. Bereken de power van die toets, onder de specifieke alternatieve hypothese: $\mu_X = 110$.

7.3 De power bij het toetsen van een hypothese betreffende twee verwachtingen—afhankelijke steekproeven

In Rubr. 6.5.2, vanaf p. 101, hebben we het voorbeeld van de rijfouten behandeld. Het gemiddelde aantal fouten in de groep met de pijnstiller is 30. Het is 28 in de groep zonder pijnstiller. Het verschil is 2. Ten opzichte van 28 is dit niet veel en de conclusie van de t-toets is toch dat de nulhypothese verworpen moet. Dit betekent dat de power van de toets vrij hoog moet zijn. Laten we dit berekenen. We moeten eerst een specifieke alternatieve hypothese bepalen; niet zomaar $\mu_M - \mu_Z > 0$ maar $\mu_M - \mu_Z = 1$ of 2 of 5 . . . Een verschil $\mu_M - \mu_Z$ gelijk aan 5 (t.o.v. 28) is iets dat we bijna zeker willen detecteren. Indien autobestuurders met pijnstillers zo slecht aan het stuur zijn, dan is het belangrijk om het te weten. Een verschil $\mu_M - \mu_Z$ gelijk aan 1 (t.o.v. 28) is niet zo relevant. Vanaf twee of drie wordt het relevant. Laten we dus de power berekenen voor een verschil van 3. We gebruiken het argument `type = "paired"`. We schatten eerst de standaarddeviatie van het verschil:

```
> sd <- sd(rijfoutenData$rijfoutenMet-rijfoutenData$rijfoutenZonder)
> sd
[1] 3.281651
```

Nu berekenen we de power.

```
> power.t.test(n = 40, delta = 3, sd = sd, sig.level = 0.05,
alternative  = "one.sided", type = "paired")

     Paired t test power calculation

              n = 40
          delta = 3
             sd = 3.281651
      sig.level = 0.05
```

```
          power = 0.9999725
    alternative = one.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
within pairs
```

Dit bevestigt onze intuïtie: de power van deze toets is zeer hoog. We hadden dus het experiment kunnen doen met een kleinere steekproef. Hoe groot moet de steekproef zijn om een power van 90% te garanderen, met een verschil gelijk aan 3?

```
> power.t.test(power = 0.90, delta = 3, sd = sd, sig.level = 0.05,
alternative  = "one.sided", type = "paired")

     Paired t test power calculation

              n = 11.74024
          delta = 3
             sd = 3.281651
      sig.level = 0.05
          power = 0.9
    alternative = one.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
within pairs
```

Een veel kleiner steekproef ($n = 12$) was genoeg.

7.4 De power bij het toetsen van een hypothese betreffende twee verwachtingen—onafhankelijke steekproeven

68. Bij oef. 60 heb je getoetst of Facebook gebruikers in het algemeen meer likes hebben gedaan in 2017 dan in 2016. Stel een relevante specifieke alternatieve hypothese en bereken de minimale steekproegrootte om een power van 90% te garanderen.

De procedure om een kind te adopteren duurt in veel landen zeer lang. Dit is ook het geval in New York city. Een groep onderzoekers [Festinger and Pratt, 2002] is nagegaan of een bepaalde aanpassing van de in N.Y. gebruikte procedure korter zou zijn. In 1998 en 1999 hebben ze een steekproef van 175 adoptieaanvragen getrokken. Uit die gevallen werden 119 aanvragen at random geselecteerd en hebben een gewijzigde procedure gevolgd (met toelating van de lokale overheid). De overige 56 aanvragen volgden de gewone procedure. De onderzoekers hebben een aantal variabelen gemeten maar we gaan hier slechts op de lengte van één van de stappen van de procedure focussen: van *adoptive home placement* tot *petition to free filed*. Wat die termen precies betekenen is voor ons niet zeer belangrijk. De duur van die stap in de controlegroep (gewone procedure) wordt door het symbool X_1 aangeduid; in de experimentele groep (aangepaste procedure), door X_2. De alternatieve hypothese luidt als "$\mu_1 > \mu_2$" en de nulhypothese als "$\mu_1 = \mu_2$".

Je vindt de gegevens in het data frame `adoptieData`.

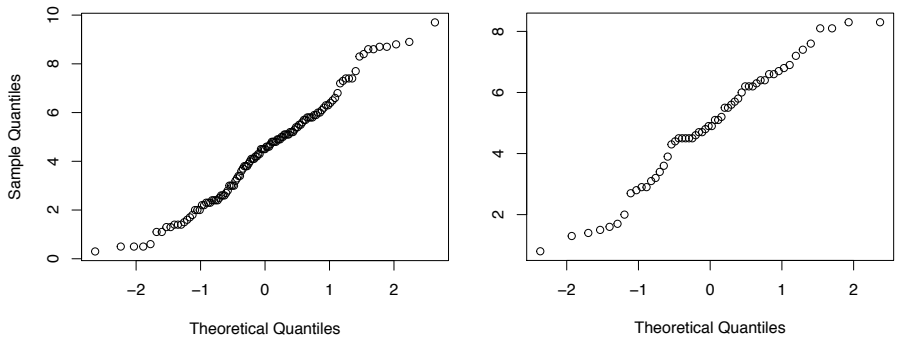
```
> adoptieData
    duur     conditie
1    7.4      control
2    2.9      control
3    2.0      control
...   ...    ...
173  2.5 experimental
174  2.0 experimental
175  4.8 experimental
```

We maken twee vectoren aan met de duur in beide groepen en we berekenen de gemiddelden:

```
> exp <- adoptieData$duur[adoptieData$conditie == "experimental"]
> con <- adoptieData$duur[adoptieData$conditie == "control"]
> mean(exp)
[1] 4.40084
> mean(con)
[1] 4.9
```

Op het eerste zicht wijst dit aan dat de aangepaste procedure inderdaad sneller is. De onderzoekers willen nagaan of dit misschien het effect van het toeval is. Ze voeren een t-toets uit. Ze gaan dus eerst na of de toevalsvariabele `duur` normaalverdeeld is in beide condities.

```
> qqnorm(exp)
> qqnorm(con)
```



Figuur 7.1: De normale qq-plot voor de duur in de experimentele (links) en controle groep (rechts).

Beide plots (Fig. 7.1) tonen geen grote afwijking t.o.v. de diagonaal. We besluiten dus dat beide steekproeven getrokken zijn uit een normale verdling. Nu de t-toets:

```
> t.test(x = exp, y = con, alternative = "less")

Welch Two Sample t-test

data:  exp and con
t = -1.5253, df = 119.48, p-value = 0.06491
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.0433243
sample estimates:
mean of x mean of y
  4.40084   4.90000
```

De p-waarde is opnieuw groter dan 5% en we aanvaarden derhalve de nulhypothese, t.t.z. de twee verwachtingen verschillen niet van elkaar. M.a.w. de aangepaste procedure is niet sneller.

Nu kunnen we ons de vraag stellen: stel dat het verschil tussen beide procedures één jaar zou zijn (gemiddeld gezien); wat zou de kans zijn om de nulhypothese te verwerpen? M.a.w., wat is de power of het onderscheidingsvermogen van de toets? We kunnen het berekenen m.b.v. de R functie `pwr.t2n.test`.[6] Deze functie is in principe bedoeld om de power te berekenen bij een t-toets met twee onafhankelijke steekproeven onder de assumptie dat de varianties identiek zijn in beide populaties. In het adoptie-onderzoek weten we niet of deze assumptie voldaan is, maar we gaan toch de functie `pwr.t2n.test` gebruiken. De bekomen power zal niet correct zijn maar kan bij benadering gebruikt worden.

We bespreken één argument van deze functie, met name `d`[7]: dit is de schatting van de effectgrootte[8] (Engels: effect size), i.e.

$$\frac{\mu_1 - \mu_2}{s_{\text{pooled}}},$$

waarbij s_{pooled} de schatting is van de gemeenschappelijke standaarddeviatie, i.e. de vierkantswortel uit

$$s^2_{\text{pooled}} = \frac{\text{SS}_X + \text{SS}_Y}{n_1 + n_2 - 2}.$$

Let op, het argument `delta` van de functie `power.t.test` is de absolute waarde van het verschil $\mu_1 - \mu_2$ terwijl het argument `d` van de functie `pwr.t2n.test`

[6] Je moet eerst de package `pwr` installeren en opladen. Te dien einde volg je de procedure die op p.119 uitgelegd is. Als je `R` op Athena gebruikt dan is de package `pwr` al geïnstallerd. Je hoeft hem wel op te laden.

[7] Niet verwarren met d van de t-toets voor twee afhankelijke steekproeven (p. 101).

[8] Er is een traditie in de humane wetenschappen dat waarden van d rond 0.2 (in absolute waarde) met een klein effect overeenkomen. Waarden rond 0.5 (in absolute waarde) corresponderen met een medium effect en waarden boven 0.8 (in absolute waarde) duiden een groot effect aan. Dit is arbitrair en niet gegrond. Die waarden werden door [Cohen, 1988] voorgesteld, maar hij schreef zelf: "This is an operation fraught with many dangers: The definitions are arbitrary, such qualitative concepts as 'large' are sometimes understood as absolute, sometimes as relative; and thus they run a risk of being misunderstood."

gebruik maakt van het verschil $\mu_1 - \mu_2$, zonder absolute waarde. Het argument `d` kan dus positief of negatief zijn terwijl `delta` altijd positief is. In het voorbeeld van de adoptieprocedure gebruiken we $(\mu_1 - \mu_2) = 1$ en s_{pooled} gelijk aan

```
> sqrt(( 55*var(con) + 118*var(exp))/(173))
[1] 2.102448
```

We zijn nu klaar om de power te berekenen:

```
> pwr.t2n.test( n1 = length(con), n2 = length(exp), d = 1/2.1,
sig.level = 0.05, alternative = "greater")

     t test power calculation

             n1 = 56
             n2 = 119
              d = 0.4761905
      sig.level = 0.05
          power = 0.9001047
    alternative = greater
```

De power is 90%. Dus, indien het verschil tussen beide procedures 1 jaar bedraagt, en indien we veel steekproeven trekken, dan gaan we het verschil detecteren in 90% van de gevallen. Dit is niet slecht.

7.4.0.1 De power en het significantieniveau

We illustreren nog eens het verband tussen α en β. Stel dat we een significantieniveau van 1% gebruiken, i.p.v. 5%. Wat is dan de power?

```
> pwr.t2n.test( n1 = length(con), n2 = length(exp), d = 1/2.1,
sig.level = 0.01, alternative = "greater")

     t test power calculation

             n1 = 56
             n2 = 119
              d = 0.4761905
      sig.level = 0.01
          power = 0.7221235
    alternative = greater
```

De power daalt tot 72%, wat niet meer acceptabel is.

69. Bij Oef. 57 heb je een t-toets gebruikt om na te gaan of mannelijke en vrouwelijke FPPW studenten van elkaar verschillen qua IQ. Bereken de power van de toets onder de specifieke alternatieve hypothese dat het verschil tussen mannen en vrouwen 5 bedraagt.

7.4.0.2 De power en de steekproefgrootte

Stel dat de twee steekproeven kleiner zijn. bv. $n_1 = 50 = n_2$. Wat zou de power dan zijn?

```
> pwr.t2n.test( n1 = 50, n2 = 50, d = 1/2.1, sig.level = 0.05,
alternative = "greater")

     t test power calculation

             n1 = 50
             n2 = 50
              d = 0.4761905
      sig.level = 0.05
          power = 0.7641174
    alternative = greater
```

De power daalt tot 76%, wat niet echt goed is.

Stel nu dat we een power van 95% willen. Hoe groot moet de steekproef zijn? We kunnen dit gemakkelijk berekenen met dezelfde functie `pwr.t2n.test`. We laten het argument `n1` of `n2` weg en we voegen het argument `power` toe:

```
> pwr.t2n.test( power = 0.95, n1 = 56, d = 1/2.1, sig.level = 0.05,
alternative = "greater")

     t test power calculation

             n1 = 56
             n2 = 330.885
              d = 0.4761905
      sig.level = 0.05
          power = 0.95
    alternative = greater
```

We hebben een power van 95% gevraagd en we hebben `n2` weggelaten. Het softwarepakket R heeft dus de minimale waarde van `n2` berekend om een power van 95% te garanderen: het is 331.

Stel dat dit onmogelijk is omdat de overheid in New York weigert om zoveel dossiers met de experimentele procedure te behandelen. We mogen slechts 119 dossiers behandelen met de versnelde procedure. Hoeveel dossiers moeten we hebben in de controle groep om een power van 95% te garanderen?

```
> pwr.t2n.test( power = 0.95, n2 = 119, d = 1/2.1, sig.level = 0.05,
alternative = "greater")

     t test power calculation

             n1 = 80.60582
             n2 = 119
              d = 0.4761905
      sig.level = 0.05
          power = 0.95
    alternative = greater
```

We hebben een power van 95% gevraagd en we hebben `n1` weggelaten. Het softwarepakket R heeft dus de minimale waarde van `n1` berekend om een power van 95% te garanderen: het is 81.

Indien we twee steekproeven met dezelfde grootte wensen, dan kunnen we de minimale grootte berekenen met de functie `power.t.test`, met het argument `type = "two.sample"`:

```
> power.t.test(delta = 1, sd = 2.1, power = 0.95, sig.level = 0.05,
alternative = "one.sided",  type = "two.sample" )

     Two-sample t test power calculation

              n = 96.13595
          delta = 1
             sd = 2.1
      sig.level = 0.05
          power = 0.95
    alternative = one.sided

NOTE: n is number in *each* group
```

Zevenennegentig dossiers moeten de standaard procedure volgen; zevenennegentig dossiers moeten de aangepaste procedure volgen om een power van 95% te garanderen.

70. Bij Oef. 57 heb je een t-toets gebruikt om na te gaan of mannelijke en vrouwelijke FPPW studenten van elkaar verschillen qua IQ. Bereken de minimale steekproefgrootte om een power van 90% te garanderen onder de specifieke alternatieve hypothese dat het verschil tussen mannen en vrouwen 5 bedraagt. Zorg ervoor dat $n_1 = n_2$.

7.5 In het algemeen

Voor veel toetsen (maar niet allemaal) is het mogelijk de power te berekenen.

Om de power te verhogen kan je de steekproefgrootte of α verhogen. Een verhoging van de steekproefgrootte heeft gevolgen in termen van kost en tijd. Het is soms niet mogelijk. Een verhoging van α betekent een groter risico om fouten van de eerste soort te maken. Je moet dus de voor- en nadelen tegen elkaar afwegen.

Veel onderzoekers onderschatten het belang van het onderscheidingsvermogen. Ze hebben de indruk dat ze een rigoureus wetenschappelijk werk doen omdat ze statistische toetsen gebruiken. Ze denken dat ze de kans controleren om een fout te maken maar ze vergeten dat er twee soorten fouten zijn. Ze controleren enkel voor de fout van eerste soort en soms is de power zeer laag, kleiner dan 50% zonder dat ze het beseffen. Dit is wetenschappelijk niet verantwoord. Vooraleer je een toets gebruikt moet je zo vaak mogelijk de power berekenen en als de power te laag is, moet je α aanpassen of met een grotere steekproef werken.

De berekening van de power doe je best bij het plannen van je onderzoek. Je moet eerst bepalen welk effect (verschil tussen verwachtingen, effectgrootte, verschil tussen proporties, ...) je wenst te detecteren. Voor een t-toets moet je over een schatting van de standaarddeviatie beschikken (misschien a.d.h.v.

een literatuurstudie of van een pilootonderzoek). En dan ben je klaar om de power te berekenen in functie van de geplande steekproefgrootte. Of je kan de minimale steekproefgrootte berekenen in functie van de gewenste power. Het is aangewezen een paar berekeningen te maken, met verschillende waarden van de effectgrootte (dit is toch iets dat je niet precies kunt bepalen).

7.6 Oplossingen

65) Bereken de power van de binomiale toets, bij het alcoholisten-voorbeeld, met $n = 40$, $\alpha = 0.05$ en onder de specifieke alternatieve hypothese $\pi = 0.15$.

Oplossing:

```
> powerBinom(n = 40, p0 = 0.08, p1 = 0.15, sig.level = 0.05,
   alternative  = "one.sided")

     power and sample size for single binomial response

              n = 40
             p0 = 0.08
             p1 = 0.15
          power = 0.3933435
    alternative = one.sided
      sig.level = 0.05

NOTE: use rejections in correct direction only
```

De power is 39%.

66) Bereken de minimale steekproefgrootte bij het alcoholisten-voorbeeld om een power van 95% te garanderen (met $\alpha = 0.05$ en onder de specifieke alternatieve hypothese $\pi = 0.20$). Gebruik de functie `powerBinom`.

Oplossing:

```
> powerBinom(power = 0.95, p0 = 0.08, p1 = 0.20, sig.level = 0.05,
   alternative  = "one.sided")

     power and sample size for single binomial response

              n = 88
             p0 = 0.08
             p1 = 0.2
          power = 0.9535651
    alternative = one.sided
      sig.level = 0.05

NOTE: use rejections in correct direction only
```

De minimale steekproefgrootte is 88.

67) In Rubr. 6.5.1, p. 94, hebben we een t-toets gebruikt om de verwachting van het IQ van FPPW studenten te vergelijken met het IQ van de doorsnee Vlaming.

Bereken de power van die toets, onder de specifieke alternatieve hypothese: $\mu_X = 110$.

Oplossing:

```
> power.t.test(n = 30, delta = 10, sd = 15.7, sig.level = 0.05,
alternative  = "one.sided", type = "one.sample")

     One-sample t test power calculation

              n = 30
          delta = 10
             sd = 15.7
      sig.level = 0.05
          power = 0.9608578
    alternative = one.sided
```

68) Bij oef. 60 heb je getoetst of Facebook gebruikers in het algemeen meer likes hebben gedaan in 2017 dan in 2016. Stel een relevante specifieke alternatieve hypothese en bereken de minimale steekproegrootte om een power van 90% te garanderen.

Oplossing: Het gemiddelde aantal likes per jaar en per gebruiker is ongeveer 300. Als het stijgt met 10 eenheden van 2016 naar 2017, wil ik dat graag weten. Minder dan 10 vind ik niet interessant (dit is subjectief). We schatten nu de standaarddeviatie van het verschil:

```
> sd <- sd(FB$like2017-FB$like2016)
```

We zijn klaar om de steekproefgrootte te berekenen.

```
> power.t.test(power = 0.9, delta = 10, sd = sd, alternative  = "one.sided", type = "pa

     Paired t test power calculation

              n = 2018.222
          delta = 10
             sd = 153.4632
      sig.level = 0.05
          power = 0.9
    alternative = one.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

We hebben 2018 individuen nodig om een power van 90% te garanderen.

69) Bij Oef. 57 heb je een t-toets gebruikt om na te gaan of mannelijke en vrouwelijke FPPW studenten van elkaar verschillen qua IQ. Bereken de power

van de toets onder de specifieke alternatieve hypothese dat het verschil tussen mannen en vrouwen 5 bedraagt.

Oplossing: We berekenen eerst de effectgrootte:

```
> s2 <- ((nM-1)*var(iqM) + (nV-1)*var(iqV))/(nM+nV-2)
> s2
[1] 238.8524
> d <- 5/sqrt(s2)
> d
[1] 0.3235231
```

Dan de power:

```
> pwr.t2n.test(n1 = length(iqM), n2 = length(iqV), d=d,
sig.level = 0.05, alternative = "two.sided")

     t test power calculation

             n1 = 14
             n2 = 16
              d = 0.3235231
      sig.level = 0.05
          power = 0.1368174
    alternative = two.sided
```

Het is slechts 13%. Duidelijk te klein.

70) Bij Oef. 57 heb je een t-toets gebruikt om na te gaan of mannelijke en vrouwelijke FPPW studenten van elkaar verschillen qua IQ. Bereken de minimale steekproefgrootte om een power 90% te garanderen onder de specifieke alternatieve hypothese dat het verschil tussen mannen en vrouwen 5 bedraagt. Zorg ervoor dat $n_1 = n_2$.

Oplossing: De twee steekproefgrootten moeten identiek zijn. We gebruiken dus `power.t.test` en niet `pwr.t2n.test`. De schatting van de gemeenschappelijke variantie is 238.8524 (zie oplossing van oef. 69).

```
> power.t.test(delta = 5, sd = sqrt(238.8524), power = 0.90,
sig.level = 0.05, alternative = "two.sided", type= "two.sample")

     Two-sample t test power calculation

              n = 201.7431
          delta = 5
             sd = 15.45485
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

```
NOTE: n is number in *each* group
```

We hebben 202 mannen en 202 vrouwen nodig. Dit is veel meer dan wat we feitelijk getrokken hebben. Het is dus geen verrassing dat de toets leidde tot de aanvaarding van de nulhypothese.

Hoofdstuk 8

Enkelvoudige lineaire regressie

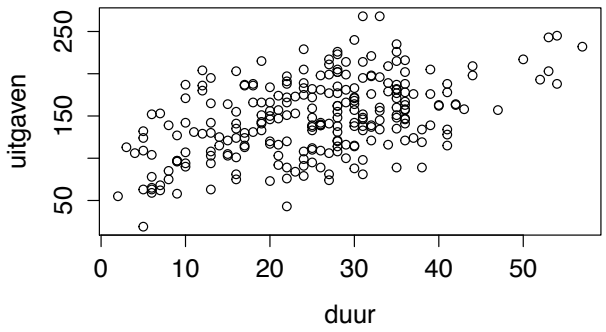
8.1 Inleiding

Een onderzoekster is geïnteresseerd in het eventuele verband tussen werkloosheid en gezondheid en in het bijzonder in de invloed van werkloosheidsduur op de gezondheid. Zij heeft een steekproef van 252 werklozen at random getrokken en heeft volgende variabelen gemeten: duur van werkloosheid (X, in maanden) en totaal bedrag gespendeerd in de laatste vier maanden aan gezondheid, met inbegrip van door het ziekenfonds en gezondheidsverzekeringen terugbetaalde bedragen (Y, in Euros). Zij heeft ook het geslacht en de leeftijd geregistreerd. Zij vermoedt dat werkloosheidsduur een gezondheidsuitgaven samenhangen. In het bijzonder vermoedt zij dat mensen die al lang werkloos zijn hogere gezondheidsuitgaven maken (positief verband). Maar ze wil een negatief verband niet uitsluiten want het zou kunnen dat werklozen gezonder zijn daar ze geen last hebben van werkgerelateerde stress.

Haar data vind je terug in het data frame `gezondheid`. We gebruiken de functie `head` om de eerste zes regels van de data frame te bekijken.

```
> head(gezondheid)
  geslacht duur uitgaven leeftijd
1        M   30      142       39
2        M   25       95       27
3        M   54      188       50
4        V   29      116       41
5        M   21      103       39
6        M   28      185       39
```

De visuele analyse van het spreidingsdiagram (fig. 8.1) bevestigt haar vermoeden: er is een stijgende tendentie. En het is plausibel dat het verband lineair



Figuur 8.1: Spreidingsdiagram van werkloosheidsduur en gezondheidsuitgaven.

is. We kunnen ook de correlatiecoëfficiënt (zie rubr. 2.3.3) berekenen om onze analyse te formaliseren:

```
> cor( gezondheid$duur, gezondheid$uitgaven )
[1] 0.4868292
```

We vinden $r_{XY} = 0.49$. We berekenen nu de regressielijn

```
> lm( formula = gezondheid$uitgaven ~ gezondheid$duur )

Call:
lm( formula = gezondheid$uitgaven ~ gezondheid$duur)

Coefficients:
    (Intercept)  gezondheid$duur
         97.204            2.001
```

71. Teken de de regressielijn van `uitgaven` *op* `duur` *op Fig. 8.1.*

We bekomen $b_0 = 97.2$ en $b_1 = 2.0$. Nu moeten we grondig overwegen vooraleer we besluiten. We hebben een positief verband gevonden in de steekproef maar kunnen we generaliseren op het niveau van de populatie van alle werklozen ? Dit is duidelijk een probleem van inductieve statistiek. Beschikken we over argumenten die sterk genoeg zijn om te kunnen besluiten dat werkloosheidsduur en gezondheidsuitgaven samenhangen in de populatie? Om deze vraag te kunnen beantwoorden moeten we eerst een probabilistisch model van de samenhang hebben.

72. Bereken de coëfficiënten van de regressielijn van `gewicht` *op* `lengte` *m.b.v.* `R` *en het data frame* `sportData`.

8.2 Het enkelvoudig lineair model—Kansrekenen

Stel dat we twee variabelen X en Y hebben en dat we de ene (Y) willen verklaren door de andere (X). We zeggen eveneens dat X de onafhankelijke variabele is en Y de afhankelijke of dat X een predictor van Y is. Voorbeeld: leeftijd is een predictor van de lengte van een kind. De socio-economische status van een man kan grotendeels verklaard worden door de SES van zijn ouders. Het

aantal uren wiskunde in het middelbaar onderwijs is een goede predictor voor de uitslagen in het eerste jaar van het hoger onderwijs. In een experiment waar verschillende doses van een bloeddruk reducerend geneesmiddel gegeven worden is de bloeddruk de afhankelijke variabele en de dosis de onafhankelijke (de bloeddruk hangt af van de dosis).

We vermoeden dat er een lineair verband bestaat tussen X en Y of kortweg dat ze gecorreleerd zijn. Als we dit formeel willen uitdrukken kunnen we niet schrijven, zoals in beschrijvende statistiek, $Y = \beta_0 + \beta_1 X$.[1] Bijvoorbeeld $Y = 97.2 + 2X$. Een bepaalde waarde van X zou dan altijd samengaan met een bepaalde waarde van Y. Bv. als $X = 50$, dan geeft de vergelijking $Y = 97.2 + 2 \times 50 = 197.2$. We weten dat de samenhang tussen twee variabelen niet zo simpel en systematisch is[2]. Laten we aan `R` de uitgaven opvragen die corresponderen met een duur gelijk aan 20 maanden:

```
> gezondheid$uitgaven[gezondheid$duur == 20]
[1] 166 150  73 156 120 146 184 117
```

Er zijn 8 individuen met een werkloosheidsduur gelijk aan 20 maanden, maar met sterk verschillende uitgaven (je kan het ook zien op fig. 8.1, alhoewel minder precies). Misschien is er een lineair verband tussen X en Y maar wordt Y ook beïnvloed door veel andere factoren of variabelen die we niet controleren; variabelen die variëren tussen proefpersonen en over de tijd en in verschillende contexten. In één woord: het toeval. We gaan dus een term bijvoegen in onze vergelijking. Deze term zal de invloed van de andere variabelen, van de ruis, van het toeval representeren. We schrijven nu

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{8.1}$$

De term ε representeert het toeval, het effect van alle variabelen (behalve X) op Y, het effect van meetfouten (daarom wordt ε soms de error of fout genoemd). De term ε is dus een toevalsvariabele[3]. Merk op dat we nu de griekse letter beta (β_0 en β_1) gebruiken en niet meer b_0 en b_1 zoals in rubr. 2.3.3 omdat we nu bezig zijn met de beschrijving van een populatie en niet meer van een steekproef. Met dit model zal een bepaalde waarde van X niet altijd leiden tot een unieke waarde van Y omdat de waarde van ε kan variëren. De coëfficiënt β_1 wordt de regressiecoëfficiënt genoemd.

Laten we vergelijking (8.1) herschrijven voor individu i op voorwaarde dat $X_i = x_i$, waar het symbool x_i de specifieke waarde van de variabele X bij individu i aanduidt. We bekomen

$$\text{Lineair model:} \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{8.2}$$

[1] β is de griekse letter *beta* en komt overeen met onze *b*.

[2] Zo'n verband noemt men deterministisch in tegenstelling tot stochastisch of probabilistisch

[3] Een toevalsvariabele wordt meestal door een latijnse hoofdletter aangeduid. Maar de foutterm is een uitzondering: die wordt door de griekse letter ε (epsilon) aangeduid. Griekse letters worden in principe voor parameters (verwachting, variantie, correlatiecoëfficiënt, regressiecoëfficiënt, ...) gebruikt.

Deze vergelijking noemen we *het enkelvoudig lineair model*. Het is enkelvoudig omdat er maar één predictor is. Een meervoudig lineair model heeft meerdere predictoren zoals bv. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$. Het is lineair omdat de parameters (β_0 en β_1) niet in een niet-lineaire vorm voorkomen, zoals bv. $\beta_1^2, \log \beta_0, \sin \beta_1$, enz. In dit hoofdstuk zullen we het meervoudig lineair model niet bespreken. We spreken dus soms van het *lineair model* tout court.

Merk op dat vergelijking (8.2) ook gebruikt kan worden indien X geen toevalsvariabele is. Bv. indien X de dosis van een geneesmiddel representeert in een experiment waar de onderzoeker zelf bepaalt welke dosis elke patiënt krijgt.

8.2.1 Assumpties

Het lineair model bevat enorm veel parameters: elke fout ε_i heeft zijn eigen verwachting en variantie (dus 2 parameters per individu); elke fout ε_i kan correleren met ε_j (dus $n(n-1)/2$ parameters); en er zijn nog de twee parameters β_0 en β_1. Dus in totaal $2 + 2n + n(n-1)/2$ parameters. Bij een steekproef van 20 individuen zijn er 232 parameters die geschat moeten worden. Zo'n complex model is onbruikbaar: niet alleen omdat de berekeningen moeilijk zouden zijn maar ook omdat het geen predictieve waarde heeft. Je kan de 232 parameters instellen zodat het model je puntenwolk zo goed mogelijk past, maar bij een andere steekproef gaat die instelling helemaal niet werken.

Om met dit model te kunnen werken gaan we een aantal assumpties (de Gauss[4]-Markov[5] assumpties) moeten maken.

1. $E(\varepsilon_i) = 0$ voor alle i. M.a.w. de verwachting van de fout hangt niet af van het individu.

2. $V(\varepsilon_i) = V(\varepsilon_j)$ voor alle i, j. M.a.w. de variantie van de fout hangt niet af van het individu (homoscedasticiteit). Deze constante variantie wordt aangeduid door σ_ε^2.

3. $COV(\varepsilon_i, \varepsilon_j) = 0$ voor alle i, j. M.a.w. de fout bij individu i is niet gecorreleerd met de fout bij individu j (geen seriële correlatie).

Het aantal parameters is dus fors gereduceerd.

Vooraleer we zien hoe we het verband in de populatie kunnen nagaan op basis van een steekproef, gaan we enkele eigenschappen van het lineair model analyseren.

8.2.2 De voorwaardelijke verwachting

De verwachting van de gezondheidsuitgaven (genoteerd $E(Y)$) is de verwachting van de variabele Y in de hele populatie, dus los van de werkloosheidsduur. We kunnen ook de verwachting van de gezondheidsuitgaven definiëren en berekenen bij een deelverzameling van de populatie: bv. bij alle individuen die 10 maanden

[4] Carl Friedrich Gauss, 1777–1855
[5] Andrey Markov, 1856–1922

werkloos zijn geweest. Deze verwachting wordt een voorwaardelijke verwachting genoemd; het is de verwachting van Y onder voorwaarde dat X gelijk is aan 10 en het wordt aangeduid door $E(Y \mid X = 10)$. In het algemeen, $E(Y \mid X = x)$ is de verwachting van de variabele Y onder voorwaarde dat $X = x$.

Laten we nu de voorwaardelijke verwachting van Y analyseren onder de hypothese dat het lineair model geldt. In dat geval

$$E(Y_i \mid X_i = x_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i)$$

en dankzij één van de stellingen in Rubr. 3.1.10

73. Welke stelling?

$$E(Y_i \mid X_i = x_i) = E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i).$$

Omdat de verwachting van ε_i nul is en omdat de verwachting van een getal gelijk aan dat getal is, vinden we uiteindelijk

$$E(Y_i \mid X_i = x_i) = \beta_0 + \beta_1 x_i. \tag{8.3}$$

Merk op dat deze vergelijking nu deterministisch is; $E(Y_i \mid X_i = x_i)$, β_0, β_1 en x_i zijn getallen en geen toevalsvariabelen. Dit komt doordat we niet meer focussen op één realisatie van Y_i maar op de verwachting van *alle* realisaties van Y_i. Het toeval speelt dan geen rol meer.

Merk ook op dat de voorwaardelijke verwachting van Y_i een lineaire functie van x_i is. Als we deze functie grafisch representeren, dan bekomen we een rechte. Deze rechte is het equivalent (in het kansrekenen) van de regressielijn die we in de beschrijvende statistiek gezien hebben. We kunnen verg. (8.3) gebruiken om voorspellingen of predicties te maken. Stel dat het volgend model geldt:

$$Y_i = 20 + 2x_i + \varepsilon_i,$$

met X het gewicht van een 10-jarige jongen en Y van een 30-jarige man. Als we het gewicht x_i van een 10-jarige jongen kennen, kunnen we dan zijn gewicht 20 jaar later voorspellen. We schrijven verg. (8.3) voor het gewicht, dat is

$$E(Y_i \mid X_i = x_i) = 20 + 2x_i$$

en we vervangen x_i door het gewicht van een jongen, bv. 28 kg. De voorspelling van zijn gewicht 20 jaar later is dan $E(Y \mid X = 28) = 20 + 2 \times 28 = 76$. Deze voorspelling is uiteraard waarschijnlijk fout. Het is maar een voorspelling maar als we deze formule herhaaldelijk gebruiken om predicties te maken dan zullen onze predicties gemiddeld gezien correct zijn.

De voorwaardelijke verwachting wordt vaak afgekort als $E(Y_i \mid x_i)$. De formule voor een predictie wordt dan

$$E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i. \tag{8.4}$$

Het verschil tussen Y_i en de predictie van Y_i is

$$Y_i - \beta_0 - \beta_1 x_i.$$

Als je dit vergelijkt met (8.2), dan vind je

$$Y_i - \beta_0 - \beta_1 x_i = \varepsilon_i.$$

Dit is een andere manier om de fout te bekijken: het is het equivalent van de residuen in de beschrijvende statistiek (zie Fig. 2.9). De variantie van ε_i, i.e., σ_ε^2, is dus de variantie van de populatie-residuen.

8.2.3 De voorwaardelijke variantie

We definiëren nu de voorwaardelijke variantie: $V(Y \mid X = x)$ is de variantie van Y onder voorwaarde dat X gelijk is aan een bepaalde waarde x. We kunnen ook de voorwaardelijke variantie analyseren onder de hypothese dat het lineair model geldt. In dat geval,

$$V(Y_i \mid X_i = x_i) = V(\beta_0 + \beta_1 x_i + \varepsilon_i)$$

74. Welke stelling?

en dankzij één van de stellingen in Rubr. 3.1.10,

$$V(Y_i \mid X_i = x_i) = V(\beta_0 + \beta_1 x_i) + V(\varepsilon_i) + 2\,COV(\beta_0 + \beta_1 x_i, \varepsilon_i).$$

De variantie van ε_i is gelijk aan σ_ε^2; de variantie van $\beta_0 + \beta_1 x_i$ is nul (het is een getal) en zijn covariantie met ε_i is dus ook nul. Bijgevolg,

$$V(Y_i \mid X_i = x_i) = \sigma_\varepsilon^2.$$

De voorwaardelijke variantie van Y_i is dus gelijk aan σ_ε^2; het is onafhankelijk van x_i.

Hoe kunnen we dit interpreteren? Stel dat we het gewicht observeren van *alle* 30-jarige mannen die 20 jaren geleden 28 kg wogen. We observeren zeker niet hetzelfde gewicht voor al die mannen: het variëert. Toch kunnen we een gewicht van 76 kg voorspellen voor die mannen. De variantie rond de voorspelling wordt verklaard door het toeval en is gelijk aan σ_ε^2. Dit wordt geïllustreerd in Fig.8.2.

8.2.4 De correlatiecoëfficiënt

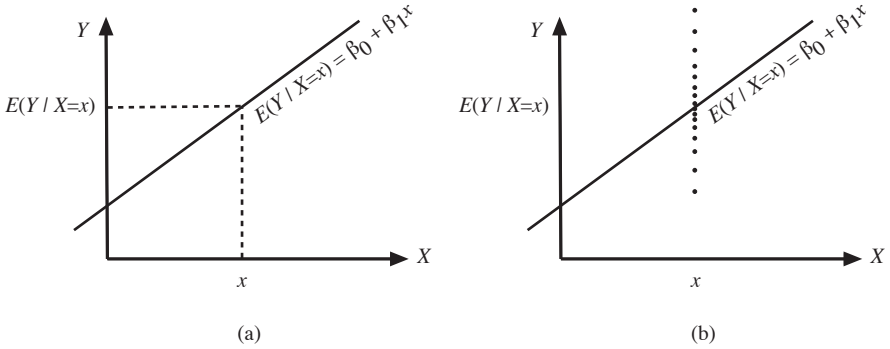
Het lineair model (8.2) heeft betrekking tot het lineair verband tussen twee toevalsvariabelen; anderzijds weten we dat het lineair verband tussen twee toevalsvariabelen gemeten kan worden door de correlatiecoëfficiënt ρ (Rubr. 3.1.9). Wat is dan het verband tussen het lineair model en de correlatiecoëfficiënt? Het is simpel:

$$\beta_1 = \rho_{XY}\,\frac{\sigma_Y}{\sigma_X}.$$

Deze formule komt overeen met de formule

$$b_1 = r_{XY}\,\frac{s_Y}{s_X}$$

in de beschrijvende statistiek.



Figuur 8.2: (a) De voorspelling van Y voor een bepaalde waarde van X en (b) de spreiding van de realisaties van Y wanneer X vast is.

8.2.5 Afsluiter

In Rubr. 8.2 hebben we het lineair model vanuit een kansrekenen-perspectief geanalyseerd. Het heeft betrekking tot toevalsvariabelen (i.t.t. geobserveerde variabelen) in populaties (niet in een specifieke steekproef). Het lineair model bevat drie parameters: β_0, β_1 en σ_ε^2 en ze zijn bijna altijd onbekend want de meeste populaties zijn te groot om volledig onderzocht te kunnen worden.

8.3 Puntschatting

Nu we over een model beschikken voor het lineair verband tussen twee variabelen kunnen we het voorbeeld omtrent gezondheidsuitgaven en werkloosheidsduur verder verwerken. Nemen we aan dat het lineair model geldt tussen X en Y, wat zijn dan de waarden van de parameters β_0, β_1 en σ_ε^2? En hoe vinden we ρ_{XY} ? Hoe kunnen we die parameters schatten op basis van een steekproef ?

8.3.1 Puntschatting van β_1

De beste schatter van β_1 is gewoon B_1; t.t.z. de toevalsvariabele die in elke steekproef gelijk is aan b_1. We mogen dus de steekproefwaarde b_1 gebruiken als schatting van β_1 in de populatie. De schatter B_1 is zuiver ($E(B_1) = \beta_1$) en efficiënt; zijn variantie is

$$V(B_1) = \frac{\sigma_\varepsilon^2}{\mathrm{SS}_X} = \frac{\sigma_\varepsilon^2}{(n-1)s_X^2}. \tag{8.5}$$

De standaardfout (zie Rubr. 4.2) van de schatter B_1 is dus

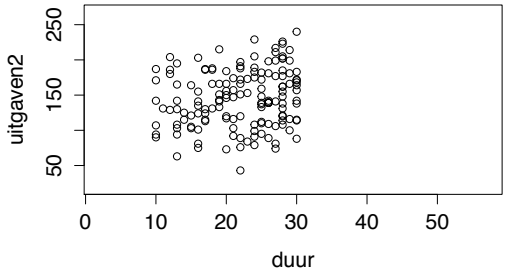
$$\mathrm{SE}_{B_1} = \sqrt{\frac{\sigma_\varepsilon^2}{\mathrm{SS}_X}} = \frac{\sigma_\varepsilon}{\sqrt{n-1}\ s_X}.$$

Om goede schattingen van β_1 uit te komen, hebben we een goede schatter nodig, i.e., een schatter waarvan de variantie zo klein mogelijk is. Hoe kunnen we deze variantie beïnvloeden?

- σ_ε^2 moet zo klein mogelijk zijn. Omdat ε het toeval representeert, i.e., alles wat we niet controleren, moeten we dus ervoor zorgen dat het effect van het toeval zo klein mogelijk is. We gebruiken dus een predictor met een hoge betrouwbaarheid; en in een experimentele setting proberen we zoveel mogelijk variabelen constant te houden.

- n moet zo groot mogelijk zijn. Dit is evident: hoe groter de steekproef, hoe beter de schatting.

- s_X^2 moet zo groot mogelijk zijn. Dit kan door een brede range van X waarden te kiezen (indien het onderzoeksopzet dit toelaat). We illustreren dit met een aangepaste versie van het werkloosheid-gezondheid voorbeeld. Stel dat de onderzoekster voor een onbepaalde reden enkel werklozen heeft geselecteerd met een werkloosheidsduur tussen 10 en 30 maanden. Haar steekproef bestaat uit dezelfde individuen als aan het begin van het hoofdstuk, maar zonder de individuen met een duur kleiner dan 10 of groter dan 30. In Fig. 8.3 vind je het spreidingsdiagram van de variabelen duur en uitgaven in deze beperkte steekproef. De stijgende tendentie is veel minder



Figuur 8.3: (a) Dezelfde gegevens als op Fig. 8.1, maar zonder de individuen waarvoor de duur kleiner dan 10 is of groter dan 30. De stijgende tendentie is nu moeilijk te zien.

duidelijk omdat we slechts een beperkt deel van het diagram observeren. Indien deze onderzoekster deze steekproef gebruikt om β_1 te schatten, dan zal haar schatting toevallig misschien goed zijn. Maar indien zij meerdere steekproeven met altijd dezelfde beperking (range restrictie), dan zullen haar schattingen sterk variëren en ze zullen dus vaak slecht zijn. Het is dus belangrijk dat we zorgen om een brede range van X te observeren.

8.3.2 Puntschatting van β_0

De beste schatter van β_0 is gewoon B_0; t.t.z. de toevalsvariabele die in elke steekproef gelijk is aan b_0. We mogen dus de steekproefwaarde b_0 gebruiken als

schatting van β_0 in de populatie. De schatter B_0 is zuiver ($E(B_0) = \beta_0$) en efficiënt; zijn variantie is

$$V(B_0) = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_X^2} \right) = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\mathrm{SS}_X} \right) \tag{8.6}$$

en zijn standaardfout (zie Rubr. 4.2):

$$\mathrm{SE}_{B_0} = \sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_X^2}} = \sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\mathrm{SS}_X}}.$$

Om goede schattingen van β_0 uit te komen, zorgen we ook ervoor dat σ_ε^2 zo klein mogelijk is terwijl n en s_X^2 zo groot mogelijk zijn (zie vorige paragraaf).

8.3.3 De predicties

We hebben gezien (Rubr. 8.2.2) dat de predicties van het lineair model gegeven worden door

$$E(Y_i \,|\, x_i) = \beta_0 + \beta_1 x_i.$$

In de praktijk kennen we β_0 en β_1 niet. Om predicties te maken gebruiken we dus de schatters B_0 en B_1 i.p.v. de parameters β_0 en β_1, in bovenstaande formule. Het resultaat is niet meer een predictie maar de schatter van een predictie, gedefinieerd door $B_0 + B_1 x_i$ en aangeduid door het symbool $\widehat{Y}_i$. Dus

$$\widehat{Y}_i = B_0 + B_1 x_i.$$

Dit wordt meestal gewoon een predictie genoemd. In een specifieke steekproef kunnen we de realisaties b_0 en b_1 van de schatters B_0 en B_1 berekenen. We bekomen dan de schattingen van de predicties (aangeduid door $\hat{y}_i$)[6]:

$$\widehat{E(Y_i \,|\, x_i)} = b_0 + b_1 x_i = \hat{y}_i.$$

Ze worden ook meestal gewoon predicties genoemd. De variantie van de schatter $\widehat{Y}_i$ is

$$V(\widehat{Y}_i) = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)s_X^2} \right) = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\mathrm{SS}_X} \right). \tag{8.7}$$

Deze variantie wordt beïnvloed door σ_ε^2, n en s_X^2, net zoals $V(B_0)$ en $V(B_1)$. Maar $V(\widehat{Y}_i)$ wordt ook beïnvloed door $(x_i - \overline{x})^2$. We zien dus dat hoe dichter x_i bij het gemiddelde $\overline{x}$, hoe kleiner de variantie en bijgevolg hoe beter de predicties.

We kunnen ook het lineair model gebruiken om predicties te maken voor nog niet geobserveerde waarden:

$$\widehat{Y} = B_0 + B_1 x.$$

[6] De notatie is misleidend: $\hat{y}_i$ is de schatting van $E(Y_i \,|\, x_i)$ en niet van y_i.

De variantie van deze schatter wordt gegeven door dezelfde formule:

$$V(\widehat{Y}) = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_X^2}\right) = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\mathrm{SS}_X}\right).$$

De predicties voor nog niet geobserveerde waarden zijn dus ook beter als x niet te ver van het gemiddelde ligt.

8.3.4 Puntschatting van σ_ε^2

We hebben gezien dat σ_ε^2 de variantie van de fouten (populatie-residuen) is (Rubr. 8.2.2), i.e., de variantie van

$$Y_i - \beta_0 - \beta_1 x_i.$$

De beste schatter ervan is

$$S_\varepsilon^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_i)^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

Deze schatter is zuiver en efficiënt. De corresponderende schatting is

$$\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Naar analogie met de notatie SS_X voor $\sum_{i=1}^n (X_i - \overline{X})^2$ gebruiken we nu de notatie $\mathrm{SS}_{\mathrm{Res}}$ voor $\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$. Het staat voor sum of squared residuals. Dezelfde notatie wordt gebruikt voor $(y_i - \widehat{y}_i)^2$. We hebben dus een equivalente formule voor de puntschatting van σ_ε^2:

$$\hat{\sigma}_\varepsilon^2 = \frac{\mathrm{SS}_{\mathrm{Res}}}{n-2}. \tag{8.8}$$

8.3.5 Puntschatting van ρ_{XY}

De beste schatter van ρ_{XY} is de steekproefgrootheid R_{XY} waarvan de realisatie in een steekproef gelijk aan r_{XY} is. Deze schatter is zuiver en efficiënt.

8.3.6 Illustratie

In het begin van dit hoofdstuk hebben we al r_{XY} berekend voor de werkloosheidsduur en de gezondheidsuitgaven. We kwamen 0.49 uit. De schatting van ρ_{XY} is dus $\hat{\rho}_{XY} = 0.49$.

We hebben ook b_0 en b_1 berekend; we kunnen dus gemakkelijk β_0 en β_1 schatten: we vinden $\hat{\beta}_0 = 97.2$ en $\hat{\beta}_1 = 2.0$.

Voor de variantie van ε zijn de berekeningen langer. In het begin van dit hoofdstuk hebben we de functie `lm` gebruikt:

```
> lm( formula = gezondheid$uitgaven ~ gezondheid$duur )

Call:
lm(formula = gezondheid$uitgaven ~ gezondheid$duur)

Coefficients:
    (Intercept)  gezondheid$duur
         97.204            2.001
```

De output van deze functie is zeer beperkt (twee coëfficiënten) maar achter de schermen heeft `R` veel andere dingen berekend. Om de uitkomst van die berekeningen te kunnen raadplegen, gaan we een naam toekennen aan het resultaat van de berekeningen:

```
> myLM <- lm( formula = gezondheid$uitgaven ~ gezondheid$duur )
```

We kunnen bv. de predicties $\hat{y}_i$ opvragen:

```
> fitted( myLM )
       1        2        3        4        5        6        7
157.2257 147.2220 205.2431 155.2249 139.2191 153.2242 135.2177
       8        9       10       11       12       13       14
147.2220 153.2242 131.2162 135.2177 129.2155 179.2336 151.2235
     ...      ...      ...      ...
     246      247      248      249      250      251      252
101.2054 115.2104 109.2083 111.2090 131.2162 127.2148 121.2126
```

En je krijgt de lijst van de 252 predicties. Of

```
> residuals( myLM )
          1           2           3           4           5
-15.2256518 -52.2220266 -17.2430529 -39.2249268 -36.2191264
          6           7           8           9          10
 31.7757983  15.7823237  34.7779734  69.7757983  54.7837738
         11          12          13          14          15
        ...         ...         ...         ...
        251         252
 36.7852239  82.7873990
```

en je krijgt de lijst van de 252 residuen $y_i - \hat{y}_i$ (i.e., de lijst van de 252 verticale afwijkingen tussen de regressielijn en de geobserveerde punten). We kunnen dit gebruiken om $\hat{\sigma}_\varepsilon^2$ te berekenen, m.b.v. de formule

$$\hat{\sigma}_\varepsilon^2 = \frac{\mathrm{SS}_{\mathrm{Res}}}{n-2}.$$

```
> sum( residuals( myLM )^2 ) / 250
[1] 1556.19
```

Bijgevolg $\hat{\sigma}_\varepsilon^2 = 1556$ en $\hat{\sigma}_\varepsilon = \sqrt{1556} = 39.45$.

8.4 Intervalschatting

75. Bereken $\hat{\sigma}_\varepsilon^2$ voor het lineair model met `gewicht` *als afhankelijke variabele en* `lengte` *als predictor (data frame* `sportData`*).*

In deze rubriek gaan we veronderstellen, naast de Gauss-Markov assumpties, dat de fouten normaal verdeeld zijn. De combinatie van deze hypothese met de Gauss-Markov assumpties leidt

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \text{voor alle } i.$$

8.4.1 Betrouwbaarheidsinterval voor β_1

In het hoofdstuk omtrent betrouwbaarheidsintervallen hebben we een algemene formule (5.2) gezien voor de meest courante betrouwbaarheidsintervallen:

$$\left[\hat{\theta} \pm t_{l;\alpha/2}\ \mathrm{SE}_Q\right].$$

Indien we deze formule hier toepassen, vinden we het tweezijdige betrouwbaarheidsinterval voor β_1 met betrouwbaarheid $1-\alpha$:

$$\left[\ b_1 - t_{n-2;\alpha/2}\ \sqrt{\widehat{V(B_1)}}\ \ ,\ \ b_1 + t_{n-2;\alpha/2}\ \sqrt{\widehat{V(B_1)}}\ \right]$$

of

$$\left[\ b_1 - t_{n-2;\alpha/2}\ \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\mathrm{SS}_X}}\ \ ,\ \ b_1 + t_{n-2;\alpha/2}\ \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\mathrm{SS}_X}}\ \right].$$

Je betrouwbaarheidsinterval zal smal zijn indien $V(B_1)$ klein is. Zorg dus ervoor dat het klein is (zie Rubr. 8.3.1).

8.4.2 Betrouwbaarheidsinterval voor β_0

Hier kunnen we ook (5.2) toepassen. Het tweezijdige betrouwbaarheidsinterval voor β_0 met betrouwbaarheid $1-\alpha$ is

$$\left[\ b_0 - t_{n-2;\alpha/2}\ \sqrt{\widehat{V(B_0)}}\ \ ,\ \ b_0 + t_{n-2;\alpha/2}\ \sqrt{\widehat{V(B_0)}}\ \right]$$

of

$$\left[\ b_0 - t_{n-2;\alpha/2}\ \hat{\sigma}_\varepsilon\ \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\mathrm{SS}_X}}\ \ ,\ \ b_0 + t_{n-2;\alpha/2}\ \hat{\sigma}_\varepsilon\ \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\mathrm{SS}_X}}\ \right].$$

We gebruiken de `R` functie `confint` om de betrouwbaarheidsintervallen voor β_0 en β_1 te bekomen:

```
> confint( myLM, level = 0.95 )
                   2.5 %    97.5 %
(Intercept)     84.88996 109.51784
gezondheid$duur  1.55357   2.44788
```

Merk op dat de functie `confint` het argument `level` gebruikt en niet `sig.level` of `conf.level`.

8.5 Toetsing

76. Bereken de betrouwbaarheidsintervallen voor β_0 en β_1 voor het lineair model met `gewicht` *als afhankelijke variabele en* `lengte` *als predictor (data frame* `sportData`*).*

In deze rubriek gaan we ook veronderstellen, naast de Gauss-Markov assumpties, dat de fouten normaal verdeeld zijn. De combinatie van deze hypothese met de Gauss-Markov assumpties leidt

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \text{voor alle } i.$$

De hypothese die men wenst te toetsen kan op verschillende maar equivalente manieren uitgedrukt worden:

$$H_0 : \rho_{XY} = 0$$

of

$$H_0 : \beta_1 = 0.$$

Ze zijn equivalent omwille van het verband tussen ρ_{XY} en β_1, uitgedrukt in onderstaande formule:

$$\beta_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X}.$$

Indien β_1 nul is, dan is ρ_{XY} ook nul, en omgekeerd. Het toetsen van de ene of van de andere maakt dus geen verschil.

Het is ook mogelijk te toetsen of $\beta_0 = 0$, maar deze hypothese is zelden interessant en wordt bijna nooit getoetst. We zien deze toets niet.

Voorwaarden. Om de toetsen van dit hoofdstuk te mogen gebruiken, moet de afhankelijke variabele (Y) continu zijn en van interval of ratio meetniveau. Voor discrete afhankelijke variabelen zijn er andere technieken, bv. logistische regressie. Die technieken worden in deze cursus niet gezien.
De onafhankelijke variabele (X) moet van interval of ratio meetniveau zijn of moet 0-1 zijn. De fouten ε_i moeten normaal verdeeld zijn (dit wordt nagegaan met een normale qq-plot) of de steekproef moet groot zijn.
De Gauss-Markov assumpties moeten ook voldaan zijn.

Als de onafhankelijke variabele dichotoom is, maar niet 0-1, dan mag je altijd de codering aanpassen. Bv. 0 voor man en 1 voor vrouw.

8.5.1 Toetsen van het lineair model via de t-verdeling

De nulhypothese $H_0 : \beta_1 = 0$ wordt getoetst a.d.h.v. de toetsingsgrootheid

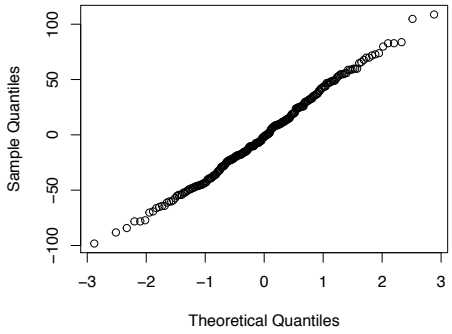
$$\frac{B_1}{\sqrt{\frac{\mathrm{SS_{Res}}}{(n-2)\ \mathrm{SS}_X}}}$$

met een t-verdeling met $n - 2$ vrijheidsgraden.

Toepassing We willen toetsen of er een lineair verband is tussen gezondheidsuitgaven en werkloosheidsduur. Om het lineair model te toetsen moeten we eerst nagaan of de fouten normaal verdeeld zijn. Dit kunnen we natuurlijk niet zeker weten omdat we de ganse populatie niet kunnen onderzoeken, maar we kunnen visueel nagaan of het plausibel is dat de geobserveerde residuen de uitkomst zijn van een normaal verdeeld proces. Dit doen we met een normale qq-plot (zie Rubr. 6.7.1).

```
> qqnorm(residuals(myLM))
```

De output wordt in Fig. 8.4 weergegeven en we zien geen aanwijzing dat de residuen niet normaal verdeeld zijn.



Figuur 8.4: De normale qq-plot voor de residuen van `myLM`.

77. Ga na of de fouten normaalverdeeld zijn bij het lineair model met `gewicht` *als afhankelijke variabele en* `lengte` *als predictor (data frame* `sportData`*).*

We berekenen nu de realisatie van de toetsingsgrootheid in onze steekproef:

$$\frac{b_1}{\sqrt{\frac{\mathrm{SS_{Res}}}{(n-2)\ \mathrm{SS}_X}}}.$$

We vragen eerst de schattingen van β_0 en β_1 op.

```
> coef( myLM )
    (Intercept) gezondheid$duur
      97.203900        2.000725
```

Dus $b_1 = 2.000725$.

```
> SSR <- sum( residuals( myLM )^2 )
> SSX <- sum( ( gezondheid$duur - mean( gezondheid$duur ) )^2 )
> 2.000725 / sqrt( SSR / ( 250 * SSX ) )
[1] 8.813427
```

De p-waarde (tweezijdig) vinden we dankzij de functie `pt`:

```
> 2 * pt(q=8.813427, df = 250, lower.tail = FALSE)
[1] 2.096873e-16
```

De p-waarde is kleiner dan 0.05 en we verwerpen de nulhypothese: er is wel een lineair verband in de populatie tussen de toevalsvariabelen gezondheidsuitgaven en werkloosheidsduur; m.a.w. $\beta_1 \neq 0$. Deze conclusie konden we ook trekken uit het betrouwbaarheidsinterval voor β_1 omdat de waarde nul niet binnen het interval $[1.55357, 2.44788]$ ligt (zie Rubr. 8.4.1).

78. Toets de nulhypothese $\beta_1 = 0$ bij het lineair model met `gewicht` *als afhankelijke variabele en* `lengte` *als predictor (data frame* `sportData`*).*

8.5.2 Toetsen van het lineair model via de F-verdeling

Er bestaat een tweede techniek om het lineair model te toetsen. Het is volledig equivalent aan de eerste techniek. Het is dus in principe niet nuttig om deze techniek te kennen. Maar deze techniek ligt aan de kern van een algemene methode om modellen te vergelijken, en je zal die methode wel nodig hebben in het volgende hoofdstuk. De nieuwe techniek is gebaseerd op de modelvergelijking- of modelselectie-aanpak.

8.5.2.1 Het nulmodel

We gaan tussen twee modellen moeten kiezen: enerzijds het lineair model, dat we al kennen, en anderzijds het nulmodel. Het nulmodel is een speciaal geval van het lineair model, of een beperkte versie van het lineair model; het is het lineair model met de beperking dat $\beta_1 = 0$.

$$\text{Nulmodel:} \quad Y_i = \beta_0 + \varepsilon_i. \tag{8.9}$$

We kunnen dit ook zien als een lineair model zonder predictor. We zeggen dat het nulmodel genest is in het lineair model met één predictor. We maken hier nog de Gauss-Markov assumpties. Met het nulmodel kunnen we predicties maken:

$$E(Y_i \,|\, x_i) = E(\beta_0 + \varepsilon_i) = E(\beta_0) + E(\varepsilon_i) = \beta_0.$$

De predictie is onafhankelijk van x_i. Dit is normaal omdat dit model veronderstelt dat er geen verband is tussen X en Y. De fout, t.t.z. het verschil tussen Y_i en de predictie van Y_i, is gelijk aan

$$Y_i - \beta_0 = \varepsilon_i.$$

Op basis van gegevens in een steekproef kunnen we ook predicties maken: te dien einde gebruiken we niet β_0 (onbekend) maar de schatter van β_0, met name B_0. We bekomen dus

$$\widehat{E(Y_i \,|\, x_i)} = B_0.$$

Het is mogelijk te bewijzen dat de schatter B_0 eigenlijk gelijk is aan $\overline{Y}$. Dus

$$\widehat{E(Y_i \,|\, x_i)} = B_0 = \overline{Y}.$$

In een specifieke steekproef hebben we

$$\widehat{E(Y_i \,|\, x_i)} = b_0 = \overline{y}.$$

De schattingen van de predicties zijn natuurlijk meestal fout en we definiëren dus de populatie-residuen

$$Y_i - \overline{Y}$$

en hun realisaties

$$y_i - \overline{y}.$$

De som van de gekwadrateerde populatie-residuen is

$$\mathrm{SS}_{\mathrm{Res}} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

en zijn realisatie is

$$\mathrm{SS}_{\mathrm{Res}} = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

8.5.2.2 Vergelijking

Het lineair model met één predictor is flexibeler dan het nulmodel. Het kan de gegevens beter passen of fitten (met een schuine rechte kan je de puntenwolk beter passen dan met de horizontale rechte die correspondeert met $\overline{y}$). De som van de gekwadrateerde residuen van het lineair model ($\mathrm{SS}_{\mathrm{Res1}}$) is dus kleiner dan de som van de gekwadrateerde residuen van het nulmodel ($\mathrm{SS}_{\mathrm{Res0}}$). Ja, maar een beetje kleiner of veel kleiner? Is het verschil groot genoeg om te kunnen beslissen dat het niet toevallig is? Om te beslissen dat het lineair model met één predictor geldig is?

Om dit na te gaan, gaan we het verschil $\mathrm{SS}_{\mathrm{Res0}} - \mathrm{SS}_{\mathrm{Res1}}$ analyseren. Is het klein? Dan wijst het aan dat de residuen van het lineair model met één predictor bijna even groot zijn als de residuen van het nulmodel. Het verschil $\mathrm{SS}_{\mathrm{Res0}} - \mathrm{SS}_{\mathrm{Res1}}$ kan dus aan het toeval toegeschreven worden. Is het groot? Dan kan het niet aan het toeval toegeschreven worden. Hoe weten we dat het groot is? De gekwadrateerde som van de residuen is afhankelijk van de meeteenheid en van de steekproefgrootte en van het aantal parameters van de modellen. Het is dus onmogelijk om het verschil rechtstreeks te interpreteren. Om die invloeden te neutraliseren, berekenen we een verhouding:

$$\frac{(\mathrm{SS}_{\mathrm{Res0}} - \mathrm{SS}_{\mathrm{Res1}})/(df_0 - df_1)}{\mathrm{SS}_{\mathrm{Res1}}/df_1}, \tag{8.10}$$

met df_0 het aantal vrijheidsgraden van het nulmodel (dat is $n-1$) en df_1 het aantal vrijheidsgraden van het lineair model met één predictor (dat is $n-2$). Het is mogelijk te bewijzen dat, onder "$H_0 : \beta_1 = 0$" (indien het nulmodel geldt), deze verhouding F-verdeeld is met $df_0 - df_1$ vrijheidsgraden in de teller en df_1 vrijheidsgraden in de noemer. In ons geval wordt de verhouding

$$\frac{\mathrm{SS}_{\mathrm{Res0}} - \mathrm{SS}_{\mathrm{Res1}}}{\mathrm{SS}_{\mathrm{Res1}}/(n-2)} \sim F_{1,n-2}.$$

Om te beslissen of $\text{SS}_{\text{Res0}} - \text{SS}_{\text{Res1}}$ groot is, moeten we gewoon de kans (p-waarde) berekenen dat F toevallig (onder H_0) groter is dan de realisatie van de verhouding in onze steekproef. Indien deze kans kleiner dan 5% is, dan betekent dit dat zo'n F-verhouding onwaarschijnlijk is als H_0 correct is. We verwerpen dus H_0.

Bij deze toets berekenen we een éénzijdige p-waarde omdat alleen hoge waarden van de F-verhouding leiden naar een verwerping van de nulhypothese.

De F toetsingsgrootheid (8.10) kan gebruikt worden om complexere geneste modellen te toetsen. Je moet natuurlijk de geschikte aantallen vrijheidsgraden df_0 en df_1 gebruiken. Dit wordt in Hoofdstuk 9 en in andere vakken gezien.

Illustratie We gaan opnieuw het lineair model toetsen, in het voorbeeld van de gezondheidsuitgaven. Deze keer gebruiken we niet de t-verdeelde toetsingsgrootheid maar de modelvergelijking-aanpak.

```
> SSRes0 <- sum( ( gezondheid$uitgaven - mean(gezondheid$uitgaven) )^2 )
> SSRes1 <- sum( residuals( myLM )^2 )
> SSRes0
[1] 509893.7
> SSRes1
[1] 389047.5
```

De som van de gekwadrateerde residuen is groter bij het nulmodel, zoals verwacht.

```
> fSter <- ( SSRes0 - SSRes1 ) / ( SSRes1 / 250)
> fSter
[1] 77.65515
```

De F-verhouding is veel groter dan 1. Dit wijst aan dat het nulmodel verworpen zal worden. We verifiëren dit m.b.v. de p-waarde:

```
> pf( q = fSter, df1 = 1, df2 = 250, lower.tail = FALSE)
[1] 2.114235e-16
```

De p-waarde is minuscuul en de nulhypothese wordt dus verworpen. Merk op dat de p-waarde dezelfde (op afrondingsfouten na) is als die van Rubr. 8.5.1, bekomen via een t-verdeelde toetsingsgrootheid.

79. Gebruik de modelvergelijking aanpak om het lineair model voor `gewicht` *met* `lengte` *als predictor te toetsen, a.d.h.v.* `sportData`.

8.6 De determinatiecoëfficiënt R^2

Het is mogelijk te bewijzen dat

$$\text{SS}_Y = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{\text{SS}_{\text{Mod}}} + \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{\text{SS}_{\text{Res}}}.$$

We kunnen dus SS_Y in twee termen splitsen. De eerste term

$$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

wordt SS_{Mod} genoteerd (SS_{Mod} staat voor *sum of squares predicted by the model*). Het is de som van de gekwadrateerde afwijkingen tussen het gemiddelde $\overline{y}$ en de predicties $\widehat{y}_i$. Het is de som van de afwijkingen die verklaard of voorspeld worden door het lineair model met één predictor.

De tweede term

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

hebben we al gezien; het is SS_{Res} (sum of squared residuals).[7] Het is de som van de gekwadrateerde residuen, van de fouten; het is wat het lineair model met één predictor niet verklaart.

We hebben dus

$$\text{SS}_Y = \text{SS}_{\text{Mod}} + \text{SS}_{\text{Res}}\,. \tag{8.11}$$

SS_Y wordt gesplitst in twee delen: één dat het model verklaart en één dat het model niet verklaart. Omdat SS_Y uit twee delen bestaat, wordt het ook SS_{Tot} genoemd (total sum of squares).

Fig. 8.5 presenteert drie spreidingsdiagrammen voor de variabelen `duur` en `uitgaven`, waarop slechts 8 punten worden gerepresenteerd om de grafiek leesbaarder te maken. Op het diagram bovenaan worden de afwijkingen tussen y_i en $\overline{y}$ aangeduid d.m.v. verticale lijnen. De som van de gekwadrateerde afwijkingen is SS_{Tot}. Onderaan links worden de afwijkingen tussen $\hat{y}_i$ en $\overline{y}$ aangeduid. De som van de gekwadrateerde afwijkingen is SS_{Mod}. Onderaan rechts worden de afwijkingen tussen y_i en $\hat{y}_i$ aangeduid. De som van de gekwadrateerde afwijkingen is SS_{Res}.

De variantie van Y in een specifieke steekproef is

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{\text{SS}_Y}{n-1}.$$

Bijgevolg,

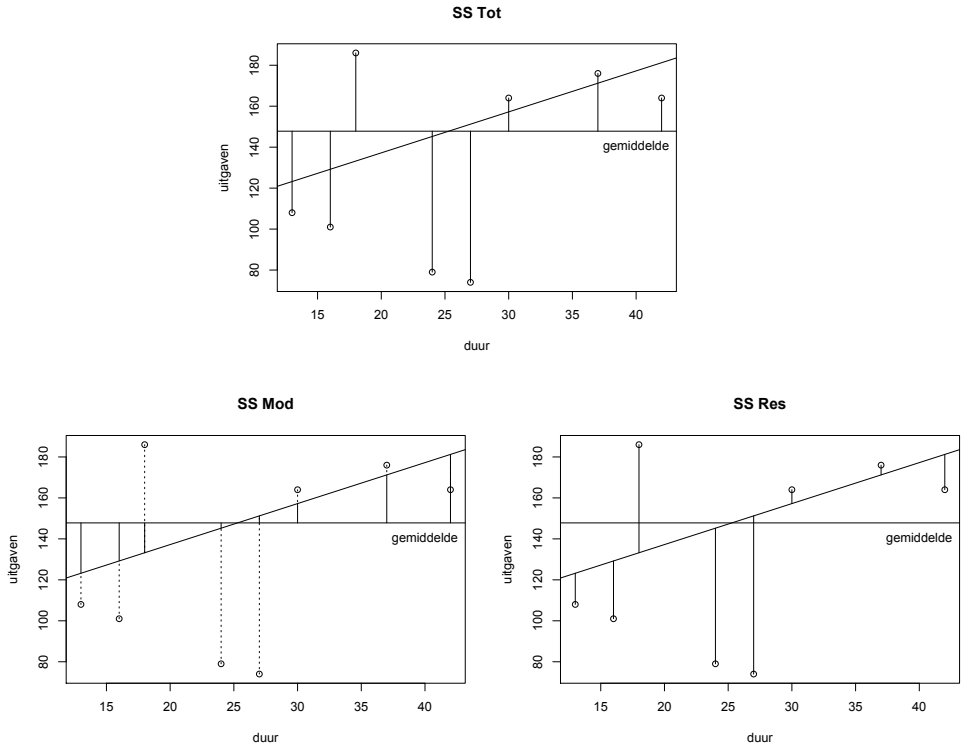
$$s_Y^2 = \frac{\text{SS}_Y}{n-1} = \frac{\text{SS}_{\text{Tot}}}{n-1} = \frac{\text{SS}_{\text{Mod}}}{n-1} + \frac{\text{SS}_{\text{Res}}}{n-1}.$$

De totale variantie van Y kan dus gesplitst[8] worden in twee delen: de verklaarde variantie $\text{SS}_{\text{Mod}}/(n-1)$ en de onverklaarde variantie $\text{SS}_{\text{Res}}/(n-1)$. Dit leidt

[7] In de context van lineaire regressie wordt soms een andere notatie gebruikt. Hieronder de correspondentie.

SS_{Mod}	SSR	R van regression (niet van residu)
SS_{Res}	SSE	E van error
SS_{Tot}	SST	T van total

[8] Dit wordt de variantiedecompositie genoemd.



Figuur 8.5: Spreidingsdiagrammen van werkloosheidsduur en gezondheidsuitgaven.

tot de definitie van een nieuwe coëfficiënt: de determinatiecoëfficiënt R^2

$$R^2 = \frac{\mathrm{SS}_{\mathrm{Mod}}}{\mathrm{SS}_{\mathrm{Tot}}} = \frac{\mathrm{SS}_{\mathrm{Tot}} - \mathrm{SS}_{\mathrm{Res}}}{\mathrm{SS}_{\mathrm{Tot}}}. \tag{8.12}$$

Let op, dezelfde notatie (met een hoofdletter) en dezelfde formule wordt gebruikt voor de steekproefgrootheid en voor zijn realisatie in een specifieke steekproef. Er is geen corresponderend symbool voor de populatieparameter.

Deze coëfficiënt is altijd positief of nul omdat $\mathrm{SS}_{\mathrm{Mod}}$ en $\mathrm{SS}_{\mathrm{Tot}}$ som van kwadraten zijn. Het is ook altijd kleiner dan of gelijk aan 1 omdat $\mathrm{SS}_{\mathrm{Mod}} \leq \mathrm{SS}_{\mathrm{Tot}}$, omwille van (8.11). Dus $0 \leq R^2 \leq 1$.

Laten we een paar gevallen bespreken.

- $R^2 = 1$. Bijgevolg $\mathrm{SS}_{\mathrm{Mod}} = \mathrm{SS}_{\mathrm{Tot}}$ en $\mathrm{SS}_{\mathrm{Res}} = 0$. Dit impliceert dat alle punten op de regressielijn liggen. Het lineair verband is perfect.
- $R^2 = 0$. Bijgevolg $\mathrm{SS}_{\mathrm{Mod}} = 0$ en $\mathrm{SS}_{\mathrm{Res}} = \mathrm{SS}_{\mathrm{Tot}}$. Dit impliceert dat de regressielijn horizontaal is. Er is geen lineair verband tussen de variabelen X en Y.

- $0 < R^2 < 1$. Een deel van SS_{Tot} wordt verklaard door het lineair model, maar niet alles. Er is een lineair verband, in de steekproef, tussen de variabelen X en Y.

Samengevat, hoge waarden van R^2 (dichtbij 1) duiden een sterk lineair verband aan en lage waarden (dichtbij 0) duiden een zwak of geen lineair verband aan. De definitie van R^2 (8.12) en zijn interpretatie blijft ongewijzigd voor lineaire modellen met meerdere predictoren (volgende hoofdstuk).

In het geval van een lineair model met één predictor (enkelvoudige lineaire regressie) is het mogelijk te bewijzen dat de de determinatiecoëfficiënt R^2 gelijk is aan r^2, het kwadraat van de correlatiecoëfficiënt r. In dit geval bevatten de twee coëfficiënten r en R^2 dus exact dezelfde informatie, behalve dat R^2 geen informatie geeft over het teken van het verband: stijgend of dalend.

De aangepaste determinatiecoëfficiënt De waarde van R^2 is gebaseerd op een steekproef van n observaties. Om een betere schatting van de corresponderende populatie-R^2 te bekomen maakt men gebruik van de volgende formule, gebaseerd op een zuivere schatter:

$$\overline{R}^2 = 1 - (1 - R^2)\,\frac{n-1}{n-p-1},$$

met p het aantal predictoren. In het geval van een lineair model met één predictor,

$$\overline{R}^2 = 1 - (1 - R^2)\,\frac{n-1}{n-2},$$

Deze aangepaste (Engels: adjusted) determinatiecoëfficiënt is altijd kleiner dan of gelijk aan R^2. In uitzonderlijke gevallen kan R^2 negatief zijn. Zo'n negatieve waarde kan niet geïnterpreteerd worden en betekent gewoon dat de schatting fout is. De beste schatting is dan gewoon nul.

8.7 De `R` functie `summary`

We hebben de functie `lm` gezien en we hebben ook gezien dat deze functie veel berekeningen doet die in de output niet weergegeven worden. Om de uitkomst van die berekeningen te kunnen raadplegen moet je dus een naam aan de uitkomst toekennen. Bv.

```
> myLM <- lm(formula= gezondheid$uitgaven ~ gezondheid$duur)
```

We kunnen dan allerlei functies gebruiken om meer details te bekomen i.v.m. de regressie. Bv. `fitted(myLM)` of `residuals(myLM)` of `confint(myLM)`. We zien nog zo'n functie:

```
> summary(myLM)

Call:
```

```
lm(formula = gezondheid$uitgaven ~ gezondheid$duur)

Residuals:
    Min      1Q  Median      3Q     Max
-98.220 -27.461  -1.725  26.538 108.774

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       97.204      6.252  15.547   <2e-16 ***
gezondheid$duur    2.001      0.227   8.812   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.45 on 250 degrees of freedom
Multiple R-squared:  0.237,Adjusted R-squared:  0.234
F-statistic: 77.66 on 1 and 250 DF,  p-value: < 2.2e-16
```

Met deze functie krijg je in een keer te zien wat we vooraf moeizaam hebben berekend. De regel "`lm(formula = gezondheid$uitgaven ~ gezondheid$duur)`" herinnert je aan welke regressie werd in `myLM` gestopt.

De volgende drie regels geven wat informatie over de residuen. Hun gemiddelde wordt niet getoond omdat het gemiddelde van de residuen altijd nul is. Hun mediaan is een beetje kleiner dan 0. Een groter verschil zou aanwijzen dat de verdeling niet symmetrisch is en dat de verdeling van ε_i misschien niet normaal is. De eerste en de derde kwartielen zijn bijna symmetrisch. Hier opnieuw zou een sterkere asymmetrie aanwijzen dat ε_i misschien niet normaal is. Hetzelfde geldt voor `Min` en `Max`.

Dan heb je een tabel. De rij '`(Intercept)`' geeft informatie i.v.m. β_0. De rij '`gezondheid$duur`' heeft betrekking op β_1. In de kolom '`estimate`' vinden we de schatting van de corresponderende parameter. Bv. $\widehat{\beta}_1 = 2.001$. In de kolom '`Std. Error`' vind je de standaardfout of standaarddeviatie van de corresponderende schatter. Bv. $\sqrt{V(B_1)} = 0.227$. In de volgende kolom heb je de waarde van de t-verdeelde statistiek die we gebruiken om de corresponderende hypothese te toetsen, i.e., $\beta_0 = 0$ (niet gezien in deze cursus) of $\beta_1 = 0$. We lezen bv. `8.812` in de rij van β_1. Dit is inderdaad de waarde die we vroeger berekenden (Rubr. 8.5.1). En in de laatste kolom vind je de corresponderende p-waarde.

Onder de tabel heb je nog `Residual standard error` (dit is $\hat{\sigma}_\varepsilon$), `Multiple R-squared` en `Adjusted R-squared` (de standaard en aangepaste R^2).

De laatste regel geeft het resultaat van de modelselectie weer. De realisatie f^* van de F-toetsings-grootheid is 77.66 (dit klopt met onze berekeningen).

80. Gebruik de functie `summary` *om het lineair model voor* `gewicht` *met* `lengte` *als predictor te analyseren (data frame* `sportData`*). Vergelijk de output van deze functie met wat je berekend hebt in de vorige oefeningen.*

8.8 De power van de toets van $H_0 : \beta_1 = 0$

We berekenen de power van deze toets m.b.v. de functie `pwr.r.test`. Een argument van deze functie is de waarde van de correlatiecoëfficiënt die je wenst te

kunnen detecteren met een hoge kans. Dit is het equivalent van de effectgrootte bij de t-toets. Om deze waarde[9] te bepalen redeneer je best in termen van de regressiecoëfficiënt β_1 omdat deze coëfficiënt een zeer concrete betekenis heeft. Het is de toename van Y die correspondeert met een toename van X met één eenheid. We illustreren dit met het voorbeeld van de werkloosheid.

81. Gebruik de functie `summary` *om het lineair model voor* `iq` *met* `gewicht` *als predictor te analyseren (data frame* `myData`*). Vergeet niet na te gaan of de fouten normaal verdeeld zijn. Anders mag je het lineair model niet toetsen.*

β_1 is het verschil (gemiddeld gezien) tussen de uitgaven van twee personen waarvan de werkloosheidsduur met één maand verschilt. M.a.w, iemand die één maand langer werkloos is zal β_1 € extra spenderen voor gezondheidsuitgaven.

Stel dat je voor het ministerie van sociale zaken werkt. Stel ook dat $\beta_1 = 1$. Wat betekent dit concreet voor jou? Iemand die vijf jaar (60 maanden) werkloos is gaat viermaandelijks[10] $60 \times 1 = 60$ € extra spenderen, t.o.v. iemand die 0 maand werklooos is. Per jaar is dit $3 \times 60 = 180$ €. In België zijn er ongeveer 300 000 werklozen met een werkloosheidsduur groter dan of gelijk aan 5 jaar. Dit impliceert minstens $300\,000 \times 180 = 54\,000\,000$ € extra per jaar dat de sociale zekerheid moet betalen. Dit is niet verwaarloosbaar. Je wil dit zeker (of bijna) kunnen detecteren. Eigenlijk, een toename van 10 miljoen Euro zou je ook willen detecteren. Je zou dan een beleid kunnen invoeren om dit bedrag te reduceren. Een toename van 5 miljoen zou je misschien niet super relevant vinden omdat een beleid om dat bedrag te reduceren waarschijnlijk meer dan 5 miljoen zou kosten. Je beschouwt dus een toename van 10 miljoen als iets dat je wil detecteren en dit komt ongeveer overeen met $\beta_1 = 0.2$ want $10\,000\,000/(300\,000 \times 3 \times 60) = 0.1851852$.

Nu dat we de relevante β_1 hebben bepaald, moeten we de relevante ρ bepalen a.d.h.v. de formule

$$\beta_1 = \rho\frac{\sigma_Y}{\sigma_X} \quad \text{of} \quad \rho = \beta_1\frac{\sigma_X}{\sigma_Y}.$$

We kennen σ_X en σ_Y niet. We gebruiken schattingen die we in de literatuur vinden of die we uit een pilootonderzoek afleiden. We gebruiken hier de schattingen op basis van onze steekproef:

```
> rho <- 0.2*sqrt(var(gezondheid$duur))/sqrt(var(gezondheid$uitgaven))
> rho
[1] 0.04866528
```

En nu gebruiken we de functie `pwr.r.test`:

```
> pwr.r.test( n = 252, r = 0.04866528 , sig.level = 0.05 )

     approximate correlation power calculation (arctangh transformation)

               n = 252
               r = 0.04866528
```

[9] [Cohen, 1988] heeft normen gesuggereerd voor waarden van de correlatiecoëfficiënt die corresponderen met kleine, medium en grote effecten. Het gebruik van die normen is niet gegrond.

[10] Viermaandelijks omdat Y gedefinieerd is als totaal bedrag gespendeerd in de laatste vier maanden aan gezondheid.

```
      sig.level = 0.05
          power = 0.1200969
    alternative = two.sided
```

We komen een power van 12% uit. Dit is natuurlijk te laag. Voor dit onderzoek hadden we best een groter steekproef getrokken. Hoe groot? We gebruiken opnieuw de functie `pwr.r.test`, met het argument `power` en zonder het argument `n`:

```
> pwr.r.test( r = 0.04866528 , sig.level = 0.05 , power = 0.9)

     approximate correlation power calculation (arctangh transformation)

              n = 4431.754
              r = 0.04866528
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

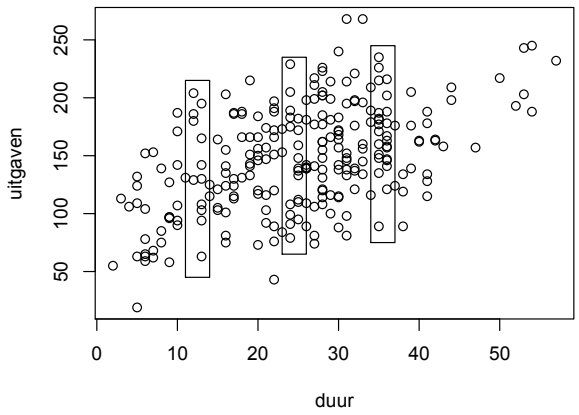
We hebben ongeveer 4 400 individuen nodig!

82. Je wil `iq` *verklaren m.b.v. de predictor* `gewicht`*, met een lineair model. Je wil een regressiecoëfficiënt β_1 gelijk aan -0.1 detecteren met een power van 90%. Welke steekproefgrootte heb je nodig?*

8.9 De validiteit van de Gauss-Markov assumpties

Dit hoofdstuk is volledig gebaseerd op de Gauss-Markov assumpties. Hoe kunnen te weten komen of ze geldig zijn? Voor de derde Gauss-Markov assumptie is er geen simpele techniek. Voor de twee anderen zijn er statistische toetsen, maar een zeer eenvoudige en intuïtieve techniek wordt bij voorkeur gebruikt: de visuele analyse van het spreidingsdiagram.

De tweede Gauss-Markov assumptie Deze assumptie (ook homoscedasticiteitassumptie genoemd) stelt dat de variantie $V(\varepsilon_i)$ onafhankelijk van x_i is. Het is gelijk aan een constante die we σ_ε^2 noteren. Dit heeft als gevolg dat de voorwaardelijke variantie $V(Y_i \mid x_i)$ ook constant is, onafhankelijk van x_i. We kunnen deze variantie visueel analyseren door "sneden" van het spreidingsdiagram te bekijken. In Fig. 8.6 zie je drie sneden, omkaderd door een rechthoek. Ze corresponderen met drie x-waarden (of eerder drie smalle x-intervallen). De spreiding van de punten parallel aan de verticale as is ongeveer dezelfde in de drie sneden. De drie geobserveerde voorwaardelijke varianties zijn dus ongeveer identiek aan elkaar. Dit wijst aan dat de drie voorwaardelijke populatievarianties waarschijnlijk ook gelijk aan elkaar zijn. We kunnen dezelfde analyse doen met andere sneden. Voor alle zulke sneden tussen 5 en 40 is de conclusie dezelfde. Als we sneden links van 5 of rechts van 40 beschouwen, dan is de geobserveerde voorwaardelijke variantie misschien wel kleiner. Maar er zijn ook weinig individuen links van 5 of rechts van 40 en dit maakt de vergelijking

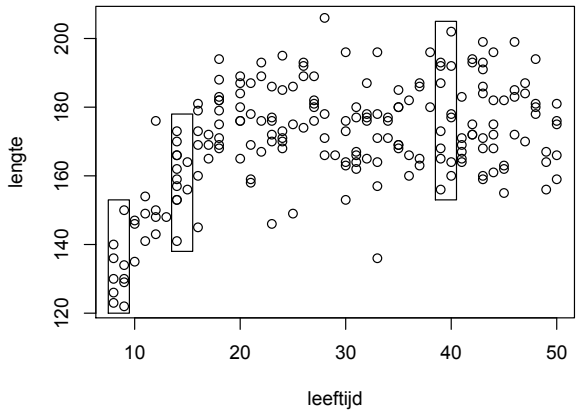


Figuur 8.6: Visuele analyse van de variantie van de residuen.

met de centrale sneden moeilijker. Het is dus plausibel dat de voorwaardelijke populatie-variantie constant is.

Laten we nu dezelfde analyse doen met andere gegevens: `leeftijd` en `lengte` in het data frame `sportData`. In Fig. 8.7 zie je ook drie sneden, omkaderd door een rechthoek. Het is hier duidelijk dat de geobserveerde voorwaar-

83. Je wil het gewicht van een persoon verklaren door het aantal uur dat hij/zij aan sport doet (`sportData`*), met een lineair model. Teken het histogram van de residuen. Op basis hiervan probeer na te gaan of de fouten normaalverdeeld zijn.*



Figuur 8.7: Visuele analyse van de variantie van de residuen.

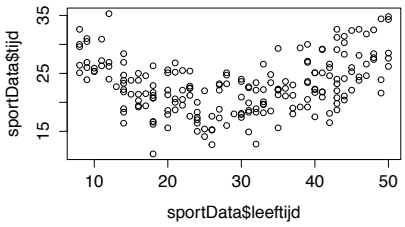
delijke variantie kleiner is in de sneden aan de linkerkant. Dit is ook iets dat we gemakkelijk kunnen verklaren: jonge kinderen hebben min of meer allemaal dezelfde lengte. Vanaf de adolescentie worden de verschillen groter. En indien onze data set ook baby's zou bevatten, dan zou de voorwaardelijke variantie nog kleiner zijn: bijna alle pasgeboren baby's zijn tussen 45 en 55 cm lang. We kunnen dus vermoeden dat de voorwaardelijke populatie-varianties niet allemaal identiek zijn. We mogen het lineair model dus niet gebruiken om het verband

84. Evalueer de validiteit van de homoscedasticiteitsassumptie met de variabelen `leeftijd` *en* `gewicht` *in het data frame* `sportData`*.*

tussen `leeftijd` en `lengte` te analyseren.

Dit probleem treedt frequent op wanneer de afhankelijke variabele van ratio meetniveau is.

De eerste Gauss-Markov assumptie Deze assumptie stelt dat de verwachting $E(\varepsilon_i)$ nul is, en dus onafhankelijk van x_i. Schendingen van deze assumptie kunnen verschillende oorzaken hebben. De belangrijkste oorzaak is de non-lineariteit van het verband tussen Y en X. We illustreren dit a.d.h.v. het data frame `sportData`. Laten we het spreidingsdiagram van `leeftijd` en `tijd` tekenen (Fig. 8.8). We voeren een lineaire regressie uit met `tijd` als afhankelijke

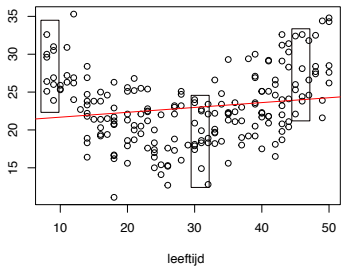


Figuur 8.8: Spreidingsdiagram van `leeftijd` en `tijd`.

variabele en `leeftijd` als predictor.

```
> LM <- lm(formula = sportData$tijd ~ sportData$leeftijd)
> LM

Call:
lm(formula = sportData$tijd ~ sportData$leeftijd)

Coefficients:
       (Intercept)  sportData$leeftijd
          21.04418             0.06421
```

Laten we de regressielijn op het spreidingsdiagram tekenen (Fig. 8.9) alsook een paar verticale sneden. In de linkersnede (`leeftijd` ≈ 9) zie je dat alle punten boven de regressielijn liggen. Alle corresponderende residuen zijn dus positief en, bijgevolg, $E(\varepsilon_i) > 0$ voor alle individuen waarbij `leeftijd` ≈ 9. Analoog zien we dat $E(\varepsilon_i) > 0$ voor alle individuen waarbij `leeftijd` ≈ 46 (rechtersnede) en $E(\varepsilon_i) < 0$ voor alle individuen waarbij `leeftijd` ≈ 31 (middensnede). We hebben dus een duidelijke schending van de eerste Gauss-Markov assumptie. Deze schending is een gevolg van het niet-lineair verband tussen `leeftijd` en `tijd`. Als het verband lineair zou zijn, dan zouden de drie rechthoeken min of meer gecentreerd zijn om de regressielijn (zoals bij Fig. 8.6) en $E(\varepsilon_i)$ zou nul zijn voor alle individuen.



Figuur 8.9: Spreidingsdiagram van `leeftijd` en `tijd`.

8.10 Opmerking m.b.t. softwarepaketten

De waarden van de regressiecoëfficiënten zijn sterk afhankelijk van de schaal waarop de variabelen X en Y gemeten worden. Bijvoorbeeld: indien we de waarden voor de variabele `duur` delen door 12 (omzetting van maand naar jaar), en we voeren de regressie opnieuw uit, dan is de waarde voor $b_1 = 24.012$ of 12 keer groter. Indien we vooraf de data (y_i en x_i voor alle i) standaardiseren, dan zullen ook de regressiecoëfficiënten gestandaardiseerd zijn. Om de scores x_i van de variabele X te standardiseren gebruik je deze formule:

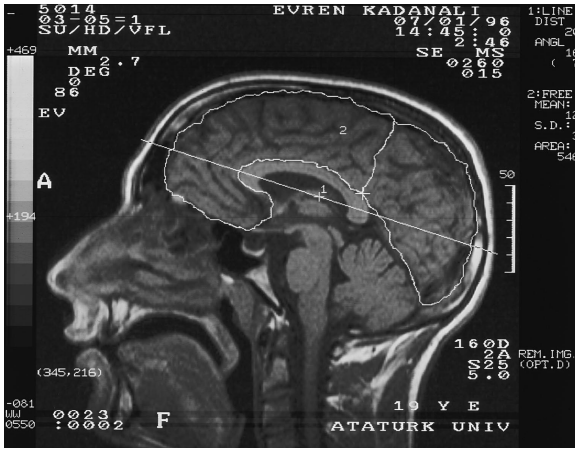
$$\frac{x_i - \overline{x}}{s}.$$

Deze gestandaardiseerde regressiecoëfficiënten kunnen we wel op een zinvolle manier onderling vergelijken. In de output van sommige softwareprogramma's noemt men deze gestandaardiseerde regressiecoëfficiënten `Beta` (niet te verwarren met de populatiewaarden van de regressiecoëfficiënten β).

8.11 Toepassing: Kunnen we het IQ voorspellen m.b.v. de hersengrootte?

Gignac and Bates [2017] analyseren de gegevens van 32 onderzoeken, gebaseerd op een totaal van 1758 individuen en vinden dat de correlatie tussen intelligentie en hersengrootte ongeveer 0.40 is. Daar hun gegevens zeer complex zijn, gaan we eenvoudigere data analyseren: die van Tan et al. [1999].

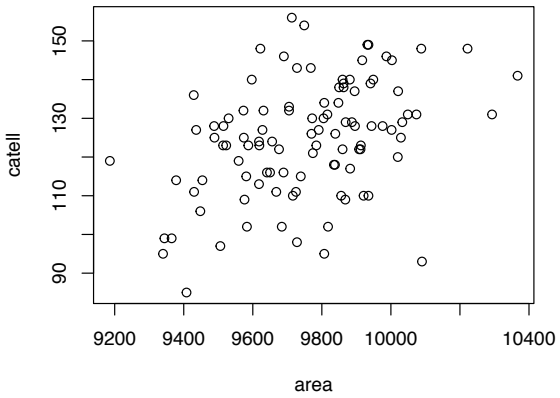
Bij dit onderzoek wordt de hersengrootte van 103 Turkse universiteitsstudenten geoperationaliseerd d.m.v. hersenoppervlakte in het mediaan vlak (midsagittal plane) van de hersenen, gemeten via MRI (zie Fig. 8.10). De variabele `area` wordt uitgedrukt in mm^2. Het IQ wordt gemeten d.m.v. de *Cattell's Culture Fair Intelligence Test battery* [Catell, 1973]. Het is genormaliseerd net zoals de WAIS ($\mu = 100, \sigma = 15$). De hersengrootte wordt ook geoperationaliseerd d.m.v. craniale capaciteit (in mm^3, gebaseerd op schedelmetingen). De dataframe `hersenen` bevat gegevens die vergelijkbaar zijn met die van Tan et al. [1999].



Figuur 8.10: Hersenoppervlakte in het mediaan vlak [Tan et al., 1999].

```
> head(hersenen)
   area catell capacity
1  9784    123     1411
2  9576    109     1397
3  9619    123     1355
4  9618    124     1375
5  9935    149     1476
6 10293    131     1471
```

In dit hoofdstuk gaan we de analyse beperken tot de variabelen `area` en `catell`. Laten we de gegevens eerst visueel analyseren m.b.v. het spreidingsdiagram (Fig. 8.11). Een stijgende tendentie is duidelijk aanwezig en we kunnen niet



Figuur 8.11: Spreidingsdiagram van de variabelen `area` en `catell`.

uitsluiten dat deze tendentie lineair is. We mogen dus de Pearson correla-

tiecoëfficiënt berekenen:

```
> cor(hersenen$catell, hersenen$area)
[1] 0.4017922
```

De correlatiecoëfficiënt is gelijk aan 0.40, zoals bij het onderzoek van Tan et al. [1999]. Dit impliceert niet dat ρ (de correlatiecoëfficiënt in de populatie) verschillend van 0 is want we hebben misschien een niet representatieve steekproef getrokken. We gaan dus een lineaire regressie uitvoeren om onderstaande lineair model te toetsen:

$$\texttt{catell}_i = \beta_0 + \beta_1\ \texttt{area}_i + \varepsilon_i.$$

De afhankelijke variabele is `catell` en de predictor is `area`. De nulhypothese is $\beta_1 = 0$ terwijl de alternatieve hypothese is $\beta_1 \neq 0$. Om deze hypothese te toetsen gebruiken we de functies `lm` en `summary`.

```
> LM <- lm(formula= catell ~ area, data=hersenen)
> summary(LM)

Call:
lm(formula = catell ~ area, data = hersenen)

Residuals:
    Min      1Q  Median      3Q     Max
-40.867  -8.186  -0.335   9.932  32.450

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.423e+02  6.057e+01  -2.349   0.0208 *
area         2.737e-02  6.207e-03   4.410 2.59e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.6 on 101 degrees of freedom
Multiple R-squared:  0.1614,Adjusted R-squared:  0.1531
F-statistic: 19.44 on 1 and 101 DF,  p-value: 2.592e-05
```

We zien dat de p-waarde die geassociëerd is met de predictor `area`, gelijk is aan 2.59e-05, dat is 0.0000259. We moeten dus de nulhypothese verwerpen en we besluiten dat het IQ voorspeld kan worden door de hersengrootte.

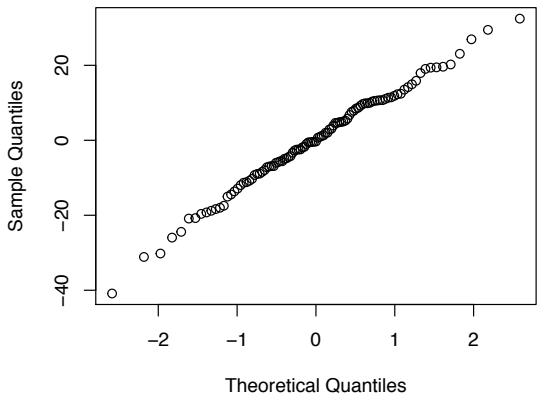
Laten we nu de parameters van het lineair model schatten. In de output van de functie `summary` lezen we $\hat{\beta}_0 = -142.3$, $\hat{\beta}_1 = 0.02737$ en $\hat{\sigma}_\varepsilon = 13.6$. De schatting van de correlatiecoëfficiënt (functie `cor`) is $\hat{\rho} = 0.4017922$. Merk op dat dit waarschijnlijk een onderschatting is omwille van de range restrictie (alle individuen zijn universiteitsstudenten en bijna alle IQs zijn groter dan 100). De aangepaste determinatiecoëfficiënt R^2 is gelijk aan 0.1531. Dit impliceert dat 85% van de variantie van `catell` niet voorspeld wordt door `area`; er is

dus nog veel ruimte voor verbetering van dit model. We zouden bvb. meerdere predictoren kunnen gebruiken.

We hoeven nog na te gaan of de voorwaarden voor de lineaire regressie voldaan zijn. We beginnen met de homoscedasticiteitsassumptie. Laten we kijken naar Fig. 8.11: we vinden geen evidentie dat de voorwaardelijke variantie variëert in functie van `area`. We gaan nu de normaliteitsassumptie na a.d.h.v. een normale qq-plot:

```
> qqnorm(residuals(LM))
```

De output wordt weergegeven in Fig. 8.12. De residuen liggen mooi op een



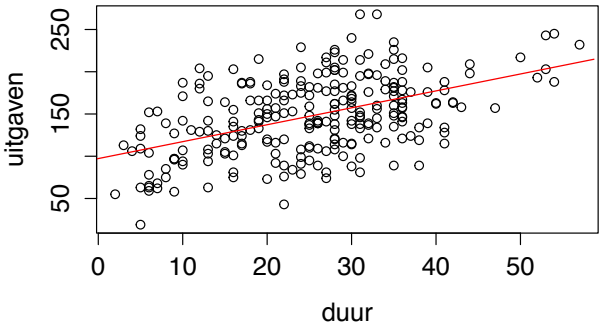
Figuur 8.12: Normale qq-plot van de residuen van het lineair model.

rechte en we kunnen de normaliteitsassumptie veilig aanvaarden.

8.12 Oplossingen

71) Teken de de regressielijn van `uitgaven` op `duur` op Fig. 8.1.

Oplossing:



72) Bereken de coëfficiënten van de regressielijn van `gewicht` op `lengte` m.b.v. R en het data frame `sportData`.

Oplossing:

```
> lm(formula = sportData$gewicht ~ sportData$lengte)

Call:
lm(formula = sportData$gewicht ~ sportData$lengte)

Coefficients:
     (Intercept)  sportData$lengte
       -19.0151            0.5917
```

Dus $b_0 = -19$ en $b_1 = 0.59$.

73) Welke stelling?

Oplossing: De verwachting van een som is de som van de verwachtingen.

74) Welke stelling?

Oplossing: De variantie van een som is gelijk aan de som van de varianties plus twee maal de covariantie.

75) Bereken $\hat{\sigma}^2_\varepsilon$ voor het lineair model met `gewicht` als afhankelijke variabele en `lengte` als predictor (data frame `sportData`).

Oplossing:

```
> sportLM <- lm(formula = sportData$gewicht ~ sportData$lengte)
> n <- length(sportData$gewicht)
> sum(residuals(sportLM)^2)/(n-2)
[1] 194.1736
```

76) Bereken de betrouwbaarheidsintervallen voor β_0 en β_1 voor het lineair model met `gewicht` als afhankelijke variabele en `lengte` als predictor (data frame `sportData`).

Oplossing:

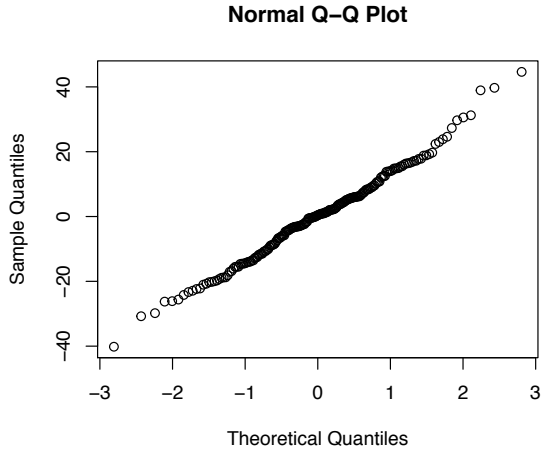
```
> confint(sportLM, level = 0.95)
                   2.5 %    97.5 %
(Intercept)      -39.7039720 1.6737539
sportData$lengte   0.4710621 0.7122637
```

77) Ga na of de fouten normaalverdeeld zijn bij het lineair model met `gewicht` als afhankelijke variabele en `lengte` als predictor (data frame `sportData`).

Oplossing:

```
> qqnorm(residuals(sportLM))
```

Output:



De qq-plot ziet er in orde uit. We mogen aannemen dat de fouten normaal verdeeld zijn.

78) Toets de nulhypothese $\beta_1 = 0$ bij het lineair model met `gewicht` als afhankelijke variabele en `lengte` als predictor (data frame `sportData`).

Oplossing: Met de functie `coef` kunnen we de schattingen van β_0 en β_1 raadplegen.

```
> coef(sportLM)
     (Intercept) sportData$lengte
     -19.0151091        0.5916629
```

We berekenen nu de t-toetsingsgrootheid:

```
> SSR <- sum(residuals(sportLM)^2)
> SSX <- sum( (sportData$lengte - mean(sportData$lengte) )^2 )
> 0.5916629 / sqrt(SSR / ((n-2)*SSX))
[1] 9.67464
```

We hebben de waarde van de t-statistiek gevonden. We berekenen nu de p-waarde.

```
> 2*pt(q=9.67464, df=n-2, lower.tail = FALSE)
[1] 2.246078e-18
```

79) Gebruik de modelselectie aanpak om het lineair model voor `gewicht` met `lengte` als predictor te toetsen, a.d.h.v. `sportData`.

Oplossing:

```
> SSRes0 <- sum((sportData$gewicht - mean(sportData$gewicht))^2)
> SSRes1 <- sum(residuals(sportLM)^2)
> F <- (SSRes0 - SSRes1)/(SSRes1/(n-2))
> pf( q = F, df1 = 1, df2 = n-2, lower.tail = FALSE)
[1] 2.246071e-18
```

De p-waarde is kleiner dan 0.05 en de nulhypothese wordt verworpen. Vergelijk deze p-waarde met de p-waarde van vorige oefening.

80) Gebruik de functie `summary` om het lineair model voor `gewicht` met `lengte` als predictor te analyseren (data frame `sportData`). Vergelijk de output van deze functie met wat je berekend hebt in de vorige oefeningen.

Oplossing:

```
> summary(sportLM)

Call:
lm(formula = sportData$gewicht ~ sportData$lengte)

Residuals:
    Min      1Q  Median      3Q     Max
-40.176  -9.601   0.391   8.316  44.632

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -19.01511   10.49122  -1.812   0.0714 .
```

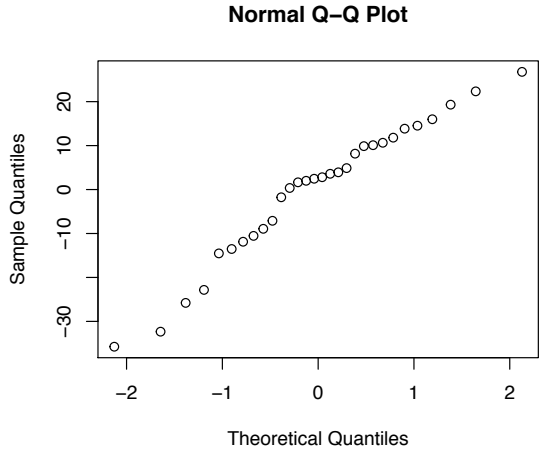
```
sportData$lengte   0.59166    0.06116   9.675   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.93 on 198 degrees of freedom
Multiple R-squared:  0.321,Adjusted R-squared:  0.3176
F-statistic:  93.6 on 1 and 198 DF,  p-value: < 2.2e-16
```

81) Gebruik de functie `summary` om het lineair model voor `iq` met `gewicht` als predictor te analyseren (data frame `myData`). Vergeet niet na te gaan of de fouten normaal verdeeld zijn. Anders mag je het lineair model niet toetsen.

Oplossing: Als het data frame `myData` niet geladen in R is, dan moet je hem eerst opladen.

```
> myLM <- lm(myData$iq~myData$gewicht)
> qqnorm(residuals(myLM))
```

Ouput:



De punten liggen min of meer op de diagonaal.

```
> summary(myLM)

Call:
lm(formula = myData$iq ~ myData$gewicht)

Residuals:
   Min     1Q Median     3Q    Max
-35.76 -10.14   2.63  10.51  26.77

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)    113.12358   15.16830   7.458 4.02e-08 ***
myData$gewicht   0.05797    0.20829   0.278    0.783
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.96 on 28 degrees of freedom
Multiple R-squared:  0.002759,Adjusted R-squared:  -0.03286
F-statistic: 0.07746 on 1 and 28 DF,  p-value: 0.7828
```

De p-waarde voor β_1 is 0.783 en ligt veel hoger dan de drempel (0.05). We aanvaarden de nulhypothese $H_0 : \beta_1 = 0$.

82) Je wil `iq` verklaren m.b.v. de predictor `gewicht`, met een lineair model. Je wil een regressiecoëfficiënt β_1 gelijk aan -0.1 detecteren met een power van 90%. Welke steekproefgrootte heb je nodig?

Oplossing: Je berekent eerst de correlatiecoëfficiënt:

```
> rho <- -0.1*sqrt(var(myData$gewicht))/sqrt(var(myData$iq))
> rho
[1] -0.09060462
```

Nu kan je de steekproefgrootte berekenen:

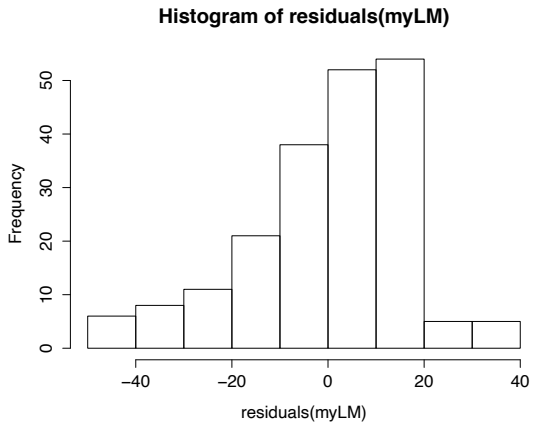
```
> pwr.r.test( r = -0.09060462 , sig.level = 0.05 , power = 0.9)

     approximate correlation power calculation (arctangh transformation)

              n = 1275.035
              r = 0.09060462
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

Je hebt 1275 individuen nodig. Het data frame `myData` is veel te klein.

Waarom heb ik $\beta_1 = -0.1$ gekozen in de opgave? Omdat het overeenkomt met bevindingen in onderzoeken omtrent het verband tussen IQ en BMI, in verscheidene populaties [Kanazawa, 2014]. Met het data frame `myData` wens ik na te gaan om zo'n verband ook aanwezig is in de FPPW populatie.

83) Teken het histogram van de residuen bij vorige oefening. Probeer visueel te begrijpen waarom we beslist hebben dat de fouten niet normaal verdeeld zijn.

Oplossing: Gebruik het commando `hist(residuals(myLM))`. De output is:

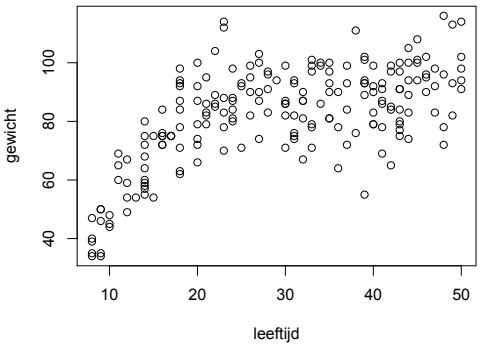


De verdeling van de residuen is zeer asymmetrisch. Dit wijst aan dat de fouten ε_i niet normaal verdeeld zijn.

84) Evalueer de validiteit van de homoscedasticiteitsassumptie met de variabelen `leeftijd` en `gewicht` in het data frame `sportData`.

Oplossing: Gebruik het commando `plot` om het spreidingsdiagram te tekenen.

```
> plot(x = sportData$leeftijd, y = sportData$gewicht)
```



Teken een aantal sneden of probeer je die sneden in te beelden. Het is hier duidelijk dat de geobserveerde voorwaardelijke variantie kleiner is aan de linkerkant. We kunnen dus vermoeden dat de voorwaardelijke populatievariantie niet constant is. We mogen het lineair model niet gebruiken.

Hoofdstuk 9

Meervoudige lineaire regressie

9.1 Inleiding

In Hoofdstuk 8 hebben we het enkelvoudig lineair model gezien. Dit model laat ons toe om het verband tussen een afhankelijke variabele en een predictor te analyseren. Stel nu dat je het verband tussen een afhankelijke variabele en meerdere predictoren wenst te analyseren. Bij het gezondheidsuitgaven-voorbeeld vermoed je bv. dat de variabele `uitgaven` afhankelijk is van `duur` en van `leeftijd`. In het bijzonder vermoed je dat `leeftijd` en `uitgaven` met elkaar positief correleren. Om dit na te gaan kan je natuurlijk twee enkelvoudige lineaire regressies uitvoeren: één voor het verband tussen `uitgaven` en `duur` en één voor het verband tussen `uitgaven` en `leeftijd`. Deze werkwijze heeft een aantal nadelen. Eerst, moet je twee toetsen uitvoeren, elk met een kans α op een fout van de eerste soort. De totale kans op een fout van de eerste soort is dus groter dan α. Een ander nadeel is dat je, met twee afzonderlijke enkelvoudige lineaire regressies, niets weet van het globaal verband tussen de drie variabelen. Nog een nadeel: bij elke toets gebruik je slechts een deel van je data (twee kolommen van de data frame); op geen enkel moment gebruik je alle beschikbare gegevens (drie kolommen van de data set). Je zou meer geloofwaardige beslissingen kunnen maken door alle relevante informatie (de drie kolommen) ineens te gebruiken.

In plaats van twee enkelvoudige lineaire regressies uit te voeren gaan we dus een meervoudig lineair model gebruiken, dat is, een model waar de afhankelijke variabele door meerdere predictoren voorspeld wordt. Bv.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \varepsilon_i$$

of, in ons voorbeeld,

$$\texttt{uitgaven}_i = \beta_0 + \beta_1\, \texttt{duur}_i + \beta_2\, \texttt{leeftijd}_i + \varepsilon_i. \tag{9.1}$$

Dit model is moeilijker om te toetsen dan het enkelvoudig lineair model. De uitkomst van de toets is niet meer binair (het model geldt of niet). Er zijn nu minstens vier mogelijke uitkomsten: (1) het model geldt, met twee predictoren, (2) het model geldt met `duur` als predictor, (3) het model geldt met `leeftijd` als predictor en (4) het model geldt helemaal niet, t.t.z. `duur` en `leeftijd` zijn geen predictoren van `uitgaven`.

We moeten nog een paar stappen overlopen vooraleer we model (9.1) statistisch kunnen toetsen.

9.2 Visuele analyse

9.2.1 Twee predictoren

Vooraleer we een statistische toets gebruiken om het lineair model (9.1) te toetsen, is het belangrijk om de gegevens visueel te analyseren. De data die we zullen gebruiken om dit model te toetsen hebben nu betrekking tot drie variabelen (en niet meer twee). Om de data te visualiseren, hebben we nu een driedimensionale spreidingsdiagram nodig. Hiervoor gebruiken we de functie `scatterplot3d` uit de package `scatterplot3d`.[1]

```
> scatterplot3d(gezondheid$duur,gezondheid$leeftijd,gezondheid$uitgaven)
```

De output van deze functie vind je in Fig. 9.1. De puntenwolk ligt nu in een

Figuur 9.1: 3D spreidingsdiagram.

driedimensionale ruimte en het stijgende verband tussen `duur` en `uitgaven` is nog duidelijk te zien. Op deze grafiek is het minder duidelijk of er een verband is tussen `leeftijd` en `uitgaven`. Om dit beter te kunnen zien, gaan we de grafiek roteren. Er zijn twee technieken om dit te doen: je kan de volgorde van de argumenten van de functie `scatterplot3d` wijzigen of je kan de functie

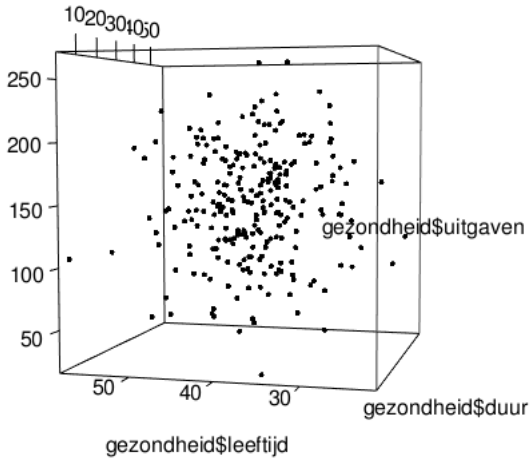
85. Wijzig de volgorde van de argumenten van de functie `scatterplot3d` *om de leeftijd op de horizontale as te krijgen.*

[1] Vergeet niet de package `scatterplot3d` te installeren (behalve als je op Athena werkt) en op te laden.

`plot3d` uit de package `rgl`[2] gebruiken. Deze functie maakt een interactive grafiek aan die je met de muis kan roteren.

```
> plot3d(gezondheid$duur,gezondheid$leeftijd,gezondheid$uitgaven)
```

In Fig. 9.2 vind je een screenshot van de interactieve grafiek in een positie waar het verband tussen `leeftijd` en `uitgaven` best gezien kan worden. Een



Figuur 9.2: geroteerde 3D spreidingsdiagram.

stijgende lineaire tendentie is op deze grafiek niet evident maar kan ook niet uitgesloten worden (we zien geen duidelijke dalende tendentie of curvilineaire tendentie).

Op de verschillende grafieken zien we geen duidelijke uitschieter.

9.2.2 Meer dan twee predictoren

Indien je het verband tussen een afhankelijke variabele en meer dan twee predictoren wil analyseren, dan ligt de puntenwolk in een ruimte met meer dan drie dimensies en hij kan niet meer ineens gevisualiseerd worden. Er is een `R` functie om alle paarsgewijze spreidingsdiagrammen te tekenen: de functie `pairs`. Als we bv. het commando `pairs(myData)` uitvoeren, dan gaat `R` een tabel tekenen met allerlei bidimensionale spreidingsdiagrammen: één per paar variabelen (probeer dit uit). Daar enkele variabelen in het data frame `myData` nominaal zijn (en de corresponderende spreidingsdiagrammen zijn dus niet relevant), gaan we aan `R` een lijst geven van de variabelen waarvoor we een spreidingsdiagram aanvragen: de variabelen 1, 2, 3, 7 en 8.

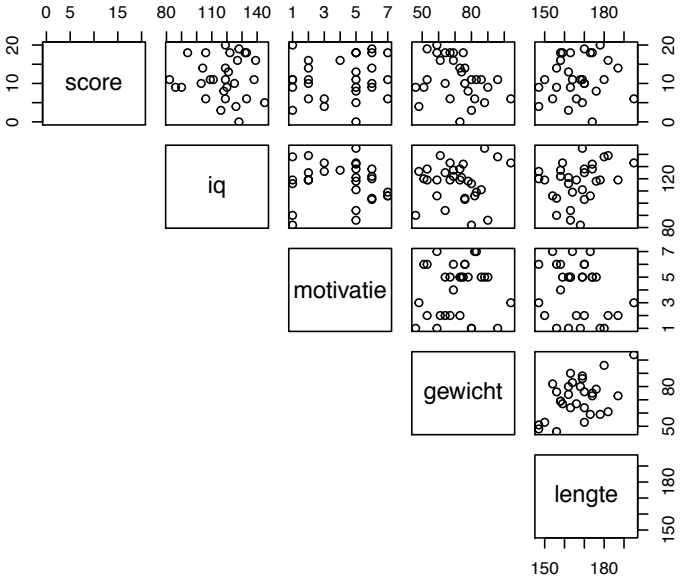
```
> pairs(myData[c(1,2,3,7,8)])
```

[2]De package `rgl` is moeilijk te installeren op Apple en Linux. Heb je een Apple of Linux computer, dan gebruik je best RStudio op Athena om de functie `plot3d` uit de package `rgl` uit te proberen.

Heb je dit commando uitgevoerd? Elke kolom en elke rij van de tabel komt overeen met een bepaalde variabele. Dit wordt aangeduid op de diagonaal. Bv. kolom 2 en rij 2 komen overeen met `iq`. Het spreidingsdiagram in rij 2 en kolom 4 geeft dus het verband weer tussen `iq` en `gewicht`. Merk op dat het spreidingsdiagram in rij 4 en kolom 2 ook het verband tussen `iq` en `gewicht` weergeeft, maar 90 graden gedraaid. Deze tabel bevat dus veel redundante informatie en we gaan de helft van de diagrammen verwijderen.

```
> pairs(myData[c(1,2,3,7,8)], lower.panel = NULL)
```

Je vindt de output van dit commando in Fig. 9.3.



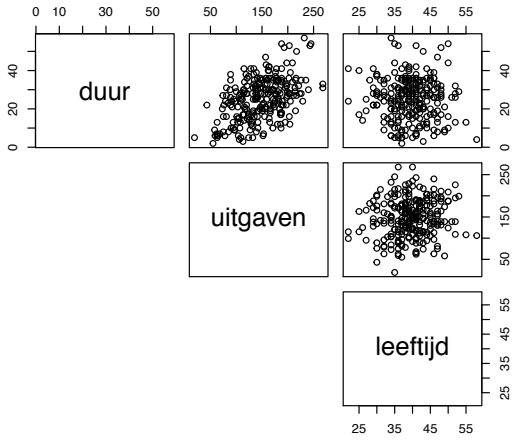
Figuur 9.3: De output van `pairs(myData[c(1,2,3,7,8)],lower.panel = NULL)`.

We kunnen de functie `pairs` gebruiken ook als er slechts twee predictoren zijn. Bv.

```
> pairs(gezondheid[c(2,3,4)], lower.panel = NULL)
```

Je vindt de output van dit commando in Fig. 9.4. Deze figuur bevestigt wat we al met Fig. 9.1 en 9.2 hebben gevonden: er is een duidelijk stijgend lineair verband tussen `duur` en `uitgaven`; zo'n verband tussen `leeftijd` en `uitgaven` is niet evident maar kan ook niet uitgesloten worden. Merk op dat er ook geen duidelijk verband is tussen de predictoren `duur` en `leeftijd`.

86. Gebruik `pairs` *en teken spreidingsdiagrammen voor* `leeftijd`, `gewicht`, `lengte` *en* `tijd` *bij* `sportData`. *Zorg ervoor dat* `tijd` *op de eerste rij staat.*



Figuur 9.4: De output van `pairs(gezondheid[c(2,3,4)], lower.panel = NULL)`.

9.3 Het meervoudig lineair model—Kansrekenen

Het model van belang in dit hoofdstuk is

$$\text{Meervoudig lineair model:} \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i. \quad (9.2)$$

Het is een model met p predictoren: $X_1, \ldots, X_p$. Het coëfficiënt van de predictor X_j is β_j $(j = 1, \ldots, p)$. De toevalsvariabele ε_i representeert nog het toeval: alles wat we niet onder controle houden of alles wat we met onze p predictoren niet verklaren.

9.3.1 Assumpties

Daar het lineair model te veel parameters gebruikt, gaan we nog de Gauss-Markov assumpties maken. Ter herinnering:

1. $E(\varepsilon_i) = 0$ voor alle i. M.a.w. de verwachting van de fout hangt niet af van het individu.

2. $V(\varepsilon_i) = V(\varepsilon_j)$ voor alle i, j. M.a.w. de variantie van de fout hangt niet af van het individu (homoscedasticiteit). Deze constante variantie wordt aangeduid door σ_ε^2.

3. $COV(\varepsilon_i, \varepsilon_j) = 0$ voor alle i, j. M.a.w. de fout bij individu i is niet gecorreleerd met de fout bij individu j (geen seriële correlatie).

Dankzij deze restricties is het aantal parameters fors gereduceerd en model (9.2) wordt bruikbaar.

9.3.2 De voorwaardelijke verwachting

In Rubr. 8.2.2 hebben we de voorwaardelijke verwachting van Y besproken. Omdat het lineair model maar één predictor telde, was de voorwaarde simpel: verwachting van de variabele Y onder voorwaarde dat de unieke predictor X gelijk is aan x. Met het meervoudig lineair model gaat de voorwaarde iets complexer zijn: voor elke predictor gaan we een waarde bepalen.

De voorwaardelijke verwachting van Y onder de hypothese dat het meervoudig lineair model geldt, is dus

$$E(Y_i \,|\, X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip}) = E(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i).$$

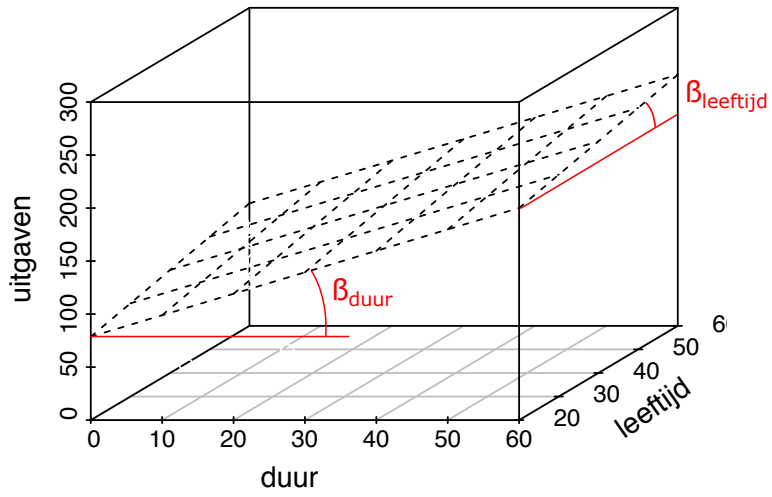
Deze voorwaardelijke verwachting wordt soms korter genoteerd als $E(Y_i \,|\, x_{i1}, \ldots, x_{ip})$. Het is mogelijk te bewijzen dat

$$E(Y_i \,|\, X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}. \qquad (9.3)$$

Merk op dat deze vergelijking deterministisch is; $E(Y_i \,|\, x_{i1}, \ldots, x_{ip})$, β_0, $\beta_1, \ldots, \beta_p$ en $x_{i1}, \ldots, x_{ip}$ zijn allemaal getallen en geen toevalsvariabelen. Dit komt doordat we niet meer focussen op één realisatie van Y_i maar op de verwachting van *alle* realisaties van Y_i. Het toeval speelt dan geen rol meer.

Merk ook op dat de voorwaardelijke verwachting van Y_i een lineaire functie van elke x_{ij} is ($j = 1, \ldots, p$). De coëfficiënt β_j van x_{ij} (die vermenigvuldigt wordt met x_{ij}) heeft een concrete betekenis: als x_{ij} meet één eenheid toeneemt en als de andere variabelen constant blijven, dan stijgt Y_i met β_j eenheden.

Als er slechts twee predictoren zijn, dan kunnen we de functie grafisch representeren in een drie-dimensionale ruimte. We bekomen dan een vlak. Fig. 9.5 presenteert zo'n vlak. Als er meer dan twee predictoren zijn, dan is het niet



Figuur 9.5: De voorwaardelijke verwachting.

mogelijk deze functie (een hypervlak) grafisch te representeren want we hebben meer dan drie dimensies nodig.

We kunnen verg. (9.3) gebruiken om voorspellingen of predicties te maken. Stel dat het volgend model geldt:

$$Y_i = 60 + 2x_{i1} + x_{i2} + \varepsilon_i,$$

met x_{i1} de werkloosheidsduur (in maand), x_{i2} de leeftijd (in jaar) en Y de gezondheidsuitgaven over vier maanden. Als we de werkloosheidsduur x_{i1} en de leeftijd x_{i2} van een individu kennen, kunnen we dan zijn gezondheidsuitgaven voorspellen. We schrijven verg. (9.3) voor de gezondheidsuitgaven, dat is

$$E(Y_i \,|\, X_{i1} = x_{i1}, X_{i2} = x_{i2}) = 60 + 2x_{i1} + x_{i2},$$

we vervangen x_{i1} door de werkloosheidsduur van een individu (bv. 36 maanden) en x_{i2} door de leeftijd van hetzelfde individu (bv. 54 jaar). De voorspelling van zijn gezondheidsuitgaven in de laatste vier maanden is dan $E(Y_i \,|\, X_{i1} = 36, X_{i2} = 54) = 60 + 2 \times 36 + 54 = 186$. Deze voorspelling is uiteraard waarschijnlijk fout. Het is maar een voorspelling maar als we deze formule herhaaldelijk gebruiken om predicties te maken dan zullen onze predicties gemiddeld gezien correct zijn.

De voorwaardelijke verwachting wordt vaak afgekort als $E(Y_i \,|\, x_{i1}, \ldots, x_{ip})$. De formule voor een predictie wordt dan

$$E(Y_i \,|\, x_{i1}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}. \tag{9.4}$$

Het verschil tussen Y_i en de predictie van Y_i is

$$Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}.$$

Dit is een populatie-residu. Als je dit vergelijkt met (9.2), dan vind je

$$Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip} = \varepsilon_i.$$

De foutterm ε_i representeert dus de populatie-residuen en de variantie van ε_i, i.e. σ^2_ε, is dus de variantie van de populatie-residuen.

9.3.3 De voorwaardelijke variantie

We definiëren nu de voorwaardelijke variantie: $V(Y_i \,|\, X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip})$. Het is is de variantie van Y onder voorwaarde dat de predictoren gelijk zijn aan bepaalde waarden. We kunnen ook de voorwaardelijke variantie analyseren onder de hypothese dat het lineair model geldt. In dat geval,

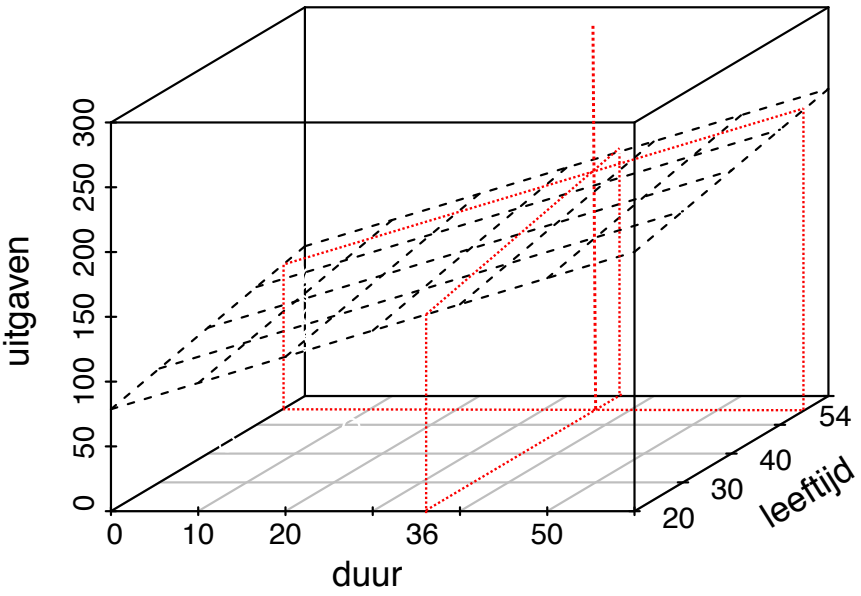
$$V(Y_i \,|\, X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip}) = V(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i).$$

Het is mogelijk te bewijzen dat

$$V(Y_i \,|\, X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip}) = \sigma^2_\varepsilon.$$

De voorwaardelijke variantie van Y_i is dus gelijk aan σ_ε^2; het is onafhankelijk van x_{ij}.

Hoe kunnen we dit interpreteren ? Stel dat we de gezondheidsuitgaven observeren van *alle* individuen die 36 maanden werkloos zijn geweest en die 54 jaar oud zijn. We observeren zeker niet dezelfde uitgaven voor al die individuen : ze variëren. Toch kunnen we uitgaven gelijk aan 186€ voorspellen voor die individuen. De variantie rond de voorspelling wordt verklaard door het toeval en is gelijk aan σ_ε^2. Dit wordt geïllustreerd in Fig.9.6.



Figuur 9.6: De voorspelling (dik rood punt) van Y voor bepaalde waarden van $X_{i1}, \ldots, X_{ip}$ en de spreiding van de realisaties (kleine groene punten) van Y wanneer X_{i1}, ..., X_{ip} vast zijn.

9.3.4 De correlatiecoëfficiënt

In Rubr. 9.2, bij het enkelvoudig lineair model, hebben we gezien dat het verband tussen de correlatiecoëfficiënt en de regressiecoëfficiënt simpel is:

$$\beta_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X}.$$

De reden voor dit simpel verband is dat beide coëfficiënten betrekking hebben tot het verband tussen twee variabelen.

Bij meervoudige lineaire regressie is de toestand complexer. De correlatiecoëfficiënt tussen de afhankelijke variabele Y en predictor j kan nog berekend worden: het is ρ_{YX_j}. Het wordt berekend los van de andere predictoren, met de formule van Rubr. 3.1.9.1. Maar de coëfficiënt β_j dat het verband tussen Y en

X_j kenmerkt is niet los van de andere predictoren: hij representeert het verband tussen Y en X_j binnen het meervoudig lineair model, dus rekening houdend met de andere predictoren. Voor die reden is er geen simpele relatie tussen ρ_{YX_j} en β_j.

9.3.5 Afsluiter

In Rubr. 9.3 hebben we het meervoudig lineair model vanuit een kansrekenenperspectief geanalyseerd. Het heeft betrekking tot toevalsvariabelen (i.t.t. geobserveerde variabelen) in populaties (niet in een specifieke steekproef). Het meervoudig lineair model met p predictoren bevat $p+2$ parameters: $\beta_0, \beta_1, \ldots, \beta_p$ en σ_ε^2 en ze zijn bijna altijd onbekend want de meeste populaties zijn te groot om volledig onderzocht te kunnen worden.

9.4 Puntschatting

Nemen we aan dat het lineair model geldt tussen Y en de predictoren $X_1, \ldots, X_p$, wat zijn dan de waarden van de parameters β_0, $\beta_1, \ldots, \beta_p$ en σ_ε^2? Hoe kunnen we die parameters schatten op basis van een steekproef? Dezelfde methode als bij het enkelvoudig lineair model wordt hier gebruikt: de kleinste kwadraten methode. Gegeven een puntenwolk (een data set) in een $p+1$-dimensionale ruimte, zoeken we naar het vlak (of hypervlak als $p > 2$) dat zo dicht mogelijk bij alle punten ligt: het (hyper)vlak dat de som van de gekwadrateerde residuen (afwijkingen tussen de punten en het hypervlak) minimaliseert (Fig. 9.7 in het geval $p = 2$). `R` heeft een functie om dit te doen: de functie `lm`. Het is dezelfde

Figuur 9.7: Het best passende vlak.

functie als bij het enkelvoudig lineair model. Het argument `formula` is gewoon een beetje anders:

```
> lm(formula = uitgaven ~ duur + leeftijd, data = gezondheid)
```

```
Call:
lm(formula = uitgaven ~ duur + leeftijd, data = gezondheid)

Coefficients:
(Intercept)         duur     leeftijd
    59.2600       2.0234       0.9468
```

Het argument `formula` bestaat nog uit twee delen, gescheiden door een tilde. Aan de linkerkant vind je de afhankelijke variabele (hier `uitgaven`) en aan de rechterkant vind je de lijst van de predictoren, gescheiden van elkaar door een "`+`". Let op, het teken "`+`" representeert hier geen som; het is gewoon een teken om de verschillende predictoren van elkaar te scheiden.

Het argument `data = gezondheid` betekent dat het data frame `gezondheid` moet gebruikt worden. Dankzij dit argument hoeven we niet `gezondheid$` te typen voor de naam van elke variabele.

De output bestaat uit de lijst van alle coëfficiënten van het best passende (hyper)vlak. Deze coëfficiënten zijn gebaseerd op een steekproef en ze worden dus aangeduid door een kleine latijnse letter. Dus $b_0 = 59.26$, $b_{\texttt{duur}} = 2.02$ en $b_{\texttt{leeftijd}} = 0.95$. We zien hieronder dat deze coëfficiënten gebruikt zullen worden om de parameters van het meervoudig lineair model te schatten.

9.4.1 Puntschatting van β_j

Voor elke predictor j is de beste schatter van β_j gewoon B_j; t.t.z. de toevalsvariabele die in elke steekproef gelijk is aan b_j. We mogen dus de steekproefwaarde b_j gebruiken als schatting van β_j in de populatie. De schatter B_j is zuiver ($E(B_j) = \beta_j$) en efficiënt. Bij het gezondheidsvoorbeeld zijn de schattingen $\hat{\beta}_{\texttt{duur}} = 2.02$ en $\hat{\beta}_{\texttt{leeftijd}} = 0.95$.

87. Is de schatting $\hat{\beta}_{\texttt{duur}}$ dezelfde als in Hoofdstuk 8 toen we een enkelvoudig lineair model hebben gebruikt? Waarom?

De formule voor de variantie van B_j is complexer dan bij enkelvoudige lineaire regressie (8.5) en wordt niet gezien. Maar de principes om de variantie van B_j klein te houden en dus om goede schattingen uit te komen, blijven geldig:

- σ_ε^2 moet zo klein mogelijk zijn.
- n moet zo groot mogelijk zijn.
- $s_{X_j}^2$ moet zo groot mogelijk zijn.

9.4.2 Puntschatting van β_0

De beste schatter van β_0 is gewoon B_0; t.t.z. de toevalsvariabele die in elke steekproef gelijk is aan b_0. We mogen dus de steekproefwaarde b_0 gebruiken als schatting van β_0 in de populatie. De schatter B_0 is zuiver ($E(B_0) = \beta_0$) en efficiënt. Bij het gezondheidsvoorbeeld is de schatting $\hat{\beta}_0 = 59.26$.

De formule voor de variantie van B_0 is complexer dan bij enkelvoudige lineaire regressie (8.6) en wordt niet gezien. Maar de principes om de variantie van

B_0 klein te houden en dus om goede schattingen uit te komen, blijven geldig: we zorgen ervoor dat σ_ε^2 zo klein mogelijk is terwijl n en $s_{X_j}^2$ (voor elke predictor) zo groot mogelijk zijn.

9.4.3 De predicties

We hebben gezien (9.4) dat de predicties van het lineair model gegeven worden door

$$E(Y_i \,|\, x_{i1}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

In de praktijk kennen we β_0 en $\beta_1, \ldots, \beta_p$ niet. Om predicties te maken gebruiken we dus de schatters B_0 en $B_1, \ldots, B_p$ i.p.v. de parameters β_0 en $\beta_1, \ldots, \beta_p$, in bovenstaande formule. Het resultaat is niet meer een predictie maar de schatter van een predictie, gedefinieerd door $B_0 + B_1 x_{i1} + \ldots + B_p x_{ip}$ en aangeduid door het symbool $\widehat{Y}_i$. Dus

$$\widehat{Y}_i = B_0 + B_1 x_{i1} + \ldots + B_p x_{ip}.$$

Dit wordt meestal gewoon een predictie genoemd. In een specifieke steekproef kunnen we de realisaties b_0 en $b_1, \ldots, b_p$ van de schatters B_0 en $B_1, \ldots, B_p$ berekenen. We bekomen dan de schatting van $E(Y_i \,|\, x_{i1}, \ldots, x_{ip})$ of ook schatting van de predictie (aangeduid door $\hat{y}_i$)[3]:

$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots, b_p x_{ip}.$$

88. Bereken de predictie $E(\texttt{uitgaven}_i \,|\, \texttt{duur}_i = 30, \texttt{leeftijd}_i = 50)$.

Dit wordt ook meestal gewoon een predictie genoemd. De variantie van de schatter $\widehat{Y}_i$ wordt gegeven door een eenvoudige formule (8.7) bij enkelvoudige lineaire regressie. Bij meervoudige lineaire regressie is de formule niet meer zo simpel en wordt hier niet gezien. Maar de principes om deze variantie klein te houden (en dus om goede predicties te maken) blijven dezelfde. We zorgen ervoor dat σ_ε^2 zo klein mogelijk is terwijl n en $s_{X_1}^2, \ldots, s_{X_p}^2$ zo groot mogelijk zijn. De predictie $E(Y_i \,|\, x_{i1}, \ldots, x_{ip})$ is ook beter (gemiddeld gezien) indien $x_{i1}, \ldots, x_{ip}$ dichtbij $\overline{x}_1, \ldots, \overline{x}_p$ liggen.

Hetzelfde geldt voor predicties voor nog niet geobserveerde waarden.

9.4.4 Puntschatting van σ_ε^2

We hebben gezien dat σ_ε^2 de variantie van de fouten (populatie-residuen) is (Rubr. 9.3.2), i.e., de variantie van

$$Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}.$$

De beste schatter ervan is

$$\begin{aligned} S_\varepsilon^2 &= \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_{i1} - \ldots - B_p x_{ip})^2 \\ &= \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2. \end{aligned}$$

[3] De notatie is misleidend: $\hat{y}_i$ is de schatting van $E(Y_i \,|\, x_{i1}, \ldots, x_{ip})$ en niet van y_i.

Deze schatter is zuiver en efficiënt. De corresponderende schatting is

$$\begin{aligned}\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 &= \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip})^2 \\ &= \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{\mathrm{SS}_{\mathrm{Res}}}{n-p-1}.\end{aligned}$$

Als er maar één predictor is, dan $n - p - 1 = n - 2$. Bovenstaande formule mag dus ook gebruikt worden bij enkelvoudige lineaire regressie (vergelijk met verg. 8.8).

Illustratie We hebben de functie `lm` al gebruikt om de coëfficiënten van het lineair model te berekenen bij het gezondheidsuitgavenvoorbeeld. De output ervan was zeer beperkt alhoewel deze functie heel veel dingen heeft berekend. Om de andere resultaten van de berekeningen te kunnen raadplegen gaan we een naam aan de uitkomst toekennen. Dan gaan we de functie `residuals` gebruiken om $\hat{\sigma}_\varepsilon^2$ te berekenen.

```
> LM <- lm( formula = uitgaven ~ duur + leeftijd, data = gezondheid)
> sum( residuals( LM )^2 ) / 249
[1] 1529.82
```

Bijgevolg $\hat{\sigma}_\varepsilon^2 = 1530$ en $\hat{\sigma}_\varepsilon = \sqrt{1530} = 39.12$.

9.4.5 Collineariteit

Collineariteit treedt op als twee (of meer) predictoren sterk met elkaar correleren. Dit wordt ook multicollineariteit genoemd. Om te begrijpen waarom dit een probleem is, gaan we veronderstellen dat twee predictoren X_1 en X_2 perfect met elkaar correleren ($\rho = 1$). De score van individu i op X_2, d.i. x_{i2}, wordt dus volledig bepaald door zijn score op X_1 (d.i. x_{i1}) want er is een volmaakt lineair verband tussen X_1 en X_2. Dit verband kan geschreven worden als $x_{i2} = \gamma_0 + \gamma_1 x_{i1}$. Stel dat het model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \tag{9.5}$$

geldt. We kunnen dit model herschrijven als

$$\begin{aligned}Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2(\gamma_0 + \gamma_1 x_{i1}) + \beta_3 x_{i3} + \varepsilon_i \\ &= \underbrace{\beta_0 + \beta_2\gamma_0}_{\beta_0'} + \underbrace{(\beta_1 + \beta_2\gamma_1)}_{\beta_1'} x_{i1} + \underbrace{0}_{\beta_2'} x_{i2} + \beta_3 x_{i3} + \varepsilon_i \\ &= \beta_0' + \beta_1' x_{i1} + 0 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.\end{aligned}$$

Dit is een nieuw lineair model (equivalent aan (9.5)), met dezelfde predictoren, maar met andere coëfficiënten. De verschillen zijn niet gering: in dit model is β_2 nul. Het is ook mogelijk dit model te herschrijven zodat $\beta_1 = 0$. En er zijn nog veel andere mogelijkheden. Het is dus onmogelijk om β_1 en β_2 te schatten.

In de praktijk is de correlatiecoëfficiënt tussen twee predictoren nooit gelijk aan 1, maar als hij groot is, dan is het moeilijk om de coëfficiënten β_j te schatten. In functie van het toeval (van de getrokken steekproef) gaat de schatting sterk variëren. De variantie van de schatter gaat groot zijn, wat impliceert dat de schattingen niet bruikbaar gaan zijn. Het lineair model met sterk gecorreleerde predictoren kan dus niet gebruikt worden.

In de praktijk is de correlatiecoëfficiënt tussen twee predictoren ook zelden gelijk aan 0. Het bovenvermelde probleem gaat zich dus altijd voordoen, maar als ρ bijna nul is, dan is het probleem niet ernstig en mag genegeerd worden.

Een andere manier om het colineariteitsprobleem te begrijpen is als volgt. Stel dat de correlatie tussen twee predictoren X_1 en X_2 zwak is (bv. `duur` en `leeftijd` bij het gezondheidsvoorbeeld). Als je het spreidingsdiagram tussen `duur` en `leeftijd` tekent, dan zie je een zeer gespreide puntenwolk. Als je dan `uitgaven` probeert te voorspellen m.b.v. die twee predictoren, dan maak je eigenlijk predicties voor individuen met allerlei combinaties van `duur` en `leeftijd`: bv. lange duur en oud; of kort-oud, lang-jong, kort-jong, matig-jong, enz. Die predicties zijn van redelijke kwaliteit want de steekproef bevat veel individuen met alle combinaties van `duur` en `leeftijd`.

89. Teken nu het spreidingsdiagram tussen `duur` *en* `leeftijd`.

Stel nu dat de correlatie tussen twee predictoren X_1 en X_2 sterk is (bv. `area` en `capacity` bij het `hersenen` dataset). Als je het spreidingsdiagram tussen `area` en `capacity` tekent, dan zie je een nauwe puntenwolk. Als je dan `catell` probeert te voorspellen m.b.v. die twee predictoren, dan maak je predicties voor individuen met allerlei combinaties van `area` en `capacity`: bv. hoog en hoog; of laag-hoog, hoog-laag, matig-hoog, laag-matig, enz. Maar in de steekproef zijn er geen individuen met de combinatie hoog-laag of laag-hoog. De corresponderende predicties zijn dus op niets gebaseerd en zijn bijgevolg van lage kwaliteit.

90. Bereken nu de correlatiecoëfficiënttussen `area` *en* `capacity` *en teken het spreidingsdiagram tussen de twee variabelen.*

Hoe weet je dat collineariteit in zo'n mate optreedt dat je model niet gebruikt kan worden? De package `car` bevat een functie `vif` om de variance inflation factor (VIF) te berekenen, voor elke predictor. De ideale situatie is wanneer elke variance inflation factor gelijk aan 1 is. Dit gebeurt als er geen correlatie is tussen de predictoren. Als sommige predictoren wel met elkaar correleren, dan gaan de VIF(s) van die predictoren toenemen. Als alle VIFs kleiner dan 2 of 3 zijn, dan is er nog geen probleem. Als één of meerdere VIF(s) groter dan 10 zijn (vuistregel), dan is dit een duidelijk teken van collineariteit en het bewuste lineair model mag niet gebruikt worden. Als alle VIFs kleiner dan 10 zijn maar niet veel kleiner dan 10, dan zit je in een grijze zone en je moet voorzichtig zijn. Als je later het model gebruikt om een p-waarde te berekenen dan is er een risico dat de p-waarde niet exact zal zijn.

Toepassing Laten we de correlatiecoëfficiënt berekenen tussen beide predictoren bij het gezondheidsvoorbeeld.

```
> cor(gezondheid$duur,gezondheid$leeftijd)
[1] -0.04367789
```

De correlatiecoëfficiënt is zeer klein en we kunnen dus vermoeden dat er geen collineariteitsprobleem is. Laten we dit verifiëren door de variance inflation factors te berekenen bij het gezondheidsvoorbeeld.

```
> LM <- lm(uitgaven ~ duur + leeftijd, data = gezondheid)
> vif(LM)
    duur leeftijd
1.001911 1.001911
```

Beide VIFs zijn bijna gelijk aan 1 en we mogen dus veilig het lineair model met twee predictoren gebruiken.

91. Je onderzoekt of jobtevredenheid bij bankbedienden verklaard kan worden door leeftijd, loon, extraversion en anciënniteit. Verwacht je een collineariteitsprobleem? Zo ja, tussen welke predictoren?

Hoe los je het probleem op? Als collineariteit optreedt, dan moet je één (of meerdere) predictoren met een groot VIF weglaten. Je hoeft niet altijd de predictor met de grootste VIF weg te laten. Inhoudelijke (niet-statistische) argumenten kunnen ook gebruikt worden om te kiezen welke predictor je wil weglaten.

9.5 Intervalschatting

In deze rubriek veronderstellen we, naast de Gauss-Markov assumpties, dat de fouten normaal verdeeld zijn. De combinatie van deze hypothese met de Gauss-Markov assumpties leidt

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \text{voor alle } i.$$

De betrouwbaarheidsintervallen voor $\beta_0, \beta_1, \ldots, \beta_p$ zijn hier nog gebaseerd op de algemene formule (5.2)

$$\left[\hat{\theta} \pm t_{l;\alpha/2}\ \mathrm{SE}_Q\right].$$

De exacte formule wordt niet gezien omdat $V(B_0), V(B_1), \ldots, V(B_p)$ moeilijker zijn om te berekenen dan bij het enkelvoudig lineair model. Maar herinner je de technieken om de variantie van die schatters klein te houden en dus om smalle betrouwbaarheidsintervallen te bekomen (zie Rubr. 9.4.1 en 9.4.2).

We gebruiken de `R` functie `confint` om de betrouwbaarheidsintervallen voor $\beta_0, \beta_1, \ldots, \beta_p$ te bekomen:

```
> confint( LM, level = 0.95 )
                 2.5 %    97.5 %
(Intercept) 24.6051496 93.914870
duur         1.5796200  2.467184
leeftijd     0.1375096  1.756017
```

De output is vanzelfsprekend.

92. Kunnen we hieruit afleiden dat $\beta_{\texttt{duur}}$ tot het interval $[1.58, 2.47]$ behoort met kans 95%?

9.6 Toetsing

In deze rubriek gaan we ook veronderstellen, naast de Gauss-Markov assumpties, dat de fouten normaal verdeeld zijn. De combinatie van deze hypothese met de Gauss-Markov assumpties leidt

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \text{voor alle } i.$$

Welke hypothese gaan we toetsen? Dat minstens één coëfficiënt β_j van ons meervoudig lineair model niet nul is (er zit wat waarheid in ons model)? Dat alle coëfficiënten β_j van ons lineair model niet nul zijn (ons model is volledig waar)? Dat een bepaalde coëfficiënt (bv. β_2) niet nul is? Al die hypotheses zijn eigenlijk relevant en kunnen statistisch getoetst worden. Een andere onderzoeksvraag is als volgt: welke predictoren (binnen onze set van p predictoren) moeten we selecteren om een optimaal meervoudig lineair model op te bouwen? Met "optimaal" bedoelen we een model dat Y zo goed mogelijk verklaart, maar zonder overbodige predictor.

Voorwaarden. Om de toetsen van dit hoofdstuk te mogen gebruiken, moet de afhankelijke variabele (Y) continu zijn en van interval of ratio meetniveau. Voor discrete afhankelijke variabelen zijn er andere technieken, bv. logistische regressie. Die technieken worden in deze cursus niet gezien.
De onafhankelijke variabelen of predictoren (X_j) moeten van interval of ratio meetniveau zijn of moeten 0-1 zijn. De fouten ε_i moeten normaal verdeeld zijn (dit wordt nagegaan met een normale qq-plot) of de steekproef moet groot zijn.
De Gauss-Markov assumpties moeten ook voldaan zijn.

Als een predictor dichotoom is, maar niet 0-1, dan mag je altijd de codering aanpassen. Bv. 0 voor man en 1 voor vrouw.

9.6.1 De coëfficiënt β_j is nul

In deze rubriek gaan we de nulhypothese "$\beta_j = 0$" toetsen tegen de alternatieve hypothese "$\beta_j \neq 0$". Met andere woorden toetsen we hier of X_j een predictor van Y is, rekening houdend met de andere variabelen in het model. Bv. is `duur` een predictor van `uitgaven`? Dit wordt getoetst a.d.h.v. een t-toets en de corresponderende p-value wordt afgelezen in de output van de functie `summary`.

```
> summary(LM)

Call:
lm(formula = uitgaven ~ duur + leeftijd, data = gezondheid)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-89.178 -29.288  -0.762  27.094 111.931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.2600    17.5954   3.368 0.000877 ***
duur          2.0234     0.2253   8.980  < 2e-16 ***
leeftijd      0.9468     0.4109   2.304 0.022035 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.11 on 249 degrees of freedom
Multiple R-squared:  0.2529,Adjusted R-squared:  0.2469
F-statistic: 42.15 on 2 and 249 DF,  p-value: < 2.2e-16
```

In de rij van `duur` lezen we "`< 2e-16`" af. Deze p-waarde is kleiner dan 0.05 en we besluiten dus dat `duur` een predictor is van `uitgaven`.

93. Schrijf het getal `2e-16` *in de klassieke decimale notatie.*

In welke zin is deze conclusie verschillend van de conclusie die we in Rubr. 8.5.1 getrokken hebben a.d.h.v. een enkelvoudige lineaire regressie? Op het eerste zicht hebben we twee keer hetzelfde getoetst: dat `duur` een predictor van `uitgaven` is. In werkelijkheid hebben we niet echt hetzelfde getoetst. In Rubr. 8.5.1 hebben we getoetst of `duur` een predictor van `uitgaven` is, los van andere variabelen. Maar hier, met het meervoudig lineair model, toetsen we of `duur` een predictor van `uitgaven` is, rekening houdend met `leeftijd`. Wat is het verschil? Het zou kunnen dat individuen met een lange werkloosheidsduur meer geld voor gezondheid spenderen gewoon omdat ze gemiddeld gezien ouder zijn en niet omdat ze langer werkloos zijn. In Rubr. 8.5.1 werd geen rekening gehouden met deze potentiële correlatie tussen `leeftijd` en `duur`. Met meervoudige lineaire regressie wordt deze correlatie wel in acht genomen. En we kunnen nu beweren dat `duur` een predictor van `uitgaven` is, ook na correctie voor verschillen in leeftijd.

We kunnen ook toetsen of `leeftijd` een predictor van `uitgaven` is. We kijken nu naar de p-waarde in de rij van `leeftijd`: we lezen 0.022 af. Dit is kleiner dan 0.05 en we besluiten dat `leeftijd` een predictor van `uitgaven` is, ook na correctie voor verschillen in werkloosheidsduur. In Rubr. 8.5.1 hadden we niet getoetst of `leeftijd` een predictor is van `uitgaven`. Laten we nu het enkelvoudig model met `leeftijd` als enige predictor toetsen.

```
> summary(lm(uitgaven ~  leeftijd, data = gezondheid))

Call:
lm(formula = uitgaven ~ leeftijd, data = gezondheid)

Residuals:
     Min       1Q   Median       3Q      Max
-125.249  -32.211   -0.426   32.967  122.966
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.7524    18.8197   6.204 2.27e-09 ***
leeftijd      0.7856     0.4714   1.667   0.0968 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44.91 on 250 degrees of freedom
Multiple R-squared:  0.01099,Adjusted R-squared:  0.007033
F-statistic: 2.778 on 1 and 250 DF,  p-value: 0.09683
```

De p-waarde is groter dan 0.05. De variabele `leeftijd` is dus geen predictor van `uitgaven` los van andere variabelen, maar wel rekening houdend met `duur`. Dit toont aan dat de conclusie van de toets van $H_0 : \beta_j = 0$ kan variëren in functie van de context (aanwezigheid van andere predictoren).

Nu zit je met volgende vraag: als er twee toetsen zijn om te beslissen of X_j een predictor van Y is en als ze niet altijd leiden tot dezelfde beslissing, welke toets moet je gebruiken? De toets a.d.h.v. het enkelvoudig lineair model of de toets a.d.h.v. het meervoudig lineair model. Het antwoord is niet altijd simpel.

Als je vermoedt dat X_j niet de enige predictor is, dat $X_k, X_l, \ldots$ ook predictoren zijn, en als je over data beschikt m.b.t. die predictoren, dan gebruik je best het meervoudig lineair model. Maar als je geen reden hebt om te denken dat er andere predictoren zijn, dan mag je het enkelvoudig lineair model gebruiken.

9.6.2 De coëfficiënten β_j zijn allemaal nul

In deze rubriek gaan we de nulhypothese "$\beta_1 = \ldots = \beta_p = 0$" toetsen tegen de alternatieve hypothese "minstens één van de coëfficiënten β_j is niet nul". Met andere woorden toetsen we hier of het meervoudig lineair model volledig fout is. Nog met andere woorden gaan we hier het meervoudig lineair model vergelijken met het nulmodel waarbij alle coëfficiënten β_j nul zijn, net zoals in Rubr. 8.5.2.

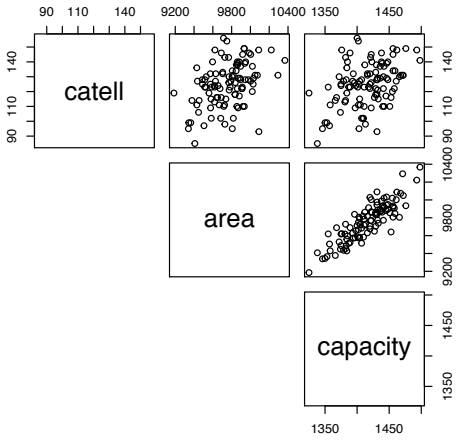
$$\begin{aligned} \text{Nulmodel: } & Y_i = \beta_0 + \varepsilon_i; \\ \text{Model 1: } & Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i. \end{aligned}$$

Dit wordt getoetst a.d.h.v. een F-toets en de corresponderende p-waarde wordt afgelezen in de output van het commando `summary(LM)`. Helemaal onderaan de output (Rubr. 9.6.1) van dit commando lezen we "`< 2.2e-16`" af. De p-waarde is kleiner dan 0.05 en we besluiten dus dat het lineair model met twee predictoren de data beter past dan het nulmodel (zonder predictor). M.a.w. bieden de gegevens genoeg evidentie aan om te besluiten dat model 1 correcter is dan het nulmodel. Minstens één van de coëfficiënten β_j verschilt dus van 0.

Toepassing Laten we de gegevens van het data frame `hersenen` opnieuw analyseren. Daar `area` en `capacity` twee operationaliseringen zijn van de hersengrootte, vermoeden we dat ze allebei predictoren zijn van `catell`. Laten we eerst de gegevens visueel inspecteren.

```
> pairs(hersenen[c("catell", "area",  "capacity")], lower.panel = NULL)
```

Je vindt de output van dit commando in Fig. 9.8. Er blijkt een lineair verband



Figuur 9.8: De output van `pairs(hersenen[c("catell", ärea", "capacity")], lower.panel = NULL)`.

te zijn tussen `area` en `catell` alsook tussen `capacity` en `catell`. Laten we nu een lineair model in `R` aanmaken en analyseren.

```
> LM.iq <- lm(catell ~ area + capacity, data = hersenen)
> summary(LM.iq)

Call:
lm(formula = catell ~ area + capacity, data = hersenen)

Residuals:
    Min      1Q  Median      3Q     Max
-37.358  -9.189   0.127   8.941  33.531

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -147.96291   60.37903  -2.451    0.016 *
area           0.01174    0.01251   0.939    0.350
capacity       0.11172    0.07776   1.437    0.154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.52 on 100 degrees of freedom
Multiple R-squared:  0.1784,Adjusted R-squared:  0.162
F-statistic: 10.86 on 2 and 100 DF,  p-value: 5.408e-05
```

94. Ga na of de normaliteitsassumptie voldaan is.

De p-waarden die corresponderen met de variabelen `area` (0.350) en `capacity` (0.154) zijn allebei groter dan 0.05. Dus geen van de twee is een predictor van `catell`. Vreemd! Dit klopt helemaal niet met onze visuele analyse. Nog vreemder: alhoewel geen van de twee variabelen een predictor is, is de p-waarde van de F-toets (`5.408e-05`) toch veel kleiner dan 0.05. We besluiten dus dat het model (met twee predictoren) geldig is, maar dat geen van de twee variabelen een predictor is. Dit is helemaal tegenstrijdig. Er is misschien een probleem met de normaliteitsassumptie. Of een collineariteitsprobleem? Laten we de variance inflation factor (VIF) berekenen.

```
> vif(LM.iq)
    area capacity
4.103844 4.103844
```

De VIF van beide predictoren is groter dan 4. Dat is niet super hoog maar toch verontrustend. Laten we nu de correlatiecoëfficiënt tussen beide variabelen berekenen.

```
> cor(hersenen$area,hersenen$capacity)
[1] 0.8696701
```

De correlatiecoëfficiënt is inderdaad zeer hoog. Dit kan je ook zien in Fig. 9.8. Laten we even stilstaan en nadenken. Beide predictoren hebben als doel om de hersengrootte te meten. Misschien is één van de twee predictoren overbodig. Laten we dan de analyse herdoen met slechts één predictor (zoals in Rubr. 8.11).

```
> summary( lm(catell ~ area , data = hersenen) )

Call:
lm(formula = catell ~ area, data = hersenen)

Residuals:
    Min      1Q  Median      3Q     Max
-40.867  -8.186  -0.335   9.932  32.450

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.423e+02  6.057e+01  -2.349   0.0208 *
area         2.737e-02  6.207e-03   4.410 2.59e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.6 on 101 degrees of freedom
Multiple R-squared:  0.1614,Adjusted R-squared:  0.1531
F-statistic: 19.44 on 1 and 101 DF,  p-value: 2.592e-05
```

Dit ziet er beter uit: `area` is een predictor van `catell`. Laten we hetzelfde doen met de andere predictor.

```
> summary( lm(catell ~ capacity , data = hersenen))

Call:
lm(formula = catell ~ capacity, data = hersenen)

Residuals:
    Min      1Q  Median      3Q     Max
-34.409  -9.774   0.591   8.935  34.022

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -123.28803   54.32254  -2.270   0.0254 *
capacity       0.17519    0.03836   4.567  1.4e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.52 on 101 degrees of freedom
Multiple R-squared:  0.1712,Adjusted R-squared:  0.163
F-statistic: 20.86 on 1 and 101 DF,  p-value: 1.4e-05
```

En `capacity` is ook een predictor van `catell`. Maar het model met beide predictoren is geen goed model omwille van de sterke correlatie tussen beide variabelen. Welk model is dan best? Het model met `area` of het model met `capacity` als predictor? We zien geen reden om één van de twee uit te sluiten. Beide modellen zijn adequaat om `catell` te verklaren. De determinatiecoëfficiënt R^2 is iets hoger met `capacity` maar het verschil is niet groot.

95. Vergelijk de coëfficiënten $\hat{\beta}_{\texttt{capacity}}$ bij het lineair model met één predictor en het lineair model met twee predictoren. Zijn ze min of meer identiek? Waarom?

9.6.3 Model vergelijking

9.6.3.1 In het algemeen

In Rubr. 8.5.2 hebben we gezien hoe we twee modellen kunnen vergelijken om de beste van de twee te kiezen. Het ging over een lineair model met één predictor en het geneste lineair model zonder predictor (nulmodel). In het algemeen kunnen we dezelfde techniek gebruiken om een lineair model met k predictoren (model A) te vergelijken met een algemener lineair model met p predictoren (model B). Men zegt ook dat model A genest is in model B. "Genest" betekent dat de k predictoren van model A een subset vormen van de p predictoren van model B.

Met elk model kunnen we predicties maken en we kunnen residuen berekenen. We bekomen dus $\mathrm{SS}_{\mathrm{Res}A}$ en $\mathrm{SS}_{\mathrm{Res}B}$. Om de modellen te vergelijken, gaan we het verschil $\mathrm{SS}_{\mathrm{Res}A} - \mathrm{SS}_{\mathrm{Res}B}$ berekenen. Als het groot is, dan hebben we evidentie dat model B beter dan model A is. Als het klein is, dan kan het verschil aan het toeval toegeschreven worden. Om te beslissen of $\mathrm{SS}_{\mathrm{Res}A} - \mathrm{SS}_{\mathrm{Res}B}$ groot is, maken we gebruik van de verhouding (zie (8.10)):

$$\frac{(\mathrm{SS}_{\mathrm{Res}A} - \mathrm{SS}_{\mathrm{Res}B})/(df_A - df_B)}{\mathrm{SS}_{\mathrm{Res}B}\,/df_B}, \tag{9.6}$$

met df_A het aantal vrijheidsgraden van model A (dat is $n-k-1$) en df_B het aantal vrijheidsgraden van model B (dat is $n-p-1$). Het is mogelijk te bewijzen dat, onder "H_0 : model A geldt", deze verhouding F-verdeeld is met $df_A - df_B$ vrijheidsgraden in de teller en df_B vrijheidsgraden in de noemer.

Toepassing We gebruiken nu een nieuw data frame (`geboorte`). Het bevat gegevens m.b.t. 599 geboortes van jongens in 1961–1962 in Oakland, Californië [Nolan and Speed, 2000][4].

```
> head(geboorte)
  gestation  wt parity race age ed ht wt.1 drace dage ded dht dwt
1       284 120      1    8  27  5 62  100     8   31   5  65 110
2       282 113      2    0  33  5 64  135     0   38   5  70 148
3       286 136      4    0  25  2 62   93     3   28   2  64 130
4       245 132      2    7  23  1 65  140     7   23   4  71 192
5       289 120      3    0  25  4 62  125     3   26   1  70 180
6       282 144      4    0  32  2 64  124     0   36   1  74 185
  marital inc smoke time number
1       1   1     0    0      0
2       1   4     0    0      0
3       1   4     2    2      7
4       1   2     0    0      0
5       0   2     0    0      0
6       1   2     1    1      3
```

De variabelen (en hun codering) zijn

- `gestation`: length of gestation (in days)
- `wt`: birth weight (in ounces)
- `parity`: total number of previous pregnancies (including fetal deaths and still births)
- `race`: mother's race: 0-5=white 6=mex 7=black 8=asian 9=mixed
- `age`: mother's age in years at termination of pregnancy
- `ed`: mother's education: 0= less than 8th grade, 1 = 8th -12th grade - did not graduate, 2= HS graduate-no other schooling, 3= HS+trade, 4=HS+some college, 5=College graduate, 6=Trade school, 7=HS unclear
- `ht`: mother's height in inches to the last completed inch
- `wt.1`: mother's prepregnancy weight (in pounds)
- `drace`: father's race (a factor with levels equivalent to mother's race)

[4]Dit is een deelverzameling van het data frame dat beschikbaar is op https://www.stat.berkeley.edu/users/statlabs/

- `dage`: father's age (in years)
- `ded`: father's education (same coding as mother's education)
- `dht`: father's height in inches to the last completed inch
- `dwt`: father's weight (in pounds)
- `marital`: marital status: 1=married, 2=legally separated, 3=divorced, 4=widowed, 5=never married
- `inc`: family yearly income in $2500 increments: 0=under 2500, 1=2500-4999, ..., 8=12,500-14,999, 9=15000+
- `smoke`: does mother smoke? 0=never, 1=smokes now, 2=until current pregnancy, 3=once did, not now
- `time`: time since quitting smoking: 0=never smoked, 1=still smokes, 2=during current preg, 3=within 1 yr, 4=1 to 2 years ago, 5= 2 to 3 yr ago, 6= 3 to 4 yrs ago, 7=5 to 9yrs ago, 8=10+yrs ago, 9=quit and don't know
- `number`: number of cigs smoked per day for past and current smokers (in de oorspronkelijke data set wordt deze variabele gecodeerd d.m.v. een getal tussen 0 en 9).

96. Wat is het meetniveau van de variabele `inc`*? Is het correct gedefinieerd in het data frame* `geboorte`*? En* `smoke`*?*

We willen graag het gewicht van de baby (`wt`) verklaren. Stel dat we weten (uit vroeger onderzoek) dat `gestation`, `parity`, `age`, `wt.1` en `number` predictoren van `wt` zijn. Dit zijn allemaal predictoren die te maken hebben met de moeder. Je vermoedt dat vader-variabelen zoals `dage` en `dwt` ook van belang zijn.[5] Laten we het verband tussen `wt`, `dage` en `dwt` visueel analyseren.

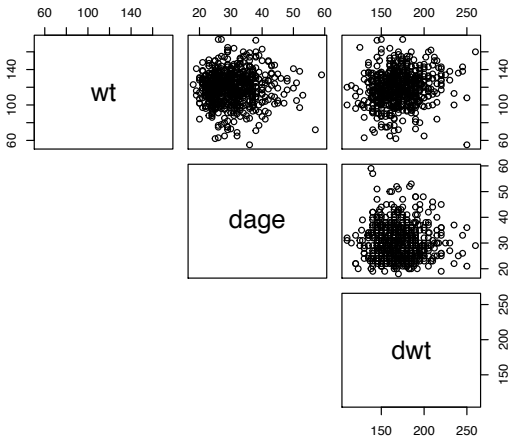
```
> pairs(geboorte[c("wt", "dage", "dwt")], lower.panel = NULL)
```

Je vindt de output van dit commando in Fig. 9.9. Er is wellicht een stijgend lineair verband tussen `wt` en `dwt`. Tussen `wt` en `dage` is er geen duidelijk verband maar ook geen evidentie van een niet lineair verband. We mogen dus meervoudige lineaire modellen gebruiken om de gegevens te analyseren. We beschikken dus over twee modellen.

$$\begin{aligned}\text{model A: } Y_i =& \beta_0 + \beta_{\texttt{gestation}}\,\texttt{gestation}_i + \beta_{\texttt{parity}}\,\texttt{parity}_i + \beta_{\texttt{age}}\,\texttt{age}_i \\ &+ \beta_{\texttt{wt.1}}\,\texttt{wt.1}_i + \beta_{\texttt{number}}\,\texttt{number}_i + \varepsilon_i;\\ \text{Model B: } Y_i =& \beta_0 + \beta_{\texttt{gestation}}\,\texttt{gestation}_i + \beta_{\texttt{parity}}\,\texttt{parity}_i + \beta_{\texttt{age}}\,\texttt{age}_i \\ &+ \beta_{\texttt{wt.1}}\,\texttt{wt.1}_i + \beta_{\texttt{number}}\,\texttt{number}_i + \beta_{\texttt{dage}}\,\texttt{dage}_i + \beta_{\texttt{dwt}}\,\texttt{dwt}_i + \varepsilon_i.\end{aligned}$$

Laten we de twee meervoudige lineaire modellen in `R` aanmaken.

[5] In dit hoofdstuk besteden we geen aandacht aan ordinale of nominale variabelen zoals `sex`, `ed`, enz.



Figuur 9.9: De output van `pairs(geboorte[c("wt", "dage", "dwt")], lower.panel = NULL)`.

```
> LMA <- lm(wt ~ gestation + parity + age + wt.1 + number, data = geboorte)
> LMB <- lm(wt ~ gestation + parity + age + wt.1 + number + dage
  + dwt, data = geboorte)
```

We kunnen ze nu vergelijken met de functie `anova`. De naam "anova" staat voor analysis of variance (variantie-analyse). Deze functie analyseert de varianties (SS_{Res}, SS_{Mod}, SS_{Tot}) van beide modellen en berekent de F-verhouding van verg. 9.6 en de aansluitende p-waarde. De argumenten van deze functie zijn de twee modellen die je met `lm` aanmaakte: eerst het geneste model en dan het algemener model (met meer predictoren).

```
> anova(LMA,LMB)
Analysis of Variance Table

Model 1: wt ~ gestation + parity + age + wt.1 + number
Model 2: wt ~ gestation + parity + age + wt.1 + number + dage + dwt
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    593 161805
2    591 158268  2    3536.4 6.6028 0.001459 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De output van deze functie is als volgt. Bovenaan vind je twee regels met de modellen die vergeleken worden. Dan vind je een anova tabel. Dit is typisch van bijna alle software paketten. Deze tabel bestaat uit twee regels: één per model. Per model heb je het aantal vrijheidsgraden (df_A en df_B) en de kwadratensom van de residuen (SS_{ResA} en SS_{ResB}). Dan, in de laatste regel vind je nog $df_A - df_B (= 2)$, $SS_{ResA} - SS_{ResB} (= 3536.4)$, de F-verhouding ($= 6.6028$) en de p-waarde ($= 0.001459$). Deze is kleiner dan 0.05 en we mogen dus besluiten dat

model B beter is dan model A.

9.6.3.2 Specifieke vergelijkingen

a) De regressiecoëfficiënten β_j zijn allemaal nul. Met de modelvergelijking-aanpak kan je allerlei modellen vergelijken. Je kan bv. een model met een aantal predictoren vergelijken met een model zonder predictor. Dit is wat we gedaan hebben in Rubr. 9.6.2. In die rubriek hebben we de functie `anova` niet gebruikt want het commando `summary(LM)` voert automatisch een F-toets uit om het model `LM` te vergelijken met het nulmodel zonder predictor. Laten we toch die vergelijking nu uitvoeren m.b.v. de functie `anova`. We moeten eerst het nulmodel en het model met twee predictoren aanmaken.

```
> nulM <- lm(formula = uitgaven ~ NULL, data = gezondheid)
> LM <- lm(formula = uitgaven ~ duur + leeftijd, data = gezondheid)
```

Nu kunnen we beide modellen vergelijken.

```
> anova(nulM,LM)
Analysis of Variance Table

Model 1: uitgaven ~ NULL
Model 2: uitgaven ~ duur + leeftijd
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    251 509894
2    249 380925  2    128969 42.151 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In de output zie je dat de F-verhouding en de p-waarde dezelfde zijn bij de output van `summary(LM)`.

b) De coëfficiënt β_j is nul We hebben al gezien hoe je deze hypothese kunt toetsen (Rubr. 9.6.1): m.b.v. een t-toets die automatisch uitgevoerd wordt door de functie `summary`. Eigenlijk kunnen we dit ook doen m.b.v. de modelvergelijking-aanpak. We gaan het lineair model met p predictoren vergelijken met het lineair model met slechts $p-1$ predictoren (zonder predictor X_j). Bv. We willen toetsen of $\beta_{\texttt{leeftijd}}$ een predictor van `uitgaven` is, rekening houdend met `duur`. We gaan twee modellen vergelijken: één met twee predictoren en één met enkel `duur`.

```
> LMA <- lm(formula = uitgaven ~ duur, data = gezondheid)
> LMB <- lm(formula = uitgaven ~ duur + leeftijd, data = gezondheid)
> anova(LMA,LMB)
Analysis of Variance Table

Model 1: uitgaven ~ duur
Model 2: uitgaven ~ duur + leeftijd
```

```
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    250 389048
2    249 380925  1    8122.4 5.3094 0.02203 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In de output zie je dat p-waarde van deze F-toets dezelfde is als de p-waarde van de t-toets in de output van `summary(LM)`.

9.6.4 Selectie van een optimale subset van predictoren

97. Gebruik de functie `anova` *om na te gaan om* `tijd` *beter verklaard wordt door* `leeftijd` *en* `sport` *dan door* `sport` *alleen (data frame* `sportData`*).*

Bij Rubr. 9.6.1–9.6.3 gingen we ervan uit dat je een bepaalde hypothese wil toetsen. Je moet dus over een model beschikken. Maar er zijn situaties waar je niet per se een specifiek model wil toetsen: je beschouwt een aantal variabelen als potentiële predictoren van Y en je wil graag weten welke van die variabelen gebruikt kunnen worden om Y te verklaren. Er zijn drie methodes om dit te doen: Achterwaartse selectie, Voorwaartse selectie en Stapsgewijze selectie.

9.6.4.1 Achterwaartse selectie

Ook achterwaartse eliminatie genoemd. Deze methode werkt stapsgewijs. Je start met een meervoudig lineair model met alle potentiële predictoren erin en je gaat één predictor uitsluiten. Als het resulterend model de data goed verklaart, dan sluit je een andere predictor uit, enz. Preciezer:

1. Start met alle predictoren in het model.
2. Verwijder de predictor waarvoor de p-waarde het grootst is en groter is dan α.
3. Beschouw het model zonder de verwijderde predictor(en) en ga opnieuw naar stap 2.
4. Stop wanneer alle p-waarden kleiner zijn dan α.

Bij elke stap gebruik je een t-toets om een predictor uit te sluiten. Bij elke stap is de kans op een fout van de eerste soort gelijk aan α. De kans op minstens één fout van de eerste soort over alle stappen is dus groter dan α. Daarom gebruik je hier best een significantie α kleiner dan 0.05. Er is geen regel om deze waarde te kiezen.

Toepassing Laten we het gezondheidsvoorbeeld gebruiken. We voeren eerst een meervoudige lineaire regressie met alle predictoren (`duur` en `leeftijd`). We hebben maar twee predictoren en de procedure zal dus maximaal uit twee stappen bestaan. We gaan dus niet veel t-toetsen uitvoeren en we gebruiken $\alpha = 0.03$.

```
> summary(LM)

Call:
lm(formula = uitgaven ~ duur + leeftijd, data = gezondheid)

Residuals:
    Min      1Q  Median      3Q     Max
-89.178 -29.288  -0.762  27.094 111.931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.2600    17.5954   3.368 0.000877 ***
duur          2.0234     0.2253   8.980  < 2e-16 ***
leeftijd      0.9468     0.4109   2.304 0.022035 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.11 on 249 degrees of freedom
Multiple R-squared:  0.2529,Adjusted R-squared:  0.2469
F-statistic: 42.15 on 2 and 249 DF,  p-value: < 2.2e-16
```

Er is geen predictor waarvoor de p-waarde groter dan α is. De procedure stopt dus na de eerste stap zonder dat we een predictor kunnen uitsluiten. Het finaal model is dus het model met twee predictoren.

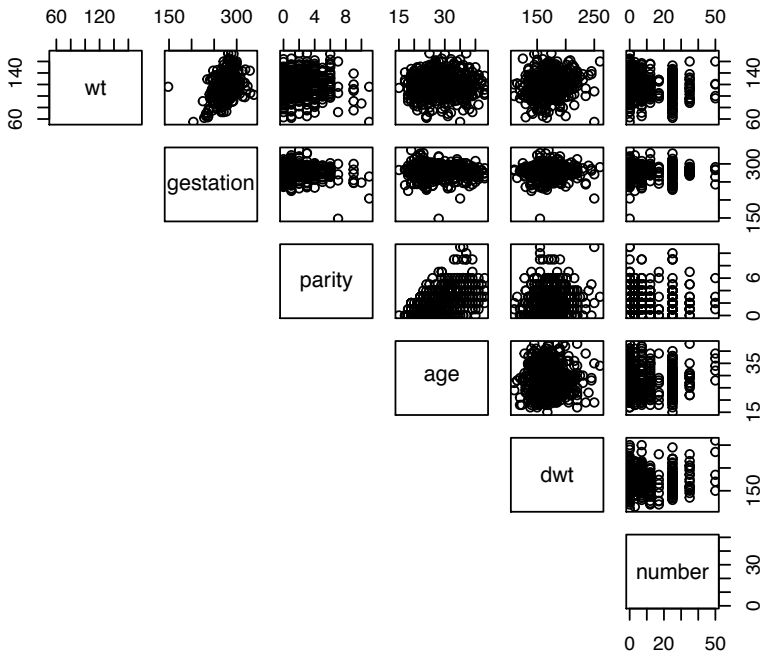
Toepassing Laten we opnieuw het geboorte-voorbeeld analyseren. In tegenstelling tot Rubr. 9.6.3 veronderstellen we hier niet dat we beschikken over een model uit vroeger onderzoek. We zijn geïnteresseerd in het gewicht van de baby (`wt`) en we vermoeden dat het mogelijks verklaard wordt door `gestation`, `parity`, `age`, `dwt` en `number`. Laten we de data visueel analyseren.

```
> pairs(geboorte[c("wt", "gestation", "parity", "age", "dwt",
  "number")], lower.panel = NULL)
```

Je vindt de output van dit commando in Fig. 9.10. We zien een duidelijk lineair stijgend verband tussen `wt` en `gestation`. Dit is geen verrassing: hoe langer de zwangerschap, hoe zwaarder de baby. We zien een (iets minder) duidelijk lineair stijgend verband tussen `parity` en `age`. Dit is ook geen verrassing: vrouwen die al meerdere keer zwanger zijn geweest, zijn meestal ouder. Voor de rest zien we geen duidelijk verband maar ook geen evidentie van een niet lineair verband. We mogen dus een meervoudige lineaire regressie gebruiken om de gegevens te analyseren.

We starten de achterwaartse eliminatie met een model met alle potentiële predictoren. Daar we veel predictoren hebben en we veel t-toetsen zullen uitvoeren, gebruiken we best een lage significantie. Bv. $\alpha = 0.01$.

```
> summary(lm(wt ~ gestation+parity+age+dwt+number, data=geboorte))
```



Figuur 9.10: De output van `pairs(geboorte[c("wt", "gestation", "parity", äge", "dwt", "number")], lower.panel = NULL)`.

```
Call:
lm(formula = wt ~ gestation + parity + age + dwt + number,
data = geboorte)

Residuals:
    Min      1Q  Median      3Q     Max
-49.977 -10.910  -0.664   9.992  54.682

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.38136   13.32494  -2.430   0.0154 *
gestation     0.46047    0.04252  10.830  < 2e-16 ***
parity        0.37189    0.41985   0.886   0.3761
age           0.12980    0.13727   0.946   0.3448
dwt           0.12460    0.02958   4.213 2.91e-05 ***
number       -0.29926    0.06192  -4.833 1.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 16.44 on 593 degrees of freedom
Multiple R-squared:  0.2166,Adjusted R-squared:   0.21
F-statistic: 32.79 on 5 and 593 DF,  p-value: < 2.2e-16
```

De variabele `parity` heeft de grootste p-waarde (0.3761) en deze is groter dan 0.01. We voeren dezelfde analyse zonder `parity`.

```
> summary(lm(wt ~ gestation + age + dwt + number, data = geboorte))

Call:
lm(formula = wt ~ gestation + age + dwt + number, data = geboorte)

Residuals:
    Min      1Q  Median      3Q     Max
-50.484 -10.653  -0.791   9.992  55.992

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.61253   13.31997  -2.448   0.0146 *
gestation     0.45614    0.04223  10.802  < 2e-16 ***
age           0.19473    0.11604   1.678   0.0938 .
dwt           0.12683    0.02946   4.305 1.95e-05 ***
number       -0.29817    0.06189  -4.818 1.85e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.43 on 594 degrees of freedom
Multiple R-squared:  0.2155,Adjusted R-squared:  0.2103
F-statistic:  40.8 on 4 and 594 DF,  p-value: < 2.2e-16
```

De variabele `age` heeft de grootste p-waarde (0.0938) en deze is groter dan 0.01. We voeren dezelfde analyse zonder `age`.

```
> summary(lm(wt ~ gestation+dwt+number, data=geboorte))

Call:
lm(formula = wt ~ gestation + dwt + number, data = geboorte)

Residuals:
    Min      1Q  Median      3Q     Max
-50.213 -10.439  -0.934   9.805  55.574

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.23421   12.78567  -2.052   0.0406 *
gestation     0.45218    0.04223  10.708  < 2e-16 ***
dwt           0.12734    0.02951   4.316 1.86e-05 ***
```

```
number       -0.29694    0.06198  -4.791 2.10e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.46 on 595 degrees of freedom
Multiple R-squared:  0.2118,Adjusted R-squared:  0.2079
F-statistic:  53.3 on 3 and 595 DF,  p-value: < 2.2e-16
```

98. Wat is de concrete interpretatie van $\hat{\beta}_{\texttt{dwt}} = 0.12734$?

Alle resterende predictoren hebben een p-waarde kleiner dan α en de achterwaartse eliminatie stopt hier. Het model

$$\texttt{wt}_i = \beta_0 + \beta_{\texttt{gestation}}\,\texttt{gestation}_i + \beta_{\texttt{dwt}}\,\texttt{dwt}_i + \beta_{\texttt{number}}\,\texttt{number}_i + \varepsilon_i$$

is ons finaal model.

9.6.4.2 Voorwaartse selectie

Deze methode werkt ook stapsgewijs. Je start met het nulmodel (zonder predictor) en je gaat één predictor toevoegen. Als het resulterend model beter is, dan voeg je een andere predictor toe, enz. Preciezer:

1. Start met geen enkele predictor in het model.
2. Voer een enkelvoudige regressie-analyse uit voor elke predictor. Dit betekent dat een set van regressie-analyses uitgevoerd wordt waarbij de uitkomst op elke predictor afzonderlijk geregresseerd wordt. Stop de predictor met de kleinste p-waarde die kleiner is dan α in het model. Stop indien geen enkele predictor in het model kan opgenomen worden.
3. Voer een set van meervoudige regressie-analyses uit voor alle predictoren die niet in het model zitten door ze telkens afzonderlijk in het reeds opgebouwde regressiemodel te stoppen. Voeg de predictor toe met de kleinste p-waarde die kleiner is dan α.
4. Herhaal stap 3 tot geen enkele predictor in het model toegevoegd kan worden (d.i. kleinste p-waarde groter dan α).

Met deze procedure moeten we meer regressies uitvoeren dan met achterwaartse eliminatie. We zijn dus nog strenger bij de keuze van het significantie niveau α.

Toepassing Laten we het gezondheidsvoorbeeld gebruiken. We voeren twee enkelvoudige lineaire regressies, elke met één predictor (we hanteren $\alpha = 0.01$):

```
> summary(lm(uitgaven ~ duur, data = gezondheid))

Call:
lm(formula = uitgaven ~ duur, data = gezondheid)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-98.220 -27.461  -1.725  26.538 108.774

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   97.204      6.252  15.547   <2e-16 ***
duur           2.001      0.227   8.812   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.45 on 250 degrees of freedom
Multiple R-squared:  0.237,Adjusted R-squared:  0.234
F-statistic: 77.66 on 1 and 250 DF,  p-value: < 2.2e-16
```

en

```
> summary(lm(uitgaven ~ leeftijd, data = gezondheid))

Call:
lm(formula = uitgaven ~ leeftijd, data = gezondheid)

Residuals:
     Min       1Q   Median       3Q      Max
-125.249  -32.211   -0.426   32.967  122.966

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.7524    18.8197   6.204 2.27e-09 ***
leeftijd      0.7856     0.4714   1.667   0.0968 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44.91 on 250 degrees of freedom
Multiple R-squared:  0.01099,Adjusted R-squared:  0.007033
F-statistic: 2.778 on 1 and 250 DF,  p-value: 0.09683
```

De kleinste p-waarde ($< 2e-16$) komt met `duur` overeen en ze is kleiner dan 0.01. Deze variabele wordt opgenomen in het model. We voeren nog één lineaire regressie waarbij de predictor `leeftijd` toevoegen.

```
> summary(lm(uitgaven ~ duur + leeftijd, data = gezondheid))

Call:
lm(formula = uitgaven ~ duur + leeftijd, data = gezondheid)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-89.178 -29.288  -0.762  27.094 111.931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.2600    17.5954   3.368 0.000877 ***
duur          2.0234     0.2253   8.980  < 2e-16 ***
leeftijd      0.9468     0.4109   2.304 0.022035 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.11 on 249 degrees of freedom
Multiple R-squared:  0.2529,Adjusted R-squared:  0.2469
F-statistic: 42.15 on 2 and 249 DF,  p-value: < 2.2e-16
```

De p-waarde die correspondeert met `leeftijd` is groter dan 0.01 en deze variabele wordt niet behouden in het finaal model. Het finaal model heeft dus slechts één predictor, in tegenstelling tot het model dat we bekomen hebben met achterwaartse eliminatie. Dit illustreert het feit dat de twee methodes niet equivalent zijn.

99. Selecteer een set van predictoren voor `wt` *in het data frame* `geboorte`*, a.d.h.v. voorwaartse selectie. De ptentiële predictoren zijn* `gestation`*,* `parity`*,* `age`*,* `dwt` *en* `number`*.*

Toepassing Laten we de data in `sportData` opnieuw analyseren. We willen een set van predictoren voor `tijd` selecteren. De lijst van potentiële predictoren is `sport`, `gewicht` en `lengte`. De variabele `leeftijd` wordt niet in deze lijst opgenomen omdat we hebben gezien dat het verband tussen `tijd` en `leeftijd` niet lineair is (zie Fig. 2.7). De variabelen `type` en `geslacht` worden ook niet opgenomen omdat ze nominaal zijn.

Laten we $\alpha = 0.03$ hanteren. Bij de eerste stap van de voorwaartse selectie moeten we drie enkelvoudige lineaire regressies uitvoeren. We beginnen met `sport` (de volgorde is niet belangrijk).

```
> summary(lm(tijd ~ sport, data = sportData))

Call:
lm(formula = tijd ~ sport, data = sportData)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4545  -2.8544  -0.3049   3.1706  10.3457

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.7553     0.8763  20.261   <2e-16 ***
sport         1.7997     0.2861   6.289    2e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.297 on 198 degrees of freedom
Multiple R-squared:  0.1665,Adjusted R-squared:  0.1623
F-statistic: 39.56 on 1 and 198 DF,  p-value: 1.999e-09
```

We gaan verder met `gewicht`.

```
> summary(lm(tijd ~ gewicht, data = sportData))

Call:
lm(formula = tijd ~ gewicht, data = sportData)

Residuals:
    Min      1Q  Median      3Q     Max
-11.663  -3.314  -0.349   2.705  12.167

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.59998    1.64504   15.56   <2e-16 ***
gewicht     -0.03261    0.01964   -1.66   0.0985 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.674 on 198 degrees of freedom
Multiple R-squared:  0.01373,Adjusted R-squared:  0.008744
F-statistic: 2.755 on 1 and 198 DF,  p-value: 0.09851
```

En nu `lengte`.

```
> summary(lm(tijd ~ lengte, data = sportData))

Call:
lm(formula = tijd ~ lengte, data = sportData)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9164  -3.2787  -0.2812   2.5859  12.1411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.64869    3.48848   9.072   <2e-16 ***
lengte      -0.05108    0.02034  -2.512   0.0128 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.633 on 198 degrees of freedom
Multiple R-squared:  0.03088,Adjusted R-squared:  0.02599
F-statistic: 6.309 on 1 and 198 DF,  p-value: 0.01281
```

De kleinste van de drie p-waarden is die van `sport` (2e-09). Ze is kleiner dan 0.03 en de variabele `sport` wordt dus in het finaal model opgenomen. We voeren nu twee lineaire regressies met twee predictoren. We beginnen met `sport` en `gewicht`.

```
> summary(lm(tijd ~ sport + gewicht, data = sportData))

Call:
lm(formula = tijd ~ sport + gewicht, data = sportData)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1482  -3.0359  -0.3171   2.7132  10.7519

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.85665    1.66848  12.500  < 2e-16 ***
sport        1.83387    0.28392   6.459 8.05e-10 ***
gewicht     -0.03900    0.01792  -2.177   0.0307 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.257 on 197 degrees of freedom
Multiple R-squared:  0.1861,Adjusted R-squared:  0.1778
F-statistic: 22.52 on 2 and 197 DF,  p-value: 1.555e-09
```

We gaan verder met `sport` en `lengte`.

```
> summary(lm(tijd ~ sport + lengte, data = sportData))

Call:
lm(formula = tijd ~ sport + lengte, data = sportData)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3692  -2.9836  -0.0708   2.9273  10.0960

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.37710    3.35471   7.565 1.45e-12 ***
sport        1.75865    0.28347   6.204 3.18e-09 ***
lengte      -0.04394    0.01868  -2.352   0.0197 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.249 on 197 degrees of freedom
Multiple R-squared:  0.1893,Adjusted R-squared:  0.181
```

```
F-statistic:    23 on 2 and 197 DF,  p-value: 1.056e-09
```

De p-waarden die overeenkomen met `lengte` en `gewicht` zijn 0.0197 en 0.0307. De kleinste van de twee is die van `lengte` (0.0197). Ze is kleiner dan 0.03 en de variabele `lengte` wordt dus in het finaal model opgenomen (samen met `sport`). We voeren nu één lineaire regressie met drie predictoren.

```
> summary(lm(tijd ~ sport + lengte + gewicht, data = sportData))

Call:
lm(formula = tijd ~ sport + lengte + gewicht, data = sportData)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2074  -2.9596  -0.2436   2.8231  10.3868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.84107    3.39536   7.316 6.39e-12 ***
sport        1.79051    0.28516   6.279 2.15e-09 ***
lengte      -0.03066    0.02277  -1.347    0.180
gewicht     -0.02221    0.02180  -1.019    0.309
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.248 on 196 degrees of freedom
Multiple R-squared:  0.1936,Adjusted R-squared:  0.1812
F-statistic: 15.68 on 3 and 196 DF,  p-value: 3.519e-09
```

De p-waarde van `gewicht` is groter dan 0.03 en deze variabele wordt dus niet opgenomen in het final model. Het finaal model is

$$\texttt{tijd}_i = \beta_0 + \beta_{\texttt{sport}}\,\texttt{sport}_i + \beta_{\texttt{lengte}}\,\texttt{lengte}_i + \varepsilon_i.$$

9.6.4.3 Stapsgewijze selectie

Bij voorwaartse selectie kan het gebeuren dat de p-waarde van een predictor die geselecteerd werd bij een bepaalde stap, plots niet meer kleiner dan α is bij een later stap (zoals bij het sportvoorbeeld in vorige rubriek. Maar de predictor wordt toch behouden in het model. Stapsgewijze selectie heeft een oplossing hiervoor. We starten zoals bij voorwaartse selectie maar bij iedere stap kunnen we beslissen om opgenomen predictoren die niet langer significant zijn opnieuw uit het model te verwijderen. De berekeningen zijn lang en worden niet geïllustreerd.

9.6.4.4 Opmerkingen

De drie methodes die we net gezien hebben, hebben allemaal hetzelfde nadeel: daar we veel toetsen uitvoeren, telkens met significantie α, is de kans op minstens

één fout van de eerste soort groter dan α. Hoeveel groter? Dat is moeilijk om in te schatten. Bijgevolg controleren we de kans op de fout van de eerste soort niet meer. We proberen dit probleem op te lossen door een kleine α te gebruiken, maar we hebben geen regel om α te bepalen.

Om de validiteit van deze technieken de verhogen, werk je best niet met kleine steekproeven. Een vuistregel zegt dat n/p best groter dan 40 moet zijn, waarbij n de steekproefgrootte is en p het aantal predictoren. In onze beide voorbeelden (gezondheid en geboorte) geldt deze vuistregel.

De beste attitude om de nadelen van die technieken te compenseren bestaat in het gebruik van kruisvalidatie. Stel dat je bv. voorwaartse selectie hebt gebruikt en je hebt een model bekomen, maar je vertrouwt het niet omwille van het bovenvermelde probleem. Je beschikt dus over een model maar je trekt het in twijfel. De evidente aanpak (nu dat je zo veel statistiek hebt gestudeerd) is om een nieuwe steekproef te trekken en een statistische toets (meervoudige lineaire regressie) uit te voeren. Maar deze keer, omdat je over een model beschikt, gebruik je geen van de drie laatste technieken maar wel de F-toets zoals in Rubr. 9.6.2. Deze techniek (de validatie van een model op basis van een nieuwe steekproef) heet kruisvalidatie (cross validation).

De drie selectietechnieken hebben als doel om een optimale subset van predictoren te selecteren. Niet te klein en niet te groot. Je kan waarschijnlijk raden wat "te klein" betekent. Dat is een set waarin belangrijke predictor(en) afwezig zouden zijn. Wat betekent dan "te groot"? Dat is een set waarin alle belangrijke predictoren aanwezig zijn, maar ook andere variabelen die eigenlijk geen predictor zijn of die bijna niet helpen om Y te verklaren. Is dat een probleem indien ons lineair model overbodige predictoren bevat? Zo'n model is flexibeler dan een model met minder predictoren en kan dus de puntenwolk beter passen.

Ja, maar de overbodige predictoren hebben ook nadelen. Ze leiden tot extra flexibiliteit en dus tot bijkomende parameters. De schattingen van die parameters gaan geen structurele kenmerk van het geobserveerde fenomeen weergeven; de schattingen gaan gewoon random zijn. Je gaat dus een complex (met veel predictoren) model bekomen waarvan een groot deel niets te maken heeft met het bewuste fenomeen, maar wel met het toeval. Dat is uiteraard niet de bedoeling van wetenschappelijk onderzoek.

Een tweede gevolg van te veel predictoren is dat het risico op collineariteit (zie Rubr. 9.4.5) dan hoger is. Dit is ook niet wenselijk.

9.7 De determinatiecoëfficiënt R^2

In Rubr. 8.6, bij enkelvoudige lineaire regressie hebben we gezien dat de variantie s_Y^2 $(= \mathrm{SS}_{\mathrm{Tot}}/(n-1))$ opgesplitst kan worden in twee delen: de verklaarde variantie $(\mathrm{SS}_{\mathrm{Mod}}/(n-1))$ en de onverklaarde variantie $(\mathrm{SS}_{\mathrm{Res}}/(n-1))$. Dit leidde tot de definitie van de determinatiecoëfficiënt $R^2 = \mathrm{SS}_{\mathrm{Mod}}/\mathrm{SS}_{\mathrm{Tot}}$: de proportie van verklaarde variantie.

Daar $\mathrm{SS}_{\mathrm{Tot}}, \mathrm{SS}_{\mathrm{Mod}}$ en $\mathrm{SS}_{\mathrm{Res}}$ gedefinieerd zijn in termen van $y_i, \hat{y}_i$ en $\overline{y}$ (en niet van x_i) zijn deze definities onafhankelijk van het aantal predictoren. De

definities van Rubr. 8.6 blijven dus geldig bij meervoudige lineaire regressie. De interpretatie is ook dezelfde. Vb. In de output van het commando

```
> summary(lm(uitgaven ~ duur + leeftijd, data = gezondheid))
```

lezen we" `Adjusted R-squared:  0.2469`" af. Dit betekent dat 25% van de variantie van `uitgaven` verklaard wordt door `duur` en `leeftijd`. Laten we deze proportie vergelijken met de proportie die we in Rubr. 8.7 hebben berekend (p. 157). Het was 23%. De toevoeging van de predictor `leeftijd` heeft $\overline{R}^2$ verhoogd met slechts 2%. Alhoewel `leeftijd` een predictor van `uitgaven` is, is het dus geen belangrijke predictor.

9.8 De power van meervoudige lineaire regressie

Eigenlijk mogen we niet van de power van meervoudige lineaire regressie spreken, maar wel van de power van een specifieke toets. Bv. de toets m.b.t. een specifieke regressiecoëfficiënt β_j (t-toets); of de toets m.b.t. alle regressiecoëfficiënten van een model (F-toets); of nog een andere toets. Eigenlijk zijn al die toetsen bijzondere gevallen van de algemene F-toets van de model vergelijking en we gaan dus dit eerst bespreken, vooraleer we bijzondere gevallen zien.

100. Bij Rubr. 9.6.4.2 hebben we voorwaartse selectie gebruikt om een set van predictoren te selecteren voor `tijd` *(data frame* `sportdata`*. Bij één van de stappen hebben we* `lengte` *opgenomen in het model. Welke bijkomende proportie variantie werd verklaard door het opnemen van* `lengte`*?*

9.8.1 Model vergelijking in het algemeen

Je wenst twee modellen te vergelijken: model A met k predictoren en model B met p predictoren, waarbij de predictoren van model A een subset vormen van de predictoren van model B. De nulhypothese is "H_0 : model A geldt" en de alternatieve hypothese is "H_a : model A geldt niet maar B wel." Om deze hypothese te toetsen gebruik je een F-toets met $p-k$ vrijheidsgraden in de teller en $n-p-1$ vrijheidsgraden in de noemer (8.10). Om de power van deze toets te berekenen gebruik je de functie `pwr.f2.test` (package `pwr`). Deze functie heeft drie argumenten nodig: het aantal vrijheidsgraden in de teller (`u`), het aantal vrijheidsgraden in de noemer (`v`) en de effectgrootte f^2 (`f2`). De definitie van de effectgrootte is

$$f^2 = \frac{R_B^2 - R_A^2}{1 - R_B^2},$$

waarbij R_A^2 en R_B^2 de determinatiecoëfficiënten (zie Rubr. 9.7) van modellen A en B zijn. De effectgrootte f^2, zoals de anderen, heeft geen duidelijke betekenis. Ze kan variëren tussen 0 en $+\infty$. Uitspraken zoals "$f^2 = 0.15$ representeert een matige effectgrootte" zijn dus zinloos. We hoeven dus f^2 te berekenen in functie van R_A^2 en R_B^2 om iets zinvol uit te komen. Jammer genoeg is het ook niet simpel om R_A^2 en R_B^2 te interpreteren. Waarden uit vroeger onderzoek of uit een pilootonderzoek kunnen eventueel overgenomen worden.

9.8.2 Alle regressiecoëfficiënten zijn nul

Deze hypothese toetsen komt erop neer dat je een model A zonder predictor vergelijkt met een model B met p predictoren (zie Rubr. 9.6.3.2.a). Je wil dus nagaan of model B in zijn geheel goed is; je wil weten of model B het toelaat om goede predicties te maken.

Daar model A geen predictor heeft, kan het niets verklaren. Dus $R^2_A = 0$. Om f^2 te berekenen, hoeven we dus slechts R^2_B te bepalen. Welke waarde van R^2_B wens je te detecteren? Misschien heb je kennis van een onderzoek waar hetzelfde model in een andere populatie geanalyseerd werd. Of een onderzoek waar een gelijkaardig model bij dezelfde populatie geanalyseerd werd. Dan kan je eventueel de proportie verklaarde variantie van dat onderzoek gebruiken om f^2 te berekenen.

Koopgedrag Je wil een model met 4 predictoren ontwikkelen om het koopgedrag van consumenten te voorspellen. Welk budget (Y) gaat individu i aan product P besteden indien haar scores op de vier predictoren $x_{i1}, \ldots, x_{i4}$ zijn? In de literatuur heb je een gelijkaardig model gevonden omtrent product Q in Vlaanderen, met $R^2 = 0.25$. Omdat producten P en Q vergelijkbaar zijn en omdat jouw onderzoek ook over Vlaanderen gaat, neem je die waarde over. Nu kan je f^2 berekenen: $f^2 = 0.25/(1-0.25) = 0.33$. En we gebruiken nu `pwr.f2.test` met `u` $= p - k = 4 - 0 = 4$ en `power` $= 0.9$.

```
> pwr.f2.test(u=4, f2=0.333, power = 0.9)

     Multiple regression power calculation

              u = 4
              v = 46.25834
             f2 = 0.333
      sig.level = 0.05
          power = 0.9
```

We vinden `v` $= n - p - 1 = n - 4 - 1 \approx 47$. Dus $n \approx 47 + 4 + 1 = 52$. Bijgevolg, indien je een steekproef van 52 individuen trekt en indien het model B met 4 predictoren 25% van de variantie van Y kan verklaren, dan zal je dit detecteren (het nulmodel verwerpen) met kans 90%.

101. Wat is de power indien $n = 30$?

102. Is de nodige steekproefgrootte groter of kleiner als $R^2 = 0.35$? Probeer eerst logisch te redeneren, zonder `R` te gebruiken.

9.8.3 De regressiecoëfficiënt β_j is nul

Deze hypothese toetsen komt erop neer dat je een model A met $p-1$ predictoren vergelijkt met een model B met p predictoren (zie Rubr. 9.6.3.2.b). Daar model B één extra predictor telt, kan het zeker meer variantie verklaren dan model A. Maar, als het verschil klein is, kan het toevallig zijn. We gaan du na of het verschil $R^2_B - R^2_A$ groter is dan wat toevallig plausibel is.

Bij deze toets hoeven modellen A en B geen uitstekende modellen te zijn, in de zin dat R^2_A en R^2_B relatief laag kunnen zijn. Wat belangrijk is, is het verschil

tussen R^2_A en R^2_B: is de toename te wijten aan de toevoeging van predictor j? Of is het toevallig?

Om f^2 te berekenen, hoeven we R^2_A en R^2_B te bepalen. Model A is misschien een bestaand model dat al empirisch onderzocht werd en je kent dus misschien R^2_A. Of je gaat misschien de proportie verklaarde variantie van een gelijkaardig model gebruiken als proxy voor R^2_A.

Voor R^2_B kan je waarschijnlijk niet rekenen op vroeger onderzoek, maar je gaat een realistische toename ($R^2_B - R^2_A$) beschouwen.

Koopgedrag 2 Je beschikt over een model met 4 predictoren om het koopgedrag van consumenten te voorspellen. Dit is model A, met 25% verklaarde variantie. Je wil nagaan of X_5 ook een predictor is van Y. Elke van de vier predictoren X_1 t.e.m. X_4 verklaren (gemiddeld gezien) $0.25/4 = 6.25\%$ van de variantie van Y. Je vermoedt dat X_5 een zwakkere predictor is, maar niet super zwak (anders wil je hem niet in model B opnemen). Je wenst dus een toets uit te voeren om 3% toename in verklaarde variantie te kunnen detecteren.

Nu kan je f^2 berekenen: $f^2 = 0.03/(1 - 0.28) \approx 0.042$. En we gebruiken nu `pwr.f2.test` met `u` $= p - k = 5 - 4 = 1$ en `power` $= 0.9$.

```
> pwr.f2.test(u=1, f2=0.042, power = 0.9)

     Multiple regression power calculation

              u = 1
              v = 250.1129
             f2 = 0.042
      sig.level = 0.05
          power = 0.9
```

We vinden `v` $= n - p - 1 = n - 5 - 1 \approx 250$. Dus $n \approx 250 + 5 + 1 = 256$. Bijgevolg, indien je een steekproef van 256 individuen trekt en indien het model B met 5 predictoren 3% van de variantie van Y kan verklaren boven wat model A al verklaart, dan zal je dit detecteren (model A verwerpen) met kans 90%.

9.9 Controle van modelassumpties: de functie `plot`

De assumpties van meervoudige lineaire regressie zijn dezelfde als die van enkelvoudige lineaire regressie: de Gauss-Markov assumpties en de normaliteitsassumptie[6].

Stel dat je een lineaire regressie met `R` hebt uitgevoerd en dat de resultaten in het object `myLM` gestopt zijn. Het commando `plot(myLM)` gaat vier diagrammen tekenen die ons helpen om de modelassumpties te checken. Telkens als je

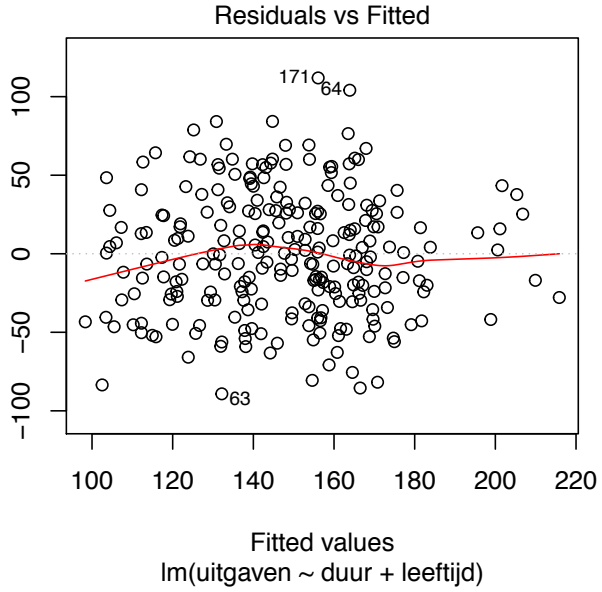
[6] Normaliteit van de residuen, niet van de predictoren of van de afhankelijke variabele

`Enter` of `Return` drukt, verschijnt een nieuw diagram. We gaan ze één per één beschrijven a.d.h.v. het gezondheidsvoorbeeld.

```
> LM <- lm(uitgaven ~ duur + leeftijd, data = gezondheid)
> plot(LM)
Hit <Return> to see next plot:
```

9.9.1 `Residuals vs fitted` — Gauss-Markov 1

Het eerste diagram is de "`Residuals vs fitted`" plot (Fig. 9.11). Op de hori-



Figuur 9.11: `Residuals vs fitted` plot — gezondheidsvoorbeeld

zontale as vind je de predicties (Engels: fitted) en op de verticale as, de residuen. Dit diagram wordt gebruikt i.p.v. het klassieke spreidingsdiagram[7] om de residuen te analyseren. Dit diagram wordt niet in detail uitgelegd; we zien wel hoe het gebruikt wordt.

Op Fig. 9.11 vind je een rode curve. Elke punt op deze curve representeert de schatting van de voorwaardelijke verwachting van ε_i. Elke punt op deze rode curve is het gemiddelde van de corresponderende verticale snede. De eerste Gauss-Markov assumptie stelt dat $E(\varepsilon_i) = 0$. Dit impliceert dat de voorwaardelijke verwachting van de residuen nul is. De rode curve moet dus min of meer horizontaal zijn, op hoogte 0 (aangeduid door de horizontale stippellijn). Op Fig. 9.11 zien we geen duidelijke afwijking t.o.v. de stippellijn. We mogen dus de eerste Gauss-Markov assumptie aanvaarden.

[7] Het klassieke spreidingsdiagram kan niet getekend worden bij meervoudige lineaire regressie indien het aantal predictoren groter dan 2 is.

Op Fig. 9.11 vind je ook een paar punten met een getal ernaast. Dit zijn punten die door `R` als outliers of speciale punten geïdentificeerd worden. Het getal naast het punt is het nummer van de corresponderende rij in het data frame. Het is aangeraden om die punten afzonderlijk te bekijken.

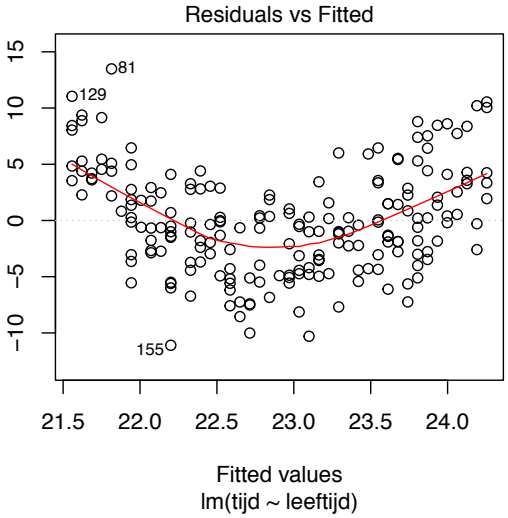
```
> gezondheid[c(63,64,171),]
    geslacht duur uitgaven leeftijd
63         M   22       43       30
64         V   33      268       40
171        V   31      268       36
```

Je kan verifiëren of er geen tikfout is of een ander probleem.

Schending van de 1ste Gauss-Markov assumptie Op Fig. 2.7 hebben we gezien dat het verband tussen `tijd` en `leeftijd` in het data frame `sportData` curvilineair is. Laten we dit analyseren met de functie `lm`.

```
> LM.tijd.leeftijd <- lm(tijd ~ leeftijd, data = sportData)
> plot(LM.tijd.leeftijd)
Hit <Return> to see next plot:
```

De eerste plot wordt in Fig. 9.12 weergegeven. We zien dat de schattingen van



Figuur 9.12: `Residuals vs fitted` plot — sportData

de voorwaardelijke verwachtingen helemaal niet constant zijn: de rode curve heeft min of meer de vorm van een parabool. De eerste Gauss-Markov assumptie wordt dus waarschijnlijk niet voldaan. Dit wijst aan dat het verband tussen de twee variabelen niet lineair is (zie ook Fig. 8.9).

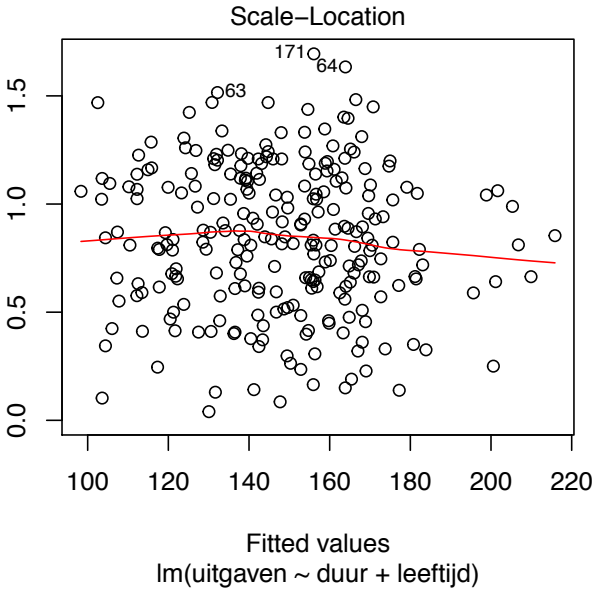
103. Teken de `Residuals vs fitted` *plot voor het lineair model dat het gewicht van de baby verklaart m.b.v. van* `gestation`, `parity`, `age`, `wt.1`, `number`, `dage` *en* `dwt`. *Is de eerste Gauss-Markov assumptie in orde?*

9.9.2 `Normal Q-Q` — Normaliteit

Het tweede diagram dat door het commando `plot(LM)` wordt getekend, is de normale qq-plot. We hebben dit diagram al vaak besproken. We maken toch een opmerking. Als je dit diagram met het commando `plot(LM)` tekent (en niet met `qqnorm(residuals(LM))`), dan worden opnieuw een paar punten door `R` geïdentificeerd als outliers. Het is aangeraden om die punten afzonderlijk te bekijken.

9.9.3 `Scale-Location` — Homoscedasticiteit

Het derde diagram dat door het commando `plot(LM)` wordt getekend, is de "`Scale-Location`" plot (Fig. 9.13). Dit diagram wordt niet in detail uitgelegd; we zien wel hoe het gebruikt wordt. Op Fig. 9.13 vind je een rode curve.

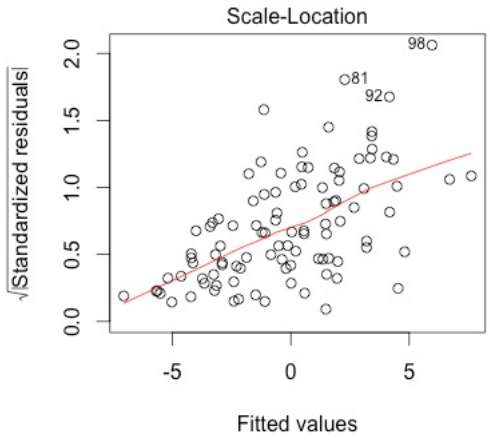


Figuur 9.13: `Scale-Location` plot — gezondheidsvoorbeeld

Elke punt op deze curve representeert de schatting van de vierkantswortel uit de voorwaardelijke variantie van Y. De tweede Gauss-Markov assumptie stelt dat de voorwaardelijke variantie van de residuen constant is. De rode curve moet dus min of meer horizontaal zijn. Op Fig. 9.13 zien we dat de rode curve min of meer horizontaal is. We mogen dus de tweede Gauss-Markov assumptie aanvaarden.

Op Fig. 9.13 vind je ook een paar punten met een getal ernaast. Dit zijn punten die door `R` als outliers geïdentificeerd worden. Het is aangeraden om die punten afzonderlijk te bekijken.

Schending van de homoscedasticiteitsassumptie Fig. 9.14 geeft de derde plot ("`Scale-Location`" plot) weer van een fictief voorbeeld. We zien dat de



Figuur 9.14: `Scale-Location` plot — fictief voorbeeld

schattingen van de voorwaardelijke variantie helemaal niet constant zijn: de rode curve stijgt. De tweede Gauss-Markov assumptie wordt dus waarschijnlijk niet voldaan.

104. Teken de `Scale-Location` *plot voor het lineair model dat het gewicht van de baby verklaart m.b.v. van* `gestation`, `parity`, `age`, `wt.1`, `number`, `dage` *en* `dwt`. *Is de tweede Gauss-Markov assumptie in orde?*

9.9.4 `Residuals vs Leverage` — Invloedrijke punten

Het vierde diagram dat door het commando `plot(LM)` wordt getekend, is de "`Residuals vs Leverage`" plot (Fig. 9.13). Dit diagram wordt niet gezien.

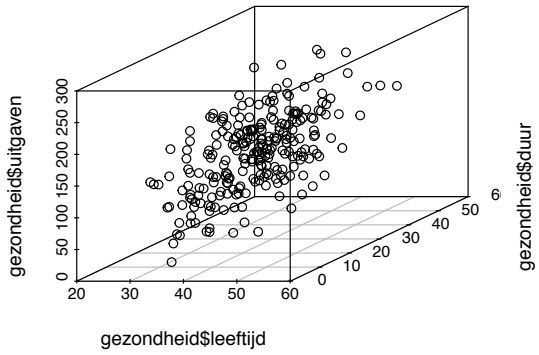
9.10 Oplossingen

85) Wijzig de volgorde van de argumenten van de functie `scatterplot3d` om de leeftijd op de horizontale as te krijgen.

Oplossing:

```
> scatterplot3d(gezondheid$leeftijd,gezondheid$duur,gezondheid$uitgaven)
```

Output:

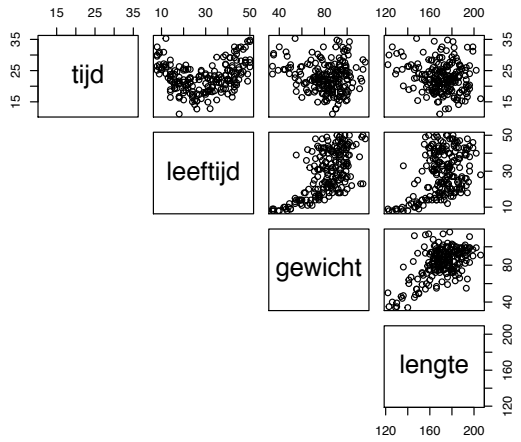


86) Gebruik `pairs` en teken spreidingsdiagrammen voor `leeftijd`, `gewicht`, `lengte` en `tijd` bij `sportData`. Zorg ervoor dat `tijd` op de eerste rij staat.

Oplossing:

```
> pairs(sportData[c(4,1,2,3)], lower.panel = NULL)
```

Output:



87) Is de schatting $\hat{\beta}_{\texttt{duur}}$ dezelfde als in Hoofdstuk 8 toen we een enkelvoudig lineair model hebben gebruikt? Waarom?

Oplossing: De schatting $\hat{\beta}_{\texttt{duur}}$ was 2.001 in Hoofdstuk 8. Nu is het 2.02. Het is verschillend omdat $\hat{\beta}_{\texttt{duur}}$ nu berekend wordt rekening houdend met `leeftijd`.

88) Bereken de predictie $E(\texttt{uitgaven}_i \,|\, \texttt{duur}_i = 30, \texttt{leeftijd}_i = 50)$.

Oplossing:

$$\begin{aligned} E(\texttt{uitgaven}_i \,|\, \texttt{duur}_i = 30, \texttt{leeftijd}_i = 50) &= 59.26 + 2.0234 \times 30 + 0.9468 \times 50 \\ &= 167.302. \end{aligned}$$

89) Teken nu het spreidingsdiagram tussen `duur` en `leeftijd`.

Oplossing: `plot(gezondheid$duur,gezondheid$leeftijd)`

90) Bereken nu de correlatiecoëfficiënt tussen `area` en `capacity` en teken het spreidingsdiagram tussen de twee variabelen.

Oplossing:

```
> cor(hersenen$capacity,hersenen$area)
[1] 0.8696701
> plot(hersenen$capacity,hersenen$area)
```

91) Je onderzoekt of jobtevredenheid bij bankbedienden verklaard kan worden door leeftijd, loon, extraversion en anciënniteit. Verwacht je een collineariteitsrobleem? Zo ja, tussen welke predictoren?

Oplossing: Alle predictoren behalve extraversion gaan waarschijnlijk sterk met elkaar correleren want oudere bedienden hebben vaak meer anciënniteit en hebben ook vaak hogere lonen. Collineariteit gaat dus optreden tussen leeftijd, loon en anciënniteit.

92) Kunnen we hieruit afleiden dat $\beta_{\texttt{duur}}$ tot het interval $[1.58, 2.47]$ behoort met kans 95%?

Oplossing: Neen. Dit is een verkeerde interpretatie van het betrouwbaarheidsinterval. Ga terug naar Hoofdstuk 5 om de juiste interpretatie van het betrouwbaarheidsinterval te begrijpen.

93) Schrijf het getal `2e-16` in de klassieke decimale notatie.

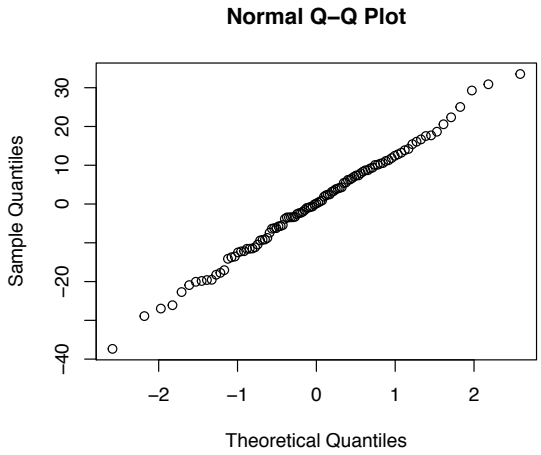
Oplossing: Het is gelijk aan 2/10000000000000000 (met 16 nullen). Het is dus 0.0000000000000002 (15 nullen na de comma of 16 nullen in totaal). Onduidelijk? Surf dan naar https://nl.wikipedia.org/wiki/Wetenschappelijke_notatie.

94) Ga na of de normaliteitsassumptie voldaan is.

Oplossing:

```
> qqnorm(residuals(LM.iq))
```

Output:



De normaliteitsassumptie is in orde.

95) Vergelijk de coëfficiënten $\hat{\beta}_{\texttt{capacity}}$ bij het lineair model met één predictor en het lineair model met twee predictoren. Zijn ze min of meer identiek? Waarom?

Oplossing: $\hat{\beta}_{\texttt{capacity}} = 0.17519$ bij het lineair model met één predictor terwijl $\hat{\beta}_{\texttt{capacity}} = 0.11172$ bij het lineair model met twee predictoren. Ze zijn sterk verschillend. Hiervoor zijn er minstens twee redenen.

- 0.17519 is de schatting van $\beta_{\texttt{capacity}}$ zonder rekening te houden met `area`.
- Daar de twee predictoren van het meervoudig lineair model sterk correleren, is de variantie van de schatter $B_{\texttt{capacity}}$ zeer hoog. Deze schatter levert dus vaak slechte schattingen en het is dus mogelijk dat 0.11172 een slechte schatting is.

96) Wat is het meetniveau van de variabele `inc`? Is het correct gedefinieerd in het data frame `geboorte`? En `smoke`?

Oplossing: Het meetniveau van `inc` is ordinaal. Typ `geboorte$inc`. Helemaal onderaan de output vind je `Levels: 0 < 1 < 2 < 3 < 4 < 5 < 6 < 7 < 8 < 9`. Dit is in orde.

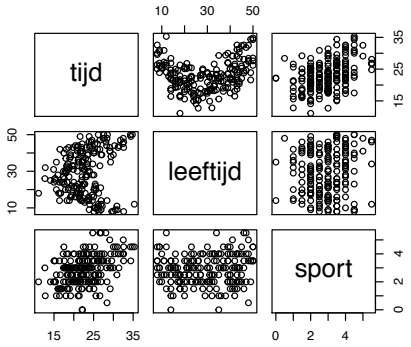
Het meetniveau van `smoke` is nominaal. Typ `geboorte$smoke`. Helemaal onderaan de output vind je `Levels: 0 1 2 3`. Dit is ook in orde.

97) Gebruik de functie `anova` om na te gaan om `tijd` beter verklaard wordt door `leeftijd` en `sport` dan door `sport` alleen.

Oplossing: We analyseren eerst de paarsgewijze spreidingsdiagrammen:

```
> pairs(sportData[c("tijd","leeftijd", "sport")])
```

Output:



We zien een duidelijke curvilineair verband tussen `tijd` en `leeftijd`. We mogen dus geen lineaire regressie gebruiken.

98) Wat is de concrete interpretatie van $\hat{\beta}_{\texttt{dwt}} = 0.12734$?

Oplossing: Als we twee geboortes vergelijken waarbij `gestation` en `number` identiek zijn en waarbij `dwt` met één kilo verschillen, dan kunnen we voorspellen dat het gewicht van de baby van de zwaardere vader hoger zal zijn met 127 gram.

99) Selecteer een set van predictoren voor `wt` in het data frame `geboorte`, a.d.h.v. voorwaartse selectie.

Oplossing: Bij de eerste stap van de voorwaartse selectie moeten we vijf enkelvoudige lineaire regressies uitvoeren. Bij de tweede stap moeten we vier meervoudige lineaire regressies uitvoeren. Enz. De berekeningen zijn te lang en worden hier niet getoond. Bij de eerste stap wordt `gestation` geselecteerd. Bij de tweede stap wordt `number` geselecteerd. Bij de derde stap wordt `dwt` geselecteerd en de p-waarden van alle niet-geselecteerde variabelen zijn groter dan 0.01. Het finaal model is dus hetzelfde als bij achterwaartse eliminatie.

100) Bij Rubr. 9.6.4.2 hebben we voorwaartse selectie gebruikt om een set van predictoren te selecteren voor `tijd` (data frame `sportdata`. Bij één van de stappen hebben we `lengte` opgenomen in het model. Welke bijkomende proportie variantie werd verklaard door het opnemen van `lengte`?

Oplossing: Bij het opnemen van `lengte` was `sport` al in het model. Met `sport` alleen is $\overline{R}^2$ gelijk aan 0.1623. Met `sport` en `lengte` is $\overline{R}^2$ gelijk aan 0.181. Het verschil is$0.181 - 0.1623 = 0.0187$. Slechts 1.87% van de variantie van `tijd` wordt dus verklaard door het opnemen van `lengte` (boven `sport`) in het model.

101) Wat is de power indien $n = 30$?

Oplossing: We berekenen eerst `v`.

$$\texttt{v} = n - p - 1 = 30 - 4 - 1 = 25.$$

En nu de power:

```
> pwr.f2.test(u=4, v=25, f2=0.333)

     Multiple regression power calculation

              u = 4
              v = 25
             f2 = 0.333
      sig.level = 0.05
          power = 0.6225809
```

De power is 62%. Nauwelijks beter dan de worp van een muntstuk.

102) Is de nodige steekproefgrootte groter of kleiner als $R^2 = 0.35$? Probeer eerst logisch te redeneren, zonder `R` te gebruiken.

Oplossing: Als $R^2 = 0.35$ i.p.v. 0.25, dan weet je dat de verbanden tussen de variabelen sterker zijn en het is dus gemakkelijker om ze te detecteren. Je mag dus met een kleinere steekproef werken. We verifiëren dit met `R`.

```
> pwr.f2.test(u=4, f2=0.35/(1-0.35), power = 0.9)

     Multiple regression power calculation

              u = 4
              v = 28.74507
             f2 = 0.5384615
      sig.level = 0.05
          power = 0.9
```

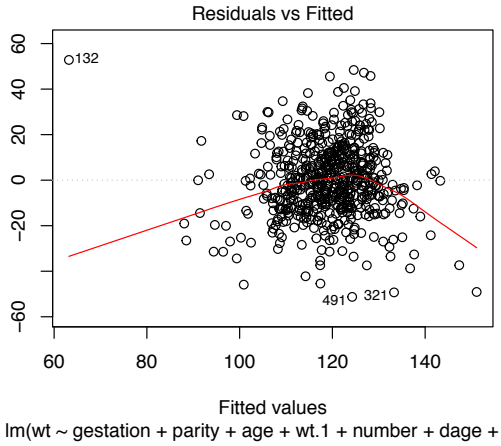
We vinden $\mathtt{v} = n - p - 1 = n - 4 - 1 \approx 29$. Dus $n \approx 29 + 4 + 1 = 34$. Dit is inderdaad kleiner dan 52.

103) Teken de `Residuals vs fitted` plot voor het lineair model dat het gewicht van de baby verklaart m.b.v. van `gestation`, `parity`, `age`, `wt.1`, `number`, `dage` en `dwt`. Is de eerste Gauss-Markov assumptie in orde?

Oplossing: We maken eerst een lineair model aan in `R` en dan gebruiken we de functie `plot`.

```
> LM <- lm(wt ~ gestation + parity + age + wt.1 + number + dage + dwt,
  data = geboorte)
> plot(LM)
Hit <Return> to see next plot:
```

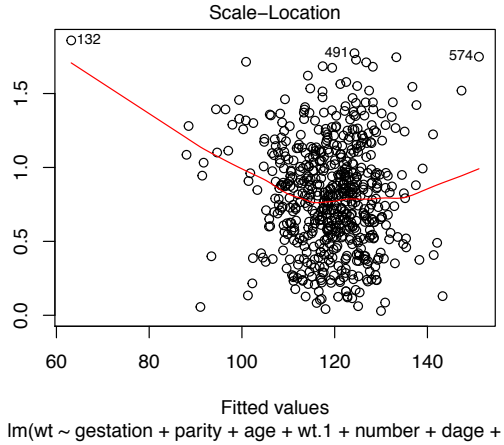
Ouput:



Interpretatie: de rode curve is helemaal niet horizontaal. Er blijkt een probleem te zijn met de eerste Gauss-Markov assumptie. Misschien een lineariteitsprobleem. Elke conclusie die gebaseerd is op dit model moet dus voorzichtig geïnterpreteerd worden.

104) Teken de `Scale-Location` plot voor het lineair model dat het gewicht van de baby verklaart m.b.v. van `gestation`, `parity`, `age`, `wt.1`, `number`, `dage` en `dwt`. Is de tweede Gauss-Markov assumptie in orde?

Oplossing: He hebben het lineair model al aangemaakt, bij vorige oefening. We hoeven due nog ééns op `Return` te drukken en we komen deze grafiek uit:



Interpretatie: de rode curve is helemaal niet horizontaal. Er blijkt een probleem te zijn met de homoscedasticiteitsassumptie. Elke conclusie die gebaseerd is op dit model moet dus voorzichtig geïnterpreteerd worden.

Hoofdstuk 10

Lineaire regressie met nominale predictoren

In Hoofdstuk 9 (p. 186) hebben we gezien dat lineaire regressie gebruikt mag worden met afhankelijke variabelen (Y) van interval of ratio meetniveau en met predictoren (X_j) van interval of ratio meetniveau of dichotoom (0-1). Stricto sensu is dit correct, maar er bestaat eigenlijk een truc om lineaire regressie toch te kunnen gebruiken met nominale predictoren. Deze truc gaan we in dit hoofdstuk zien.

Daarvoor gaan we in Rubr. 10.1 het gebruik van lineaire regressie met dichotome[1] predictoren illustreren en bespreken. In Rubr. 10.2 zullen we dan een techniek zien om lineaire regressie ook met nominale variabelen met meer dan twee niveau's te gebruiken.

10.1 Lineaire regressie met dichotome predictoren

10.1.1 Eén dichotome predictor

Laten we het adoptievoorbeeld opnieuw analyseren. We willen nagaan of de duur van de adoptieprocedure (variabele `duur`) verklaard kan worden door `conditie`. M.a.w. willen we nagaan of `conditie` een predictor van `duur` is. Laten we `conditie` hercoderen met de numerieke waarden 0 (control) en 1 (experimental) en laten we het corresponderend lineair model schrijven:

$$\texttt{duur}_i = \beta_0 + \beta_{\texttt{conditie}}\ \texttt{conditie}_i + \varepsilon_i. \tag{10.1}$$

[1] Herinner je dat een dichotome variabele een nominale variabele is met slechts twee mogelijke waarden.

Laten we nu de corresponderende voorwaardelijke verwachtingen berekenen.

$$E(\texttt{duur}_i \,|\, \texttt{control}) = \beta_0 + \beta_{\texttt{conditie}}\ 0 = \beta_0;$$
$$E(\texttt{duur}_i \,|\, \texttt{experimental}) = \beta_0 + \beta_{\texttt{conditie}}\ 1 = \beta_0 + \beta_{\texttt{conditie}}.$$

Ons lineair model maakt dus eenvoudige predicties: de predicties in de controle groep zijn allemaal gelijk aan een vast getal (β_0) en de predicties in de experimentele groep zijn allemaal gelijk aan een ander vast getal ($\beta_0 + \beta_{\texttt{conditie}}$). We kunnen die vaste getallen herdopen: $\beta_0 \rightarrow \mu_{\texttt{control}}$ en $\beta_0 + \beta_{\texttt{conditie}} \rightarrow \mu_{\texttt{experimental}}$. Je ziet nu dat het lineair model met `conditie` als predictor eigenlijk equivalent is aan de tweezijdige variant van de alternatieve hypothese die we in Rubr. 7.4 hebben getoetst; namelijk dat $\mu_{\texttt{experimental}} \neq \mu_{\texttt{control}}$.[2]

Laten we nu het lineair model (10.1) toetsen met `R`. We gebruiken nog de functie `lm`.

```
> LM.adopt <- lm(duur ~ conditie, data = adoptieData)
```

De argumenten zijn net zoals in Hoofdstuk 9. `R` weet dat `conditie` een nominale variabele is en gaat dus automatisch die variabele numeriek hercoderen met 0 en 1 volgens de alfabetische volgorde. Dus 0 voor `control` en 1 voor `experimental`. We kijken nu naar de resultaten.

```
> summary(LM.adopt)

Call:
lm(formula = duur ~ conditie, data = adoptieData)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1008 -1.7504  0.0992  1.3992  5.2992

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.9000     0.2810  17.441   <2e-16 ***
conditieexperimental  -0.4992     0.3407  -1.465    0.145
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.102 on 173 degrees of freedom
Multiple R-squared:  0.01226,Adjusted R-squared:  0.006546
F-statistic: 2.146 on 1 and 173 DF,  p-value: 0.1447
```

105. Ga na of de residuen min of meer normaal verdeeld zijn, met een normale qq-plot. Ga ook na of de homoscedasticiteitsassumptie geldig is, met de `Scale-Location` *plot.*

De output is zoals vroeger. Bovenaan vind je het model. Dan beschrijvende statistieken m.b.t. de residuen (geen aanwijzing van asymmetrie). Dan een tabel met een rij per coëfficiënt. In de eerste rij vinden we de schatting van β_0 : 4.9. Dit is de schatting van de verwachting in de controle groep. De tweede

[2] In Rubr. 7.4 was het $\mu_{\texttt{experimental}} < \mu_{\texttt{control}}$.

rij gaat over $\beta_{\texttt{conditie}}$. Deze rij wordt `conditieexperimental` gelabeld. Dit betekent gewoon dat de waarde 1 aan de experimentele conditie toegekend is. De schatting van $\beta_{\texttt{conditie}}$ is -0.4992. Dit betekent dat de verwachting in de experimentele groep met 0.4992 lager is dan in de controle groep. In dezelfde rij vind je ook de aansluitende p-waarde: 0.145. Ze is groter dan 0.05 en we besluiten dus dat $\beta_{\texttt{conditie}}$ niet verschillend is van 0. M.a.w. is `conditie` geen predictor van `duur`.

106. Is de schatting van de verwachting in de controle groep identiek aan de waarde die we in Rubr. 7.4 hebben berekend?

Als je de p-waarde van deze lineaire regressie vergelijkt met de p-waarde van de t-toets van Rubr. 7.4, zie je een groot verschil: 0.145 bij de lineaire regressie en 0.07236 bij de t-toets van Rubr. 7.4. Het verschil is gewoon te wijten aan het feit dat de hypothese bij Rubr. 7.4 eenzijdig was terwijl de hypothese van de lineaire regressie tweezijdig is. Als we de p-waarde van de t-toets van Rubr. 7.4 met twee vermenigvuldigen, dan komen we 0.14472 uit en dat is bijna exact gelijk aan de p-waarde van de lineaire regressie. Er is nog een klein verschil want de lineaire regressie wordt uitgevoerd onder de assumptie dat de voorwaardelijke variantie dezelfde is in beide groepen (homoscedasticiteit) terwijl de t-toets bij Rubr. 7.4 los van die assumptie is.

107. Hoe groot was het verschil tussen $\hat{\mu}_{\texttt{experimental}}$ en $\hat{\mu}_{\texttt{control}}$ in Rubr. 7.4.

108. Lopen vrouwen en mannen even snel? Toets deze hypothese m.b.v. van een t-toets en van enkelvoudige lineaire regressie.

Merk op dat we geen spreidingsdiagram hebben getekend om na te gaan of een lineair verband plausibel is. De reden is dat een niet-lineair verband onmogelijk is als de predictor dichotoom is: door twee punten kan je altijd een rechte tekenen.

10.1.2 In het algemeen

In vorige rubriek hebben we de gegevens van het adoptieonderzoek geanalyseerd a.d.h.v. lineaire regressie en we hebben gezien dat het equivalent is aan de analyses die we in Rubr. 7.4 hebben uitgevoerd met een eenvoudige t-toets. In zich is dit niet echt interessant. Waarom zou je een complexe techniek gebruiken als een eenvoudige techniek even goed is?

Stel dat we niet alleen één dichotome predictor hebben, maar ook andere predictoren (al dan niet dichotoom). Dan is de t-toets van Rubr. 7.4 niet meer bruikbaar maar we kunnen meervoudige lineaire regressie wel gebruiken.

10.1.2.1 Toepassing — gezondheidsuitgaven

We wensen na te gaan of `geslacht` een predictor van `uitgaven` is, rekening houdend met `duur` en `leeftijd`. De predictor `geslacht` is dichotoom, maar daar er meerdere predictoren zijn, mogen we de t-toets van Rubr. 7.4 niet gebruiken. We gebruiken dus meervoudige lineaire regressie en het relevante model is dus

$$\texttt{uitgaven}_i = \beta_0 + \beta_{\texttt{duur}}\,\texttt{duur}_i + \beta_{\texttt{leeftijd}}\,\texttt{leeftijd}_i + \beta_{\texttt{geslacht}}\,\texttt{geslacht}_i + \varepsilon_i$$

en we willen toetsen of $\beta_{\texttt{geslacht}}$ al dan niet nul is. Er zijn twee mogelijk technieken om dit te doen. De t-toets van Rubr. 9.6.1 of de F-toets van Rubr. 9.6.3. We mogen vrij kiezen omdat beide technieken equivalent zijn. We gaan ze allebei illustreren.

a) t-toets We maken een lineair model aan met de functie `lm` en we kijken naar de resultaten van de berekeningen.

```
> LM <- lm(uitgaven ~ duur + leeftijd + geslacht, data = gezondheid)
> summary(LM)

Call:
lm(formula = uitgaven ~ duur + leeftijd + geslacht, data = gezondheid)

Residuals:
    Min      1Q  Median      3Q     Max
-86.264 -26.156  -0.582  26.548 106.305

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.2078    17.4511   3.450 0.000658 ***
duur          1.9430     0.2261   8.592 9.68e-16 ***
leeftijd      0.8282     0.4107   2.017 0.044807 *
geslachtV    11.4408     4.9804   2.297 0.022443 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.78 on 248 degrees of freedom
Multiple R-squared:  0.2685,Adjusted R-squared:  0.2596
F-statistic: 30.34 on 3 and 248 DF,  p-value: < 2.2e-16
```

De schatting van de coëfficiënt $\beta_{\texttt{geslacht}}$ is 11.4408. Dit betekent dat vrouwen gemiddeld gezien 11.44€extra betalen, om de vier maanden, rekening houdend met `duur` en `leeftijd`. M.a.w. het verschil in uitgaven tussen een man en een vrouw met dezelfde werkloosheidsduur en dezelfde leeftijd is gemiddeld gezien 11.44€. Dit zou toevallig kunnen zijn; we hebben misschien toevallig zo'n steekproef getrokken. Wat is de kans om zo'n verschil te observeren als `geslacht` geen echte predictor is (als $\beta_{\texttt{geslacht}} = 0$). Het is gelijk aan de p-waarde die we lezen in de rij van `geslachtV`. Dat is 0.022443. Deze p-waarde is kleiner dan 0.05 en we besluiten dat `geslacht` een predictor van `uitgaven` is.

Merk op dat $\overline{R}^2 = 26\%$. Het was 23% met de unieke predictor `duur` en 25% met twee predictoren (`duur` en `leeftijd`, Rubr. 9.7). Het opnemen van `geslacht` laat dus niet toe om veel extra variantie te verklaren.

b) F-toets We maken twee lineaire modellen aan met de functie `lm` en we vergelijken ze met de functie `anova`.

```
> LM <- lm(uitgaven ~ duur + leeftijd, data = gezondheid)
> LM.geslacht <- lm(uitgaven ~ duur + leeftijd + geslacht,
  data = gezondheid)
> anova(LM,LM.geslacht)
Analysis of Variance Table
```

```
Model 1: uitgaven ~ duur + leeftijd
Model 2: uitgaven ~ duur + leeftijd + geslacht
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    249 380925
2    248 372989  1    7936.4 5.2769 0.02244 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

109. Ga na of de residuen (met 3 predictoren) min of meer normaal verdeeld zijn, met een normale qq-plot. Ga de eerste en tweede Gauss-Markov assumpties ook na.

We komen dezelfde p-waarde uit. Het nadeel van deze methode is dat je geen schatting van $\beta_{\texttt{geslacht}}$ krijgt.

10.1.2.2 Toepassing — geboortes

Laten we dezelfde analyse uitvoeren als in Rubr. 9.6.4.1. We gaan een optimale subset van predictoren selecteren m.b.v. achterwaartse selectie. Deze keer beschouwen we ook `marital` (burgerlijke staat) als een potentiële predictor. We gebruiken opnieuw een lage significantie: $\alpha = 0.01$. We analyseren het volledige model:

```
> summary(lm(wt ~ gestation+parity+age+dwt+number+marital,
  data=geboorte))

Call:
lm(formula = wt ~ gestation + parity + age + dwt + number + marital,
    data = geboorte)

Residuals:
    Min      1Q  Median      3Q     Max
-50.030 -10.711  -0.606   9.909  51.734

Coefficients:
             Estimate Std. Error t value Pr(> |t| )
(Intercept) -30.03038   13.47913  -2.228   0.0263 *
gestation     0.46550    0.04273  10.893  < 2e-16 ***
parity        0.36441    0.41979   0.868   0.3857
age           0.12660    0.13727   0.922   0.3568
dwt           0.12456    0.02957   4.213 2.92e-05 ***
number       -0.29902    0.06190  -4.831 1.73e-06 ***
marital      -3.58554    3.13489  -1.144   0.2532
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.43 on 592 degrees of freedom
Multiple R-squared:  0.2183,Adjusted R-squared:  0.2104
F-statistic: 27.56 on 6 and 592 DF,  p-value: < 2.2e-16
```

De variabele `parity` heeft de grootste p-waarde (0.3857) en deze is groter dan 0.01. We voeren dezelfde analyse zonder `parity`.

```
> summary(lm(wt ~ gestation+age+dwt+number+marital, data=geboorte))

Call:
lm(formula = wt ~ gestation + age + dwt + number + marital, data = geboorte)

Residuals:
    Min      1Q  Median      3Q     Max
-50.527 -10.249  -0.799   9.867  52.982

Coefficients:
             Estimate Std. Error t value Pr(>|t| )
(Intercept) -30.22908   13.47439  -2.243   0.0252 *
gestation     0.46132    0.04245  10.867  < 2e-16 ***
age           0.19017    0.11607   1.638   0.1019
dwt           0.12675    0.02945   4.303 1.97e-05 ***
number       -0.29795    0.06187  -4.815 1.87e-06 ***
marital      -3.62790    3.13386  -1.158   0.2475
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.43 on 593 degrees of freedom
Multiple R-squared:  0.2173,Adjusted R-squared:  0.2107
F-statistic: 32.93 on 5 and 593 DF,  p-value: < 2.2e-16
```

De variabele `marital` heeft de grootste p-waarde (0.2475) en deze is groter dan 0.01. We voeren dezelfde analyse zonder `marital`.

```
> summary(lm(wt ~ gestation+age+dwt+number, data=geboorte))

Call:
lm(formula = wt ~ gestation + age + dwt + number, data = geboorte)

Residuals:
    Min      1Q  Median      3Q     Max
-50.484 -10.653  -0.791   9.992  55.992

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.61253   13.31997  -2.448   0.0146 *
gestation     0.45614    0.04223  10.802  < 2e-16 ***
age           0.19473    0.11604   1.678   0.0938 .
dwt           0.12683    0.02946   4.305 1.95e-05 ***
number       -0.29817    0.06189  -4.818 1.85e-06 ***
---
```

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

Residual standard error: 16.43 on 594 degrees of freedom
Multiple R-squared:  0.2155,Adjusted R-squared:  0.2103
F-statistic:  40.8 on 4 and 594 DF,  p-value: < 2.2e-16
```

De variabele `age` heeft de grootste p-waarde (0.0938) en deze is groter dan 0.01. We voeren dezelfde analyse zonder `age`.

```
> summary(lm(wt ~ gestation+dwt+number, data=geboorte))

Call:
lm(formula = wt ~ gestation + dwt + number, data = geboorte)

Residuals:
    Min      1Q  Median      3Q     Max
-50.213 -10.439  -0.934   9.805  55.574

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.23421   12.78567  -2.052   0.0406 *
gestation     0.45218    0.04223  10.708  < 2e-16 ***
dwt           0.12734    0.02951   4.316 1.86e-05 ***
number       -0.29694    0.06198  -4.791 2.10e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.46 on 595 degrees of freedom
Multiple R-squared:  0.2118,Adjusted R-squared:  0.2079
F-statistic:  53.3 on 3 and 595 DF,  p-value: < 2.2e-16
```

Alle variabelen hebben een p-waarde kleiner dan 0.01 en de achterwaartse selectie stopt hier. De geselecteerde predictoren zijn `gestation`, `dwt` en `number`.

10.2 Lineaire regressie met nominale predictoren

Stel dat we een model willen analyseren waarbij een predictor nominaal is, met meer dan twee niveaus. We gaan die nominale predictor vervangen door meerdere 0-1 predictoren, die we hulpveranderlijken gaan noemen. In het algemeen geldt dat we een nominale predictor met I niveaus moeten hercoderen tot $I-1$ nieuwe hulpveranderlijken (0-1 variabelen) die we vervolgens in het regressiemodel kunnen stoppen. We illustreren dit met een nieuw voorbeeld.

Reactietijd bij bepaalde cognitieve taken spelen een belangrijke rol bij het begrijpen van depressie (zie bv. Kaiser et al. [2008]). Je bent onderzoeker in klinische psychologie en je wil vier types behandeling vergelijken en je gebruikt

een steekproef van 37 patiënten. Ze worden in vier gerandomiseerde groepen ingedeeld en ze volgen één van de vier behandelingen A, B, C of D. Na drie maanden worden ze uitgenodigd om een aantal cognitieve taken uit te voeren. In het data frame `depressie` vind je de reactietijden van de 37 patiënten bij één van die cognitieve taken.

```
> depressie
   behandeling reactietijd
1            A       0.925
2            D       0.875
3            A       0.825
4            B       0.950
...        ...         ...
36           C       1.170
37           C       1.155
```

Daar de variabele `behandeling` nominaal is met meer dan twee niveau's mogen we niet zomaar een lineaire regressie uitvoeren om te weten of verschillen in `reactietijd` verklaard kunnen worden door `behandeling`. We kunnen ook geen t-toets van hoofdstuk 6 gebruiken omdat de t-toets om verwachtingen te vergelijken alleen met twee groepen werkt, niet met meer dan twee. We gaan dus de variabele `behandeling` hercoderen tot 3 $(= 4 - 1)$ nieuwe hulpveranderlijken met twee niveaus: 0 en 1.

10.2.1 Hercodering

We kunnen een onderscheid maken tussen twee hercoderingen: dummy-codering en effect-codering.

- Bij dummy-codering kiest men 1 van de I niveaus als referentieniveau en worden de andere niveaus via een 0-1 variabele gecodeerd.

  In het geval van het voorbeeld betekent dit dat we 3 hulpveranderlijken X_1, X_2 en X_3 moeten aanmaken. Wanneer we behandeling D als referentieniveau beschouwen[3], dan bekomen we de volgende codering:

Behandeling	X_1	X_2	X_3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

  Dit betekent concreet dat voor een individu i

  - die behandeling A volgt, geldt: $x_{i1} = 1$, $x_{i2} = 0$, $x_{i3} = 0$.
  - die behandeling B volgt, geldt: $x_{i1} = 0$, $x_{i2} = 1$, $x_{i3} = 0$.

[3] De keuze van het referentieniveau is vrij

  - die behandeling C volgt, geldt: $x_{i1} = 0$, $x_{i2} = 0$, $x_{i3} = 1$.
  - die behandeling D volgt, geldt: $x_{i1} = 0$, $x_{i2} = 0$, $x_{i3} = 0$.

- Bij effect-codering wordt ook een groep gekozen maar deze groep wordt steeds met -1 gecodeerd i.p.v. met 0. Deze groep wordt niet als referentie beschouwd. Voor het voorbeeld bekomen we:

Behandeling	X_1	X_2	X_3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

  Dit betekent dat de codering hetzelfde is als de dummy-codering voor individuen die behandeling A, B of C volgen maar voor een individu i die behandeling D volgt, geldt: $x_{i1} = -1$, $x_{i2} = -1$, $x_{i3} = -1$.

Het effect van de behandeling op de verwachting van de reactietijd Y kunnen we als volgt modelleren:

$$E(Y_i \,|\, x_{i1}, x_{i2}, x_{i3}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Naargelang het coderingsschema dat gehanteerd wordt, hebben de regressieparameters een andere betekenis.

- Dummy-codering
  - $E(Y_i \,|\, \mathrm{D}) = E(Y_i \,|\, x_{i1} = 0, x_{i2} = 0, x_{i3} = 0) = \beta_0$. De coëfficiënt β_0 stelt dus de verwachte reactietijd voor bij behandeling D.
  - $E(Y_i \,|\, \mathrm{A}) = E(Y_i \,|\, x_{i1} = 1, x_{i2} = 0, x_{i3} = 0) = \beta_0 + \beta_1$. De coëfficiënt β_1 stelt dus het verschil voor van de verwachte reactietijd bij behandeling A en de verwachte reactietijd bij behandeling D.
  - Analoog stellen β_2 en β_3 het verschil in verwachte reactietijd tussen behandeling B en behandeling D en tussen behandeling C en behandeling D.

- Effect-codering
  - In dit geval kan aangetoond worden dat β_0 het *marginale gemiddelde*[4] van de reactietijd voorstelt, i.e. het gemiddelde van de verwachtte reactietijden over de verschillende behandelingen heen:

$$\beta_0 = \big(\; E(Y_i \,|\, \mathrm{A}) + E(Y_i \,|\, \mathrm{B}) + E(Y_i \,|\, \mathrm{C}) + E(Y_i \,|\, \mathrm{D}) \;\big)/4.$$

110. Bereken het marginale gemiddelde van `reactietijd`.

[4] Let op; dezelfde uitdrukking "marginale gemiddelde" wordt gebruikt voor het gemiddelde van de verwachtingen en voor het gemiddelde van de corresponderende gemiddelden.

– $E(Y_i \mid \mathrm{A}) = E(Y_i \mid x_{i1} = 1, x_{i2} = 0, x_{i3} = 0) = \beta_0 + \beta_1$. De coëfficiënt β_1 stelt dus het verschil voor tussen de verwachte reactietijd bij behandeling A en het marginale gemiddelde.

– In het algemeen, β_ℓ ($\ell = 1, 2, 3$) drukt het verschil uit tussen de verwachte reactietijd bij behandeling ℓ en het marginale gemiddelde.

– De verwachte reactietijd bij behandeling D is

$$E(Y_i \mid \mathrm{D}) = E(Y_i \mid x_{i1} = -1, x_{i2} = -1, x_{i3} = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3.$$

Dit betekent dat het verschil tussen de verwachte reactietijd bij behandeling D en het marginale gemiddelde gelijk is aan $-\beta_1 - \beta_2 - \beta_3$.

10.2.2 Welke hypothese?

De variabele X_1 heeft geen betekenis in zich. Stel dat we weten dat $x_{1i} = 0$ bij individu i. Dit geeft ons geen duidelijke informatie over dat individu. Hetzelfde geldt voor elke hulpveranderlijke. We gaan dus nooit toetsen of β_j al dan niet nul is. Als we modellen vergelijken, gaan we ook nooit een model beschouwen met de hulpveranderlijke X_1 en zonder de hulpveranderlijke X_2. Idem bij predictorenselectie: we beschouwen alleen modellen die alle hulpveranderlijken bevatten of geen.

Dus, indien we willen toetsen of een nominale variabele een predictor van Y is, dan gaan we het model met alle hulpveranderlijken vergelijken met het model zonder de hulpveranderlijken, a.d.h.v. een F-toets. Merk op dat het resultaat van deze toets onafhankelijk is van het gehanteerde coderingsschema: dummy-codering en effect-codering geven dezelfde resultaten. Het resultaat is ook onafhankelijk van het gekozen referentieniveau.

10.2.3 Berekeningen met `R`

Dankzij `R` hoeven we niet zelf hulpveranderlijken te definiëren. We hoeven ook niet zelf een coderingsschema en een referentieniveau te kiezen. `R` doet het allemaal voor ons. We gaan nog de functie `lm` gebruiken en als één (of meerdere) van de predictoren nominaal is, gaat `R` automatisch dummy-codering gebruiken met het eerste niveau als referentie. `R` gaat ook zelf hulpveranderlijken definiëren. Voorbeeld:

```
> LM.depressie <- lm( reactietijd ~ behandeling, data = depressie)
> summary(LM.depressie)

Call:
lm(formula = reactietijd ~ behandeling, data = depressie)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15812 -0.04550 -0.01611  0.05889  0.13050
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.91111    0.02744  33.207  < 2e-16 ***
behandelingB  0.07839    0.03782   2.073  0.04608 *
behandelingC  0.27939    0.03782   7.387 1.74e-08 ***
behandelingD  0.12201    0.04000   3.051  0.00448 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08231 on 33 degrees of freedom
Multiple R-squared:  0.6418,Adjusted R-squared:  0.6093
F-statistic: 19.71 on 3 and 33 DF,  p-value: 1.674e-07
```

111. Ga de normaliteitsassumptie, de 1ste en de 2de Gauss-Markov assumpties na.

Het commando is vanzelfsprekend. De eerste regels van de output zijn zoals vroeger. De tabel met de coëfficiënten bevat vier regels: één per hulpveranderlijke. In de regel van het intercept vinden we $\hat{\beta}_0 = 0.91111$: de schatting van de verwachte reactietijd bij behandeling A. De corresponderende p-waarde is, zoals bijna altijd, niet relevant. In de regel `behandelingB` vinden we $\hat{\beta}_B = 0.07839$: de schatting van het verschil tussen de verwachte reactietijd bij behandelingen A en B. De corresponderende p-waarde is niet betekenisvol omdat één hulpveranderlijke in zich geen betekenis heeft. Dan hebben we nog twee analoge regels voor behandelingen B en C. Helemaal onderaan vinden we de p-waarde van de F-toets die ons model vergelijkt met het nulmodel. Deze p-waarde (`1.674e-07`) is wel relevant. Ze is kleiner dan 0.05 en we besluiten dus dat `behandeling` een predictor is van `reactietijd`. Daar de p-waarde veel kleiner is dan 0.05 is de schending van homoscedasticiteit (zie oefening 111) niet belangrijk.

112. Gebruikmakend van dezelfde codering als `R`*, wat zijn de waarden van de hulpveranderlijken X_1, X_2 en X_3 bij individu 27?*

Er zijn contexten waar effect-codering handiger is dan dummy-codering. Het is dan mogelijk om dat coderingschema te hanteren bij de berekeningen met `R`; Dit wordt in deze cursus niet gezien.

10.2.4 Een voorbeeld met meerdere predictoren — `sportData`

We willen nagaan of de variabele `type` een predictor van `tijd` is, rekening houdend met `lengte` en `sport`. De variabele `type` is een nominale variabele met vijf niveaus: `andere`, `basketbal`, `tennis`, `voetbal` en `zwemmen`. Het referentieniveau zal dus `andere` zijn en `R` gaat vier hulpveranderlijken definiëren. Laten we het model met de drie predictoren analyseren.

```
> LM.sport <- lm(tijd ~ lengte + sport + type, data = sportData)
> summary(LM.sport)

Call:
lm(formula = tijd ~ lengte + sport + type, data = sportData)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-9.8812 -2.8280 -0.2145  2.6244  9.7401

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.97628    3.47839   7.180 1.47e-11 ***
lengte        -0.04436    0.01887  -2.350   0.0198 *
sport          1.75088    0.28944   6.049 7.42e-09 ***
typebasketbal  0.82205    1.04452   0.787   0.4322
typetennis     0.83061    1.00160   0.829   0.4080
typevoetbal    0.86430    0.88839   0.973   0.3318
typezwemmen   -0.17403    0.99309  -0.175   0.8611
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.268 on 193 degrees of freedom
Multiple R-squared:  0.1987,Adjusted R-squared:  0.1737
F-statistic: 7.975 on 6 and 193 DF,  p-value: 1.076e-07
```

Daar er meerdere predictoren zijn, kunnen we hier het intercept (24.97628) niet interpreteren als de schatting van $\mu_{\texttt{tijd}}$ bij individuen die aan een andere sport doen. De correcte interpretatie is: het intercept (24.97628) is de schatting van $\mu_{\texttt{tijd}}$ bij individuen die aan een andere sport doen en waarbij de variabelen `lengte` en `sport` nul zijn. Zulke individuen bestaan uiteraard niet en het intercept heeft dus geen concrete en intuïtieve betekenis.

De coëfficiënt van `lengte` is -0.04436. Dit is de schatting van $\beta_{\texttt{lengte}}$. Dit betekent dat twee individuen met een verschil van één cm op `lengte` en met identieke scores op alle andere variabelen een verschil van -0.04436 seconde zullen ervaren op `tijd` (gemiddeld gezien). Daar de coëfficiënt negatief is, zal `tijd` lager zijn bij het langere individu. Het langere individu loopt dus sneller (gemiddeld gezien).

De coëfficiënt van `typebasketbal` (0.82205) representeert het gemiddelde tijdsverschil tussen een individu die aan basketbal doet en een individu die aan "andere" doet, indien ze identieke scores hebben op de andere variabelen. Een analoge interpretatie geldt voor de coëfficiënten van `typetennis`, `typevoetbal` en `typezwemmen`.

De p-waarde (`1.076e-07`) van de F-toets helemaal onderaan de output heeft niets te maken met onze onderzoeksvraag (is `type` een predictor van `tijd`, rekening houdend met `lengte` en `sport`?). Deze p-waarde heeft betrekking tot de vergelijking van het nulmodel (zonder predictor) met het model met drie predictoren.

Om onze onderzoeksvraag te beantwoorden moeten we het model met drie predictoren vergelijken met hetzelfde model maar zonder `type` We maken dus nu een lineair model aan zonder de predictor `type`, maar wel met de twee andere predictoren..

```
> LM.sportZonderType <- lm(tijd ~ lengte + sport, data = sportData)
```

Nu kunnen we beide modellen vergelijken m.b.v. een F-toets, dankzij de functie `anova`.

```
> anova(LM.sportZonderType,LM.sport)
Analysis of Variance Table

Model 1: tijd ~ lengte + sport
Model 2: tijd ~ lengte + sport + type
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    197 3556.1
2    193 3514.9  4    41.155 0.5649 0.6884
```

De p-waarde (0.6884) van deze toets is groter dan 0.05 en we besluiten dat `type` geen predictor van `tijd` is.

10.2.5 Nog een voorbeeld — `microbusiness`

In de Verenigde Staten zijn er veel programma's om vrouwen met een laag inkomen die een microbusiness stichten financieel te ondersteunen. In een onderzoek [Sanders, 2004] wil de auteur nagaan of die programma's efficiënt zijn. Drie steekproeven worden getrokken: een steekproef van vrouwen die de steun van zo'n programma krijgen ($n = 62$), een steekproef van vrouwen die een microbusiness hebben maar geen steun krijgen ($n = 57$) en een steekproef van vrouwen die geen microbusiness hebben maar wel werken ($n = 178$). De toename (of afname) van het inkomen tussen 1991 en 1995 wordt geregistreerd, samen met het ras. De gegevens (data frame `microbusiness`) zien er als volgt uit:

```
> microbusiness
            groep inkomenWijziging   race
1          GeenMB             8941 latino
2          GeenMB            -5798  black
3          GeenMB            19240  black
4          GeenMB            -9746  white
5          GeenMB             3023  black
6       MBMetSteun             8200  black
...           ...              ...    ...
295 MBZonderSteun            19712  white
296        GeenMB            22082  white
297        GeenMB             2656 latino
```

We gaan eerst de gemiddelde inkomenwijziging in de drie groepen berekenen. Als we het commando `mean(microbusiness$inkomenWijziging)` gebruiken, dan komen we het gemiddelde van alle vrouwen. Dat is niet wat we nodig hebben. Om de gemiddelden in de drie groepen apart te krijgen gebruiken we de functie `aggregate`:

```
> aggregate( formula = microbusiness$inkomenWijziging ~ microbusiness$groep,
```

```
   FUN = mean )
  microbusiness$groep microbusiness$inkomenWijziging
1              GeenMB                            8652
2           MBMetSteun                            5708
3        MBZonderSteun                            6455
```

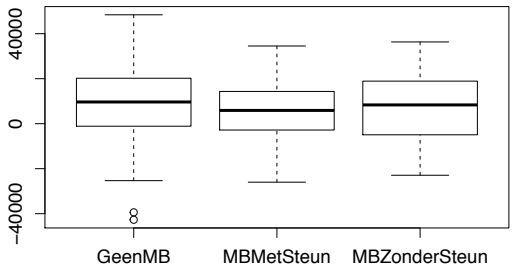
Het argument `formula` werkt zoals in het vorige hoofdstuk. Het maakt duidelijk dat we wensen de variabele `inkomenWijziging` te voorspellen m.b.v. de predictor `groep`. Het argument `FUN` is de afkorting van 'functie' en wordt gebruikt om `R` te zeggen wat hij moet berekenen in elke groep. Met het argument `FUN = mean` weet `R` dat hij het gemiddelde in elke groep moet berekenen. Met het argument `FUN = var` zou `R` de schatting van de variantie in elke groep berekenen. Met `FUN = median` zou `R` de mediaan berekenen. Enz.

113. Voer het commando `aggregate( tijd geslacht+type, FUN = mean, data = sportData)` *uit. Begrijp je de output?*

We kijken nu naar de output van het commando. We zien dat de drie gemiddelden niet identiek zijn en dat de verschillen niet gering zijn (ongeveer 3000$ tussen groep 1 en 2). Kunnen we hieruit afleiden dat de drie verwachtingen in de populaties niet identiek zijn? Niet zomaar. Laten we voorzichtig zijn en laten we de gegevens visueel analyseren met de `boxplot` functie:

```
> boxplot(formula=microbusiness$inkomenWijziging ~ microbusiness$groep)
```

De output wordt in Fig. 10.1 weergegeven. De boxplots tonen dat de medianen ook van elkaar verschillen (ongeveer zoals de gemiddelden) maar vooral dat



Figuur 10.1: Boxplot van de inkomenwijzigingen in de drie groepen.

de variatie binnen elke steekproef zeer groot is: veel groter dan de verschillen tussen de medianen of tussen de gemiddelden. De verschillen tussen de medianen lijken bijna verwaarloosbaar t.o.v. de variatie binnen elke steekproef. Dit geeft de indruk dat de verschillen tussen de drie groepen toevallig zijn.

We berekenen nu de gemiddelden bij de drie rassen.

114. Gebruik de functies `aggregate` *en de* `boxplot` *om de verdeling van de toevalsvariabele* `score` *te vergelijken in de drie opleidingen psy, ped en soc, m.b.v. het data frame* `myData`.

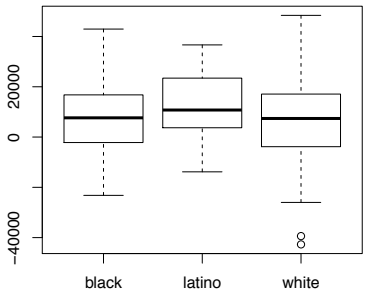
```
> aggregate( formula=inkomenWijziging ~ race,FUN = mean,
  data=microbusiness)
    race inkomenWijziging
1  black         7460.496
2 latino        12168.581
3  white         6709.917
```

Er zijn grote verschillen tussen de drie rassen. Laten we nu de boxplots tekenen.

```
boxplot(formula=microbusiness$inkomenWijziging ~ microbusiness$race)
```

115. Maak een lineair model aan om `score` *te verklaren m.b.v.* `opleiding`*. Gebruik de functie* `plot` *om de homoscedasticiteitsassumptie na te gaan. Vergelijk met de boxplot van oefening 114.*

De output wordt in Fig. 10.2 weergegeven. Zoals bij Fig. 10.1 lijken de verschillen tussen de medianen bijna verwaarloosbaar t.o.v. de variatie binnen elke steekproef. Dit geeft de indruk dat de verschillen tussen de drie rassen ook toevallig zijn. We gaan een lineair model gebruiken om de verschillen tussen



Figuur 10.2: Boxplot van de inkomenwijzigingen bij de drie rassen.

groepen en rassen te verklaren en we gaan dit model toetsen. We willen dus een lineair model toetsen met `groep` en `race` als predictoren. Omdat beide predictoren nominaal zijn, worden ze ook factor genoemd.

```
> LM.mb <- lm(inkomenWijziging ~ groep + race, data = microbusiness)
> summary(LM.mb)

Call:
lm(formula = inkomenWijziging ~ groep + race, data = microbusiness)

Residuals:
   Min     1Q Median     3Q    Max
-50418  -9537    412  10428  40684

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         8486.6     1503.2   5.646  3.9e-08 ***
groepMBMetSteun    -2773.1     2227.3  -1.245    0.214
groepMBZonderSteun -2200.2     2298.4  -0.957    0.339
racelatino          4555.0     3011.4   1.513    0.131
racewhite           -770.7     1852.5  -0.416    0.678
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15090 on 292 degrees of freedom
Multiple R-squared:  0.01786,Adjusted R-squared:  0.004411
```

```
F-statistic: 1.328 on 4 and 292 DF,  p-value: 0.2596
```

Zoals bij de andere voorbeelden met nominale variabelen heeft `R` hulpveranderlijken gedefinieerd: twee om `groep` te hercoderen en twee om `race` te hercoderen. Het referentieniveau is alfabetisch bepaald: `GeenMB` voor de groep en `black` voor het ras.

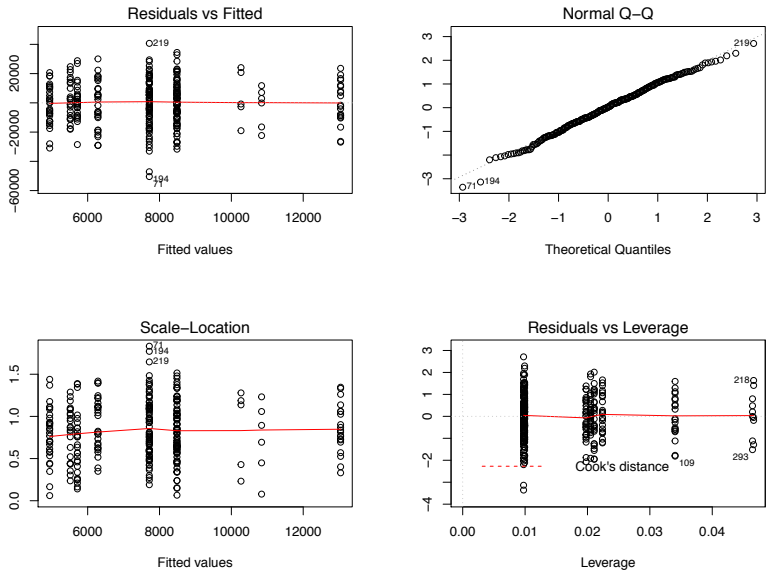
Laten we de output van `summary(LM.mb)` bespreken. In de rij van het intercept lezen we 8486.6 af. Dit is $\hat{\beta}_0$: de schatting van de voorwaardelijke verwachting van Y voor zwarte vrouwen die geen microbusiness hebben.

Wat is de schatting van de voorwaardelijke verwachting van Y voor zwarte vrouwen die wel een microbusiness hebben, maar geen steun? Het is $8486.6 - 2200.2 = 6286.4$. En de schatting van de voorwaardelijke verwachting van Y voor latino vrouwen die een microbusiness hebben, zonder steun? Het is $8486.6 - 2200.2 + 4555.0 = 10841.4$. Enz.

De p-waarde van de F-toets is 0.2596. Ons lineair model maakt dus predicties die niet significant beter zijn dan die van het nulmodel en we aanvaarden dus het nulmodel.

We kijken nu naar de output (Fig. 10.3) van het commando `plot(LM.mb)` om de assumpties van lineaire regressie na te gaan.

```
> plot(LM.mb)
Hit <Return> to see next plot:
```



Figuur 10.3: Assumptiecheck: output van `plot(LM.mb)`.

De "`Residuals vs Fitted`" plot ziet er niet uit zoals bij de andere voorbeelden. We zien niet echt een puntenwolk, maar verticale lijnen. De reden is dat de

predictoren nominaal zijn. Het lineair model maakt dan een beperkt aantal predicties: één per combinatie van de niveaus van de factoren. We hebben bij dit voorbeeld twee factoren, elk met drie niveaus: dus $3 \times 3 = 9$ combinaties en er zijn inderdaad 9 verticale lijnen. Voor de rest is de interpretatie van deze grafiek zoals vroeger. De rode curve is min of meer horizontaal en de eerste Gauss-Markov assumptie is dus in orde.

De normale qq-plot is in orde.

Op de "`Scale-Location`" plot zien we ook 9 verticale lijnen omdat de predictoren nominaal zijn. Voor de rest is de rode curve min of meer horizontaal en de tweede Gauss-Markov assumptie is dus in orde. De vierde grafiek wordt niet besproken.

10.3 Historische nota — variantie-analyse (anova)

Een techniek om verwachtingen in twee groepen te vergelijken werd in het begin van de 20ste eeuw ontwikkeld: de t-toets (zie Rubr. 6.5.2). Deze techniek werd in de eerste helft van de 20ste eeuw veralgemeend om verwachtingen in p groepen te vergelijken. Deze techniek heet variantie-analyse (analysis of variance, anova). Indien de groepen bepaald worden op basis van één nominale variabele, dan spreekt men van one-way anova. Bv. zijn de verwachte scores op het examen Statistiek II identiek in de groepen van psychologie studenten, ped. wetenschappen studenten en sociaal werk studenten? De nominale variabele van belang is `opleiding`.

Indien de groepen bepaald worden op basis van twee nominale variabelen, dan spreekt men van two-way anova. Bv. zijn de verwachte scores op het examen Statistiek II identiek in de groepen van vr. psy., man. psy., vr. ped., man. ped., vr. soc. en man. soc. studenten? De nominale variabelen van belang zijn nu `opleiding` en `geslacht`. Als we die twee nominale variabelen (of factoren) kruizen, dan komen we 6 ($= 3 \times 2$) groepen uit. Three-way, four-way, ... anova worden op dezelfde manier gedefinieerd.

Bij een variantie-analyse wordt, zoals bij een t-toets, nagegaan of de verschillen tussen de groepen het effect van het toeval kunnen zijn (nulhypothese) of niet. Schattingen van de verwachtingen in elke groep worden dan berekend en ze kunnen gebruikt worden om predicties te maken. Bv. als student A een man is en psychologie studeert, dan kan je voorspellen dat zijn score op Statistiek II gelijk zal zijn aan de geschatte verwachting van de corresponderende groep. Met regressie-analyse met nominale variabelen doen we eigenlijk hetzelfde en het is mogelijk te bewijzen dat beide technieken equivalent zijn, met nominale variabelen. De p-waarde van de F-toets bij een variantie-analyse is identiek aan de p-waarde van de F-toets bij een lineaire regressie. Maar lineaire regressie laat ook toe om continue predictoren van ratio of interval meetniveau te gebruiken. Lineaire regressie is dus algemener (of krachtiger) en er is geen reden om beide technieken te studeren en te gebruiken. In deze cursus wordt dus geopteerd om geen variantie-analyse te zien.

Daar de berekeningen simpler zijn bij variantie-analyse dan bij lineaire re-

gressie en daar krachtige computers duur waren tot het einde van de 20ste eeuw, is variantie-analyse populair gebleven tot in het begin van de 21ste eeuw en wordt vandaag nog gebruikt en gedoceerd. In de wetenschappelijke literatuur zal je dus waarschijnlijk artikels lezen waarin de resultaten van een variantie-analyse gerapporteerd worden. Hieronder vind je een paar tips om die artikels te begrijpen.

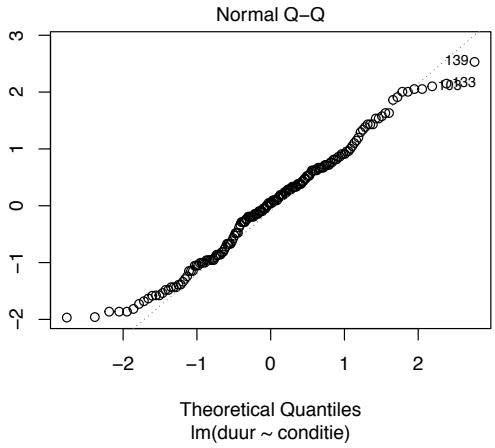
Bij variantie-analyse wordt niet van $\mathrm{SS}_{\mathrm{Mod}}$ gesproken, maar van $\mathrm{SS}_{\mathrm{between}}$. Ze zijn gelijk aan elkaar. Er wordt ook niet van $\mathrm{SS}_{\mathrm{Res}}$ gesproken, maar van $\mathrm{SS}_{\mathrm{within}}$. Die twee termen zijn ook synoniem. De determinatiecoëfficiënt wordt ook aangeduid door R^2 en is ook gelijk aan $\mathrm{SS}_{\mathrm{Mod}} \,/\, \mathrm{SS}_{\mathrm{Tot}}$. De toets van het model met p factoren (p nominale variabelen) is volledig equivalent aan de toets van het lineair model met p nominale predictoren. Deze toets is ook gebaseerd op een F-verhouding, met een F-verdeling, met dezelfde aantallen vrijheidsgraden als bij lineaire regressie. De p-waarde van beide toetsen is dezelfde en wordt op dezelfde manier geïnterpreteerd.

Covariantie-analyse (ancova) is een uitbreiding van variantie analyse om verwachtingen in verschillende groepen te kunnen vergelijken (zoals anova), rekening houdend met andere variabelen (covariaten) die eventueel van interval of ratio meetniveau zijn. Deze techniek wordt in deze cursus niet gezien omdat ze ook een speciaal geval is van meervoudige lineaire regressie, met een mengeling van nominale variabelen en variabelen van interval of ratio meetniveau.

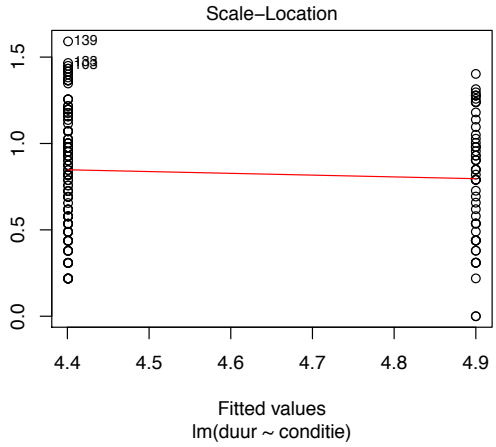
10.4 Oplossingen

105) Ga na of de residuen min of meer normaal verdeeld zijn, met een normale qq-plot. Ga ook na of de heteroscedasticiteitsassumptie geldig is, met de `Scale-Location` plot.

Oplossing: Voor beide plots gebruik je `plot(LM.adopt)`. Voor de normaliteit kijk je naar de tweede grafiek:



Geen ernstige afwijking van de diagonaal. Voor de heteroscedasticiteit kijk je naar de derde grafiek:



De rode curve is bijna horizontaal. Geen schending van de heteroscedasticiteitsassumptie.

106) Is de schatting van de verwachting in de controle groep identiek aan de waarde die we in Rubr. 7.4 hebben berekend?

Oplossing: Ja, het was ook 4.9.

107) Hoe groot was het verschil tussen $\hat{\mu}_{\texttt{experimental}}$ en $\hat{\mu}_{\texttt{control}}$ in Rubr. 7.4.

Oplossing: Het was $4.9 - 4.40084 = 0.49916$.

108) Lopen vrouwen en mannen even snel? Toets deze hypothese m.b.v. van een t-toets en van enkelvoudige lineaire regressie.

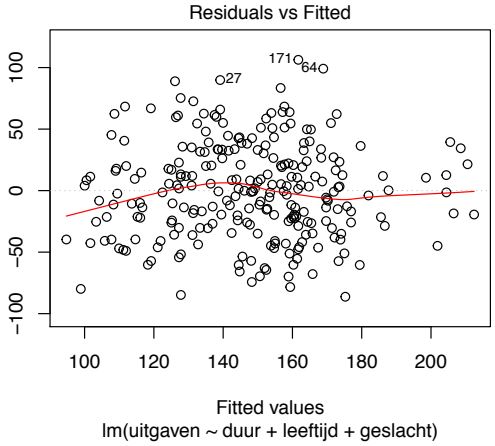
Oplossing: Voor de t-toets, zie Rubr. 6.5.2.1. Voor enkelvoudige lineaire regressie maken we een lineair model aan en we gebruiken de functie `summary`.

```
> summary(LM)

Call:
lm(formula = tijd ~ geslacht, data = sportData)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5096  -3.4059  -0.4096   2.8191  12.6904

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.6096     0.4845  46.667   <2e-16 ***
geslachtV     0.5951     0.6655   0.894    0.372
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.697 on 198 degrees of freedom
Multiple R-squared:  0.004023,Adjusted R-squared:  -0.001007
F-statistic: 0.7998 on 1 and 198 DF,  p-value: 0.3722
```

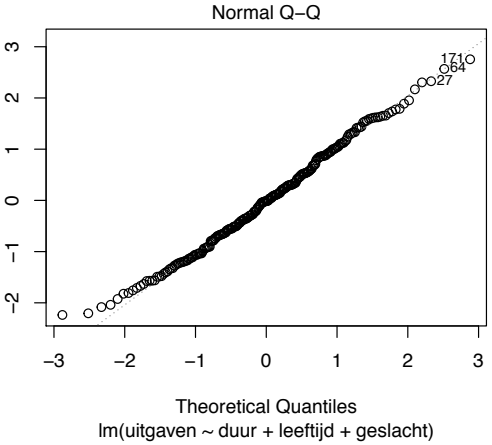
De p-waarde is groter dan 0.05 en we aanvaarden de nulhypothese. Merk op dat de p-waarde bijna identiek is aan die van de t-toets. Er is een klein verschil want lineaire regressie is gebaseerd op de homoscedasticiteitsassumptie (identieke voorwaardelijke variantie in beide groepen) terwijl deze hypothese niet gedaan wordt bij de Welch t-toets.

109) Ga na of de residuen (met 3 predictoren) min of meer normaal verdeeld zijn, met een normale qq-plot. Ga de eerste en tweede Gauss-Markov assumpties ook na.

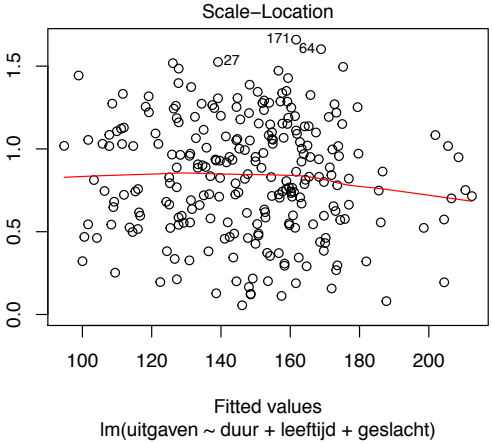
Oplossing: Gebruik het commando `plot(LM.geslacht)`. We kijken naar de eerste drie grafieken:



De rode curve is bijna horizontaal: de eerste Gauss-Markov assumptie is in orde.



De normaliteitsassumptie is in orde.



De homoscedasticiteitsassumptie is in orde.

110) Bereken het marginale gemiddelde van `reactietijd`.

Oplossing: We maken vier vectoren aan met de reactietijden in de vier groepen.

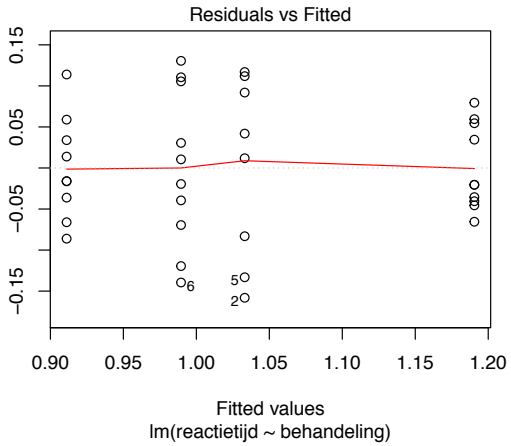
```
> rA <- depressie$reactietijd[depressie$behandeling == "A"]
> rB <- depressie$reactietijd[depressie$behandeling == "B"]
> rC <- depressie$reactietijd[depressie$behandeling == "C"]
> rD <- depressie$reactietijd[depressie$behandeling == "D"]
```

Nu berekenen we het gemiddelde in elke groep en het gemiddelde van de gemiddelden:

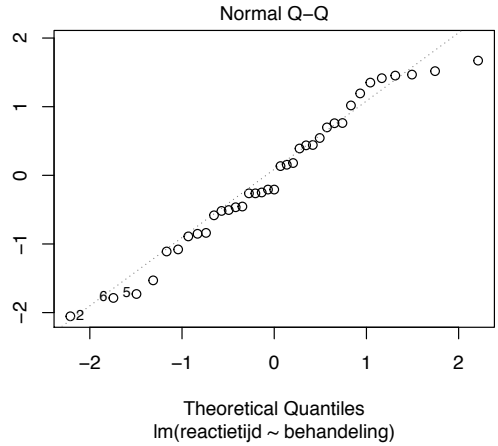
```
> (mean(rA)+mean(rB)+mean(rC)+mean(rD))/4
[1] 1.031059
```

111) Ga de normaliteitsassumptie, de 1ste en de 2de Gauss-Markov assumpties na.

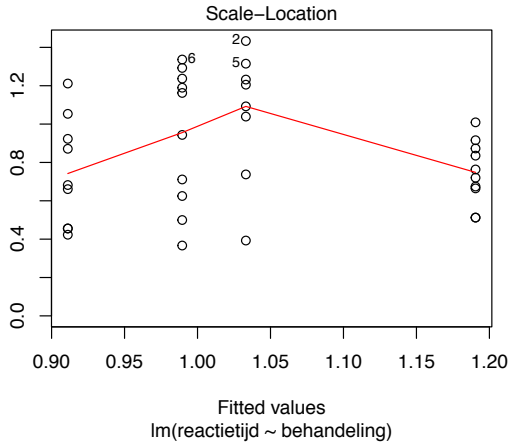
Oplossing: Gebruik het commando `plot(LM.depressie)` en kijk naar de eerste drie grafieken.



In orde.



In orde.



De voorwaardelijke variantie blijkt groter te zijn in de groep waarvoor de fitted value (predictie) ongeveer gelijk aan 1.03 is (welke groep is dit?). We hoeven voorzichtig te zijn bij de interpretatie van de resultaten.

112) Gebruikmakend van dezelfde codering als `R`, wat zijn de waarden van de hulpveranderlijken X_1, X_2 en X_3 bij individu 27?

Oplossing: Tot welke groep behoort individu 27?

```
> depressie[27,]
   behandeling reactietijd
27           C       1.145
```

`R` gebruikt groep A als referentieniveau. Dus $x_{27,1} = 0, x_{27,2} = 1$ en $x_{27,3} = 0$.

113) Voer het commando `aggregate( tijd  geslacht+type, FUN = mean, data = sportData)` uit. Begrijp je de output?

Oplossing:

```
> aggregate( tijd ~ geslacht+type, FUN = mean, data = sportData)
   geslacht      type     tijd
1         M    andere 21.41667
2         V    andere 24.03000
3         M basketbal 23.69231
4         V basketbal 22.34444
5         M    tennis 21.79091
6         V    tennis 23.90417
7         M   voetbal 23.66486
8         V   voetbal 23.14545
9         M   zwemmen 21.10000
10        V   zwemmen 22.45455
```

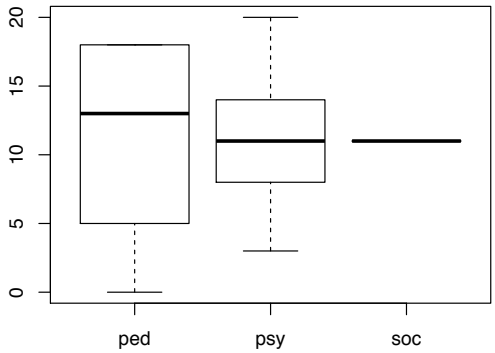
`R` heeft het gemiddelde berekend in de 10 groepen die resulteren uit het kruisen van `geslacht` met `type`.

114) Gebruik de functies `aggregate` en de `boxplot` om de verdeling van de toevalsvariabele `score` te vergelijken in de drie opleidingen psy, ped en soc, m.b.v. het data frame `myData`.

Oplossing:

```
> aggregate( score ~ opleiding, FUN = mean, data = myData)
  opleiding    score
1       ped 11.60000
2       psy 11.33333
3       soc 11.00000
> boxplot(formula=myData$score ~ myData$opleiding)
```

Output:



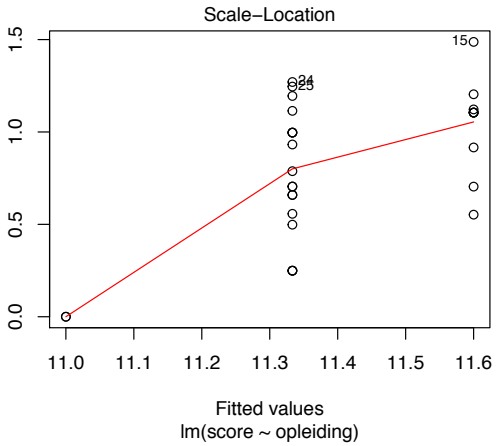
De drie gemiddelden zijn bijna identiek. De verschillen tussen de drie medianen zijn iets groter. De varianties zijn helemaal niet identiek.

115) Maak een lineair model aan om `score` te verklaren m.b.v. `opleiding`. Gebruik de functie `plot` om de homoscedasticiteitsassumptie na te gaan. Vergelijk met de boxplot van oefening 114.

Oplossing:

```
> LM.score <- lm(score ~ opleiding, data = myData)
> plot(LM.score)
```

We kijken naar de derde grafiek:



De voorwaardelijke variantie is helemaal niet constant. Dit komt overeen met onze besluit bij oefening 114. We mogen dus geen lineaire regressie gebruiken om het verband tussen `score` en `opleiding` te analyseren.

Hoofdstuk 11

Categorische data-analyse

Categorische variabelen zijn variabelen die ofwel nominaal ofwel ordinaal (met weinig verschillende niveaus) zijn. In het geval van een ordinale variabele zijn de categorieën geordend en in het geval van een nominale variabele zijn ze niet geordend. Het probleem of de moeilijkheid met die variabelen is dat de waarden van de variabelen niet echt relevant of belangrijk zijn. Wat echt relevant is, is het aantal observaties (frequentie) met bepaalde waarden of de proportie (relatieve frequentie) met bepaalde waarden. Als we bv. de variabele 'haarkleur' analyseren, dan gaan we de waarden van de variabele (blond, bruin, enz.) niet kunnen optellen of vermenigvuldigen. We zullen enkel met het aantal blonden en het aantal bruinen kunnen werken.

Tot nu toe hebben we veel technieken in de inductieve statistiek gezien die niet met categorische variabelen kunnen gebruikt worden. Een voorwaarde voor veel technieken is dat de variabelen tenminste van interval meetniveau moeten zijn. We hebben toch één techniek gezien die wel gebruikt kan worden met categorische variabelen : de toets voor één proportie (par. 6.6). We hebben die techniek gebruikt met het voorbeeld omtrent de proportie van alcoholisten. In dat voorbeeld hadden we inderdaad een categorische variabele : het al dan niet alcoholistisch zijn. Die variabele is categorisch maar heeft ook iets speciaals : ze heeft slechts twee categorieën. Men zegt dat ze dichotoom is.

In dit hoofdstuk gaan we een techniek zien die geschikt zijn voor variabelen met twee of meer dan twee categorieën.

11.1 Inleidend voorbeeld

Een onderzoeker wil nagaan of er verschillen zijn tussen de verdelingen van de studierichtingen van universiteitstudenten afkomstig van het katholieke en het officiële onderwijs. Hij trekt een steekproef van 300 studenten in Vlaanderen en registreert de twee variabelen `onderwijsnet` (officiëel of katholiek) en `richting` (wetenschappen, sociale wetenschappen, letteren en andere). Hieronder gebruik ik het commando `head` om de eerste zes rijen van de data frame `netRichting`

te tonen.

```
> head(netRichting)
  onderwijsnet  richting
1    officieel      wet.
2        katho  letteren
3    officieel soc. wet.
4    officieel soc. wet.
5        katho soc. wet.
6    officieel soc. wet.
```

De bivariate frequentieverdeling van de twee variabelen krijg je met volgende commando (zie Rubr. 2.1):

```
> table(netRichting$onderwijsnet, netRichting$richting )

           andere letteren soc. wet. wet.
  katho         20       20        75   15
  officieel     12       25       115   18
```

In dit geval, omdat de data frame slechts twee variabelen bevat, kan je ook de bivariate frequentieverdeling als volgt bekomen:

```
> table(netRichting)
            richting
onderwijsnet andere letteren soc. wet. wet.
   katho         20       20        75   15
   officieel     12       25       115   18
```

Om de gegevens gemakkelijker te kunnen interpreteren gaan we dezelfde tabel met proporties i.p.v. frequenties opstellen.

```
> prop.table(x = table(netRichting), margin = 1)
            richting
onderwijsnet     andere   letteren  soc. wet.       wet.
   katho     0.15384615 0.15384615 0.57692308 0.11538462
   officieel 0.07058824 0.14705882 0.67647059 0.10588235
```

116. Gebruik de functie `prop.table` *met het argument* `margin = 2` *en zonder het argument* `margin`*. Probeer te begrijpen wat* `R` *doet.*

Met het argument `margin = 1` zeggen we `R` dat hij de proporties per rij moet berekenen. Je kan verifiëren dat de som van de proporties in elke rij gelijk is aan 1.

We zien nu dat de proporties per richting niet dezelfde zijn in beide onderwijsnetten. Bv. uit het officieel onderwijs hebben 67.6% (115/170) de wetenschappen gekozen, maar slechts 57.7% (75/130) uit het katholiek onderwijs. Wat moeten we concluderen? Is het aannemelijk zulke verschillen te observeren als de twee steekproeven uit een homogene populatie komen ? Er bestaat een statistische toets om dit na te gaan: de Pearson chi kwadraat toets.

11.2 De Pearson chi kwadraat toets.

k populaties zijn in p categorieën verdeeld. We willen controleren of de proporties in de p categorieën identiek zijn in de k populaties (nulhypothese). In elke populatie i trekken we een steekproef van n_i proefpersonen waar de frequenties in de p categorieën $f_{i,j}$ zijn (de geobserveerde frequenties). De geobserveerde proporties in steekproef i zijn $f_{i,1}/n_i, \ldots, f_{i,p}/n_i$. Indien de geobserveerde proporties niet identiek zijn over de steekproeven heen (dit is meestal het geval) is het mogelijk na te gaan of de verschillen het resultaat van het toeval zijn.

De toets is gebaseerd op een χ^2-verdeelde toetsingsgrootheid:

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{(f_{i,j} - n_i \hat{\pi}_{\cdot j})^2}{n_i \hat{\pi}_{\cdot j}} \sim \chi^2_{(p-1)(k-1)} \tag{11.1}$$

waarbij p het aantal categorieën is en k het aantal populaties. Het symbool $\hat{\pi}_{\cdot j}$ representeert de proportie van individuen in categorie j (over alle steekproeven). Het is een schatting van de theoretische of verwachtte prorportie in categorie j indien de nulhypothese correct is. De schatting is gelijk aan het totaal aantal individuen in categorie j gedeeld door de totale steekproefgrootte. Dus

$$\hat{\pi}_{\cdot j} = \frac{\sum_{i=1}^{k} f_{i,j}}{\sum_{i=1}^{k} n_i}.$$

De `R` functie `chisq.test` laat ons toe om de p-waarde van deze toets te berekenen. Deze functie heeft één argument nodig: een tabel met de frequentieverdelingen in de k populaties. Je hoeft geen argument `alternative` door te geven want deze toets is altijd tweezijdig: de proporties zijn identiek in de k populaties of ze zijn niet identiek.

Voorwaarden. Als p groot is en als één van de steekproeven klein is, dan kan het gebeuren dat de berekening van de p-waarde door de functie `chisq.test` niet precies is. `R` geeft dan een melding.

11.2.1 Vb. Onderwijsnetten

Laten we de functie `chisq.test` gebruiken bij het voorbeeld van de richtingen in de verschillende onderwijsnetten.

```
> chisq.test(x=table(netRichting))

Pearson's Chi-squared test

data:  table(netRichting)
X-squared = 6.0231, df = 3, p-value = 0.1105
```

De p-waarde is 11%, wat betekent dat het plausibel is dat de twee steekproeven uit één homogene populatie getrokken zijn. We aanvaarden dus de nulhypothese, namelijk dat de proporties identiek zijn in de twee onderwijsnetten.

11.2.2 Vb. Invloed van het ras op het vonnis

Heeft het ras van de verdachte een invloed op het vellen van het doodvonnis? Om deze onderzoeksvraag te beantwoorden analyseert een onderzoeker drieduizend gelijkaardige gevallen (volwassen mannelijke moordenaar, één slachtoffer, geen antecedent) in de laatste 10 jaren. Je vindt de gegevens in de data frame `vonnis`.

```
> vonnis
     ras doodvonnis
1      D       Neen
2      B       Neen
3      D       Neen
...   ...        ...
2999   C         Ja
3000   A       Neen
```

De relatieve frequentieverdeling van de twee variabelen krijg je met volgende commando:

```
> prop.table( x = table(vonnis), margin = 1)
   doodvonnis
ras         Ja       Neen
  A 0.01445783 0.98554217
  B 0.01276596 0.98723404
  C 0.01031746 0.98968254
  D 0.01136364 0.98863636
```

117. Gebruik de functie `prop.table` *om de relatieve frequentieverdeling van* `type` *bij mannen en bij vrouwen te bekomen.*

Je ziet dat de relatieve frequentie van het doodvonnis in de steekproef niet dezelfde is bij de vier rassen. Zijn de verschillen groot genoeg om te besluiten dat de kans op het doodvonnis niet dezelfde is in de populatie, naar gelang de ras? We gebruiken de Pearson chi kwadraat toets om deze vraag te beantwoorden.

```
> chisq.test(table(vonnis))

Pearson's Chi-squared test

data:  table(vonnis)
X-squared = 0.76205, df = 3, p-value = 0.8585
```

De p-waarde is groter dan 0.05 en we aanvaarden de nulhypothese: de kans op het doodvonnis is dezelfde bij de vier rassen.

118. Gebruik de data frame `myData` *en ga na of de verdeling van de toevalsvariabele* `opleiding` *identiek is bij de populaties van mannen en vrouwen.*

11.3 De power van de Pearson χ^2 toets

Om de power te berekenen moet je altijd over een specifieke alternatieve hypothese beschikken. Onze nulhypothese houdt in dat de proporties identiek zijn in alle k populaties. Onze specifieke alternatieve hypothese gaat dus bepalen

wat de proporties van de p categorieën zijn in de k populaties. We hebben dus een $(k \times p)$-tabel nodig om de specifieke alternatieve hypothese uit te drukken. Deze tabel gaan we in een kansverdeling omzetten en m.b.v. de functie `ES.w2` zullen we dan een effectgrootte berekenen en we zullen uiteindelijk de functie `pwr.chisq.test` gebruiken om de power te berekenen.

11.3.1 Vb. Onderwijsnetten

Laten we zo'n tabel opstellen met verschillen (tussen rijen) die we graag wensen te detecteren. Dit zal een (2×4)-tabel zijn. Voor de eerste rij kunnen we eventueel de proporties van het katholieke net overnemen (na afronding, voor het gemak). Voor de tweede rij kiezen we waarden die verschillen van de eerste rij zodanig dat we de verschillen als relevant beschouwen. Bijvoorbeeld

	andere	letteren	soc.wet.	wet.
katho	0.15	0.15	0.58	0.12
officieel	0.15	0.10	0.45	0.30

We gaan de overeenkomende tabel in `R` aanmaken, dankzij de functie `matrix`. Deze functie transformeert een vector bestaande uit alle waarden van de tabel in de gewenste tabel.

```
> prop <-matrix(data = c(0.15, 0.15, 0.58, 0.12, 0.15, 0.10, 0.45,
  0.30), nrow=2, byrow=TRUE)
> prop
     [,1] [,2] [,3] [,4]
[1,] 0.15 0.15 0.58 0.12
[2,] 0.15 0.10 0.45 0.30
```

Het argument `data` is een vector, aangemaakt m.b.v. de functie `c`, en bestaat gewoon uit de lijst van alle proporties in de tabel, rij per rij. Het argument `nrow` bepaalt het aantal rijen van de tabel en het argument `byrow` bepaalt hoe de tabel ingevuld moet worden (rij per rij). Indien je het argument `byrow` weg laat, dan gaat `R` ervan uit dat de tabel kolom per kolom ingevuld moet worden.

119. Gebruik de functie `matrix` *zoals hierboven maar zonder het argument* `byrow` *en probeer de output te begrijpen.*

120. Gebruik de functie `matrix` *zoals hierboven maar met het argument* `nrow=4` *en probeer de output te begrijpen.*

We gaan nu de tabel `prop` transformeren in een bivariate kans verdeling, dat is, een tabel waarvoor de som van alle cellen gelijk aan 1 is. Hiervoor delen we elke cel van de tabel door het aantal rijen, dus door 2.

```
> prop.k <- prop/2
> prop.k
      [,1]  [,2]  [,3] [,4]
[1,] 0.075 0.075 0.290 0.06
[2,] 0.075 0.050 0.225 0.15
```

We berekenen nu de corresponderende effectgrootte:

```
> w <- ES.w2(prop.k)
[1] 0.2275419
```

en we geven deze door aan de functie `pwr.chisq.test`:[1]

```
> pwr.chisq.test(w = w, N = 300, df = 3, sig.level = 0.05)

     Chi squared power calculation

              w = 0.2275419
              N = 300
             df = 3
      sig.level = 0.05
          power = 0.9265456

NOTE: N is the number of observations
```

Het argument `df` is het aantal vrijheidsgraden en is gelijk aan $(p-1)(k-1)$. De power bedraagt 93%. Dit is prima!

11.3.2 Vb. Doodvonnis

Wat is de power van de Pearson χ^2 toets bij deze toepassing? Om dit te berekenen moeten we eerst een specifieke alternatieve hypothese opstellen. We moeten dus een tabel opstellen met vier regels (één per ras) en twee proporties in elke rij. De regels mogen niet allemaal identiek zijn. Hoe gaan we de proporties kiezen? In dit geval lijkt het aangewezen om zeer kleine verschillen te beschouwen omdat elk verschil (hoe klein het is) een discriminatie aanwijst en is dus niet acceptabel. We gebruiken bijvoorbeeld

	Ja	Neen
A	0.015	0.985
B	0.014	0.986
C	0.013	0.987
D	0.011	0.989

We maken de overeenkomende tabel aan in R, dankzij de functie `matrix`.

```
> prop <-matrix(data = c(0.0150, 0.9850, 0.014, 0.986, 0.013, 0.987,
  0.011, 0.989), nrow=4, byrow=TRUE)
> prop
      [,1]  [,2]
[1,] 0.015 0.985
[2,] 0.014 0.986
[3,] 0.013 0.987
[4,] 0.011 0.989
```

We transformeren deze tabel in een kansverdeling. Te dien einde delen we elke cel door het aantal rijen, dat is 4.

[1] Let op, bij deze functie wordt de steekproefgrootte aangeduid door `N` en niet door `n`.

```
> prop.k <- prop/4
> prop.k
        [,1]    [,2]
[1,] 0.00375 0.24625
[2,] 0.00350 0.24650
[3,] 0.00325 0.24675
[4,] 0.00275 0.24725
```

We berekenen nu de corresponderende effectgrootte:

```
> w <- ES.w2(prop.k)
> w
[1] 0.01293488
```

en we geven deze door aan de functie `pwr.chisq.test`:

```
> pwr.chisq.test(w = w, N = 3000, df = 3, sig.level = 0.05)

     Chi squared power calculation

              w = 0.01293488
              N = 3000
             df = 3
      sig.level = 0.05
          power = 0.08125159

NOTE: N is the number of observations
```

De power is veel te laag. Hoe groot zou de steekproef moeten zijn om een degelijke power te garanderen?

```
> pwr.chisq.test(w = w, power = 0.9, df = 3, sig.level = 0.05)

     Chi squared power calculation

              w = 0.01293488
              N = 84701.36
             df = 3
      sig.level = 0.05
          power = 0.9

NOTE: N is the number of observations
```

We hebben 84701 individuen nodig. Onze steekproef van 3000 is dus veel te klein.

11.4 Afhankelijkheid van twee categorische variabelen

121. Gebruik de functie `prop.table` *om de bivariate relatieve frequentieverdeling van* `type` *en* `geslacht` *te bekomen.*

In deze paragraaf gaan we zien hoe we het verband tussen twee categorische variabelen kunnen analyseren. Stel bijvoorbeeld dat we willen weten of er een verband is tussen de variabelen opleiding (ped, psy, soc) en geslacht. We kunnen dit gemakkelijk herleiden naar een probleem dat we al vroeger hebben behandeld. We kunnen de twee niveaus van de variabele geslacht als twee populaties beschouwen. Dan gaan we na of de verdelingen van de variabele opleiding in de twee populaties identiek zijn met behulp van de Pearson χ^2 toets (Rubr. 11.2). Zo ja, dan is er geen verband tussen opleiding en geslacht; zo neen, dan is er wel een verband.

We kunnen ook anders werken. We kunnen de drie niveaus van de variabele opleiding beschouwen als drie populaties. Dan gaan we na of de verdelingen van de variabelen geslacht identiek zijn in de drie populaties. Die twee werkwijzen zijn uiteraard equivalent.

We kunnen dus altijd het onderzoek naar een verband tussen twee categorische variabelen herleiden naar een probleem van homogeniteit tussen meerdere populaties. We hebben geen nieuwe toets nodig. In sommige boeken worden twee technieken gepresenteerd : één voor het verschil tussen de verdelingen van één variabele in verschillende populaties en één voor de onafhankelijkheid van twee variabelen in één populatie. Die technieken zijn in feite identiek.

11.4.1 Vb. Geslacht en opleiding

We wensen na te gaan of de variabelen `geslacht` en `opleiding` al dan niet afhankelijk zijn in de populatie van FPPW studenten. Hiervoor gebruiken we de data frame `myData`. Laten we eerst de data bekijken:

```
> table(myData$geslacht,myData$opleiding)

   ped psy soc
  M   7   7   0
  V   3  11   2
> prop.table(table(myData$geslacht,myData$opleiding),margin=1)

      ped    psy    soc
  M 0.5000 0.5000 0.0000
  V 0.1875 0.6875 0.1250
```

Er zijn grote verschillen tussen mannen en vrouwen. Maar de steekproef is klein en het is dus mogelijk dat de verschillen aan het toeval te wijten zijn. We gaan dit na met de Pearson χ^2 toets.

```
> chisq.test(table(myData$geslacht,myData$opleiding))

Pearson's Chi-squared test
```

```
data:  table(myData$geslacht, myData$opleiding)
X-squared = 4.375, df = 2, p-value = 0.1122

Warning message:
In chisq.test(table(myData$geslacht, myData$opleiding)) :
  Chi-squared approximation may be incorrect
```

R geeft een melding omdat de steekproef te klein is. De berekende p-waarde is dus misschien niet correct. Was de p-waarde super klein (bv. 0.000001) of super groot (bv. 0.84) dan zouden we wel iets kunnen besluiten. Maar in dit geval is de p-waarde te dicht bij de 5%-drempel om een besluit te kunnen nemen.

122. Gebruik de functie `chisq.test` *om na te gaan of er een verband is tussen de toevalsvariabelen* `type` *en* `geslacht` *(data frame* `sportData`*). Interpreteer de output.*

11.5 Opmerking betreffende de meetniveaus

Alle technieken die we in dit hoofdstuk hebben gezien zijn gebaseerd enkel op de frequenties en nooit op de waarden van de variabele en ook nooit op hun volgorde. We mogen ze dus gebruiken met variabelen van alle meetniveaus. Niet alleen met categorische variabelen maar ook met variabelen van interval, ratio of absoluut meetniveau. Maar als we variabelen van interval meetniveau (of hoger) hebben, is het meestal beter andere technieken (bv. de t-toets, ANOVA of lineaire regressie) te gebruiken omdat ze een hoger onderscheidingsvermogen hebben.

11.6 Oplossingen

116) Gebruik de functie `prop.table` met het argument `margin = 2` en zonder het argument `margin`. Probeer te begrijpen wat R doet.

Oplossing:

```
> prop.table(x = table(netRichting), margin = 2)
            richting
onderwijsnet    andere  letteren soc. wet.      wet.
   katho     0.6250000 0.4444444 0.3947368 0.4545455
   officieel 0.3750000 0.5555556 0.6052632 0.5454545
```

R heeft de proporties per kolom berekend. De som in elke kolom is 1.

```
> prop.table(x = table(netRichting))
            richting
onderwijsnet     andere   letteren  soc. wet.       wet.
   katho     0.06666667 0.06666667 0.25000000 0.05000000
   officieel 0.04000000 0.08333333 0.38333333 0.06000000
```

R heeft de proporties berekend t.o.v. het totaal aantal studenten in de steekproef. De som van alle cellen in de table is 1.

117) Gebruik de functie `prop.table` om de relatieve frequentieverdeling van `type` bij mannen en bij vrouwen te bekomen.

Oplossing:

```
> prop.table(table(sportData$geslacht, sportData$type), margin = 1)

       andere basketbal    tennis   voetbal   zwemmen
  M 0.1914894 0.1382979 0.1170213 0.3936170 0.1595745
  V 0.1886792 0.1698113 0.2264151 0.2075472 0.2075472
```

118) Gebruik het data frame `myData` en ga na of de verdeling van de toevalsvariabele `opleiding` identiek is bij de populaties van mannen en vrouwen.

Oplossing:

```
> chisq.test(table(myData$geslacht , myData$opleiding ))

Pearson's Chi-squared test

data:  table(myData$geslacht, myData$opleiding)
X-squared = 4.375, df = 2, p-value = 0.1122
```

```
Warning message:
In chisq.test(xtabs(myData, formula =  geslacht + opleiding)) :
  Chi-squared approximation may be incorrect
```

R geeft een waarschuwing. De reden is niet duidelijk, maar het is omdat sommige theoretische frequenties ($n_i \hat{\pi}_{\cdot j}$) te klein zijn. Dit hadden we kunnen verwachten want de steekproef is zeer klein ($n = 30$). We mogen dus niet besluiten.

119) Gebruik de functie `matrix` zoals hierboven maar zonder het argument `byrow` en probeer de output te begrijpen.

Oplossing:

```
> prop <-matrix(data = c(0.15, 0.15, 0.58, 0.12, 0.15, 0.10, 0.45,
  0.30), nrow=2)
> prop
     [,1] [,2] [,3] [,4]
[1,] 0.15 0.58 0.15 0.45
[2,] 0.15 0.12 0.10 0.30
```

`R` heeft de tabel kolom per kolom ingevuld.

120) Gebruik de functie `matrix` zoals hierboven maar met het argument `nrow=4` en probeer de output te begrijpen.

Oplossing:

```
> prop <-matrix(data = c(0.15, 0.15, 0.58, 0.12, 0.15, 0.10, 0.45,
  0.30), nrow=4, byrow = TRUE)
> prop
     [,1] [,2]
[1,] 0.15 0.15
[2,] 0.58 0.12
[3,] 0.15 0.10
[4,] 0.45 0.30
>
```

Vanzelfsprekend.

121) Gebruik de functie `prop.table` om de bivariate relatieve frequentieverdeling van `type` en `geslacht` te bekomen.

Oplossing:

```
> prop.table(table(sportData$geslacht, sportData$type))

    andere basketbal tennis voetbal zwemmen
  M  0.090     0.065  0.055   0.185   0.075
  V  0.100     0.090  0.120   0.110   0.110
```

122) Gebruik de functie `chisq.test` om ta te gaan of er een verband is tussen de toevalsvariabelen `type` en `geslacht`. Interpreteer de output.

Oplossing:

```
> chisq.test(sportData$geslacht,sportData$type)

Pearson's Chi-squared test

data:  sportData$geslacht and sportData$type
X-squared = 10.195, df = 4, p-value = 0.03727
```

De realisatie van de toetsingsgrootheid is gelijk aan 10.195. Indien de nulhypothese (geen verband tussen de twee variabelen) correct is, dan zie je zo'n grote realisatie in slechts 3.7% van de gevallen. Dit is kleiner dan 5% (de significantie) en we verwerpen de nulhypothese.

Deel II

Appendix

Formuleblad

Beschrijvende statistiek

Regressielijn van Y op X: $b_1 = r_{XY}\frac{s_Y}{s_X}$ en $b_0 = \overline{y} - b_1\overline{x}$.

Kansrekenen

$P(B(n,\pi)=k) = \frac{n!}{k!\,(n-k)!}\,\pi^k\,(1-\pi)^{n-k}$.
$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$. $\qquad E(\overline{X}) = E(X), \quad V(\overline{X}) = V(X)/n.$

Puntschatting

$\hat{\sigma}_X^2 = s_X^2 = \frac{1}{n-1}\ \sum_{i=1}^n (x_i-\overline{x})^2 \qquad \widehat{COV}_{XY} = \frac{1}{n-1}\ \sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})$.

Intervalschatting

$X_1,\ldots,X_n \sim N(\mu_X,\sigma_X^2)$ of n groot en onafhank. trekkingen $\Rightarrow \frac{\overline{X}-\mu_X}{S_X/\sqrt{n}} \sim t_{n-1}$.
Betrouwbaarheidsinterval voor μ_X: $[\,\overline{x} \pm t_{n-1;\alpha/2}\,\frac{s_X}{\sqrt{n}}\,]$.
Formule (niet altijd geldig) voor een betrouwbaarheidsinterval voor θ:
$[\,\hat{\theta} \pm t_{l;\alpha/2}\,\mathrm{SE}_Q\,]$.

De statistische toetsen

Het toetsen van een hypothese betreffende μ_X

σ_X bekend: $\overline{X} \sim N(\mu_X,\sigma_X^2/n)$. $\qquad$ σ_X onbekend: $\frac{\overline{X}-\mu_X}{S_X/\sqrt{n}} \sim t_{n-1}$.

Het toetsen van een hypothese betreffende twee verwachtingen

Onafhankelijke steekproeven.

- σ_1 en σ_2 bekend:

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N\left(0, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right).$$

- σ_1 en σ_2 gelijk maar onbekend:

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\ \left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}.$$

- Geen hypothese m.b.t. σ_1 en σ_2:

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t_l, \text{ met } l = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)}+\frac{S_2^4}{n_2^2(n_2-1)}}.$$

$$s^2_{\text{pooled}} = \frac{\text{SS}_X + \text{SS}_Y}{n_1 + n_2 - 2}.$$

Afhankelijke steekproeven: herleiden tot het toetsen van een hypothese betreffende één verwachting.

Het toetsen van een hypothese betreffende een proportie: frequentie $\sim B(n, \pi)$.

Enkelvoudige lineaire regressie

$$V(B_1) = \frac{\sigma^2_\varepsilon}{\text{SS}_X} = \frac{\sigma^2_\varepsilon}{(n-1)s^2_X}. \qquad V(B_0) = \sigma^2_\varepsilon \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s^2_X}\right) = \sigma^2_\varepsilon \left(\frac{1}{n} + \frac{\overline{x}^2}{\text{SS}_X}\right).$$

$$V(\widehat{Y}_i) = \sigma^2_\varepsilon \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)s^2_X}\right) = \sigma^2_\varepsilon \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\text{SS}_X}\right).$$

$$\hat{\sigma}^2_\varepsilon = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{\text{SS}_{\text{Res}}}{n-2}.$$

$$\text{BI voor } \beta_1 : \left[\; b_1 \pm t^{\alpha/2}_{n-2} \sqrt{\frac{\hat{\sigma}^2_\varepsilon}{\text{SS}_X}} \; \right] \qquad \text{BI voor } \beta_0 : \left[\; b_0 \pm t^{\alpha/2}_{n-2} \; \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\text{SS}_X}} \; \right].$$

$$\text{Toetsingsgrootheden} : \frac{B_1}{\sqrt{\frac{\text{SS}_{\text{Res}}}{(n-2)\ \text{SS}_X}}} \sim t_{n-2}\,, \qquad \frac{\text{SS}_{\text{Res0}} - \text{SS}_{\text{Res1}}}{\text{SS}_{\text{Res1}}/(n-2)} \sim F_{1,n-2}.$$

$$R^2 = \text{SS}_{\text{Mod}} \,/\, \text{SS}_{\text{Tot}}\,, \qquad \overline{R}^2 = 1 - (1 - R^2)\ \tfrac{n-1}{n-2},$$

Meervoudige lineaire regressie

$$\hat{\sigma}^2_\varepsilon = \frac{\text{SS}_{\text{Res}}}{n-p-1}.$$

F-toets: model A met k predictoren en model B met $p(> k)$ predictoren:

$$\frac{(\text{SS}_{\text{Res}A} - \text{SS}_{\text{Res}B})/(p-k)}{\text{SS}_{\text{Res}B}/(n-p-1)} \sim F_{p-k,n-p-1}.$$

$$R^2 = \text{SS}_{\text{Mod}} \,/\, \text{SS}_{\text{Tot}}, \qquad \overline{R}^2 = 1 - (1 - R^2)\ \tfrac{n-1}{n-p-1}, \qquad f^2 = \tfrac{R^2_B - R^2_A}{1 - R^2_B}$$

Categorische data-analyse

Toetsingsgrootheid: $X^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{(f_{i,j} - n_i \hat{\pi}_{\cdot j})^2}{n_i \hat{\pi}_{\cdot j}} \sim \chi^2_{(p-1)(k-1)}$.

UNIVERSITEIT GENT

Studiefiche

Vanaf academiejaar 2020-2021

Statistiek II (H001440)

Wegens Covid19 kan mogelijk afgeweken worden van de onderwijs- en evaluatievormen. Dergelijke afwijkingen zullen via Ufora worden gecommuniceerd.

Cursusomvang (nominale waarden; effectieve waarden kunnen verschillen per opleiding)

Studiepunten 6.0 Studietijd 180 u Contacturen 45.0 u

Aanbodsessies en werkvormen in academiejaar 2020-2021

B (semester 2)	Nederlands	Gent	werkcollege: PC-klasoefeningen	7.5 u
			werkcollege: geleide oefeningen	7.5 u
			hoorcollege	30.0 u

Lesgevers in academiejaar 2020-2021

Marchant, Thierry PP01 Verantwoordelijk lesgever

Aangeboden in onderstaande opleidingen in 2020-2021	stptn	aanbodsessie
Bachelor of Science in de psychologie (afstudeerrichting bedrijfspsychologie en personeelsbeleid)	6	B
Bachelor of Science in de pedagogische wetenschappen (afstudeerrichting klinische orthopedagogiek en Disability Studies)	6	B
Bachelor of Science in de psychologie (afstudeerrichting klinische psychologie)	6	B
Bachelor of Science in de psychologie (afstudeerrichting onderwijs)	6	B
Bachelor of Science in de pedagogische wetenschappen (afstudeerrichting pedagogiek en onderwijskunde)	6	B
Bachelor of Science in de pedagogische wetenschappen (afstudeerrichting sociale agogiek)	6	B
Bachelor of Science in de psychologie (afstudeerrichting theoretische en experimentele psychologie)	6	B
Bachelor of Arts in de moraalwetenschappen	6	B
Bachelor of Arts in de wijsbegeerte	6	B
Gemeenschappelijk gedeelte Bachelor of Science in de pedagogische wetenschappen	6	B
Gemeenschappelijk gedeelte Bachelor of Science in de psychologie	6	B
Schakelprogramma tot Master of Science in de psychologie (afstudeerrichting bedrijfspsychologie en personeelsbeleid)	6	B
Schakelprogramma tot Master of Science in de pedagogische wetenschappen (afstudeerrichting klinische orthopedagogiek en Disability Studies)	6	B
Schakelprogramma tot Master of Science in de psychologie (afstudeerrichting klinische psychologie)	6	B
Schakelprogramma tot Master of Science in de pedagogische wetenschappen (afstudeerrichting pedagogiek en onderwijskunde)	6	B
Schakelprogramma tot Master of Science in de psychologie (afstudeerrichting theoretische en experimentele psychologie)	6	B
Schakelprogramma tot Master of Science in het sociaal werk	6	B
Voorbereidingsprogramma tot Master of Science in de psychologie (afstudeerrichting bedrijfspsychologie en personeelsbeleid)	6	B
Voorbereidingsprogramma tot Master of Science in de pedagogische wetenschappen (afstudeerrichting klinische orthopedagogiek en Disability Studies)	6	B
Voorbereidingsprogramma tot Master of Science in de psychologie (afstudeerrichting klinische psychologie)	6	B
Voorbereidingsprogramma tot Master of Science in de pedagogische wetenschappen (afstudeerrichting pedagogiek en onderwijskunde)	6	B
Voorbereidingsprogramma tot Master of Science in de psychologie (afstudeerrichting theoretische en experimentele psychologie)	6	B

Onderwijstalen

Nederlands

Trefwoorden

statistiek

Situering

Dit uitdiepend opleidingsonderdeel sluit aan bij de leerlijn rond onderzoekscompetenties binnen de opleidingen Psychologie en Pedagogische Wetenschappen. Het bouwt verder op het vak Statistiek I. Het doel is om kennis en inzicht te verschaffen in methodologische en data-analytische aspecten van empirisch wetenschappelijk onderzoek. Dit opleidingsonderdeel bouwt mee aan de competenties die het mogelijk maken om empirisch wetenschappelijk onderzoek binnen het vakgebied zelfstandig en kritisch te verwerken en om actief (mee) te werken aan wetenschappelijk onderzoek, o.a. binnen de context van de masterproef.

Inhoud

In dit opleidingsonderdeel komen volgende onderwerpen aan bod:
- Databestand en codeboek
- Parametrische inferentie: betrouwbaarheidsintervallen, two-sample en paired t-test
- Parametrische inferentie: effect sizes, steekproefgrootte en power
- Introductie algemeen lineair model: enkel- en meervoudige lineaire regressie met continue en categorische predictoren
- Introductie categorische data-analyse: toets voor 1 proportie, chi kwadraat toets voor 2 variabelen
- Statistische software R: data manipulatie, beschrijvende statistiek, grafieken, inductieve statistiek

Begincompetenties

Dit opleidingsonderdeel bouwt verder op Statistiek I

Eindcompetenties

1 De onderliggende assumpties van veel gebruikte onderzoeksmethodes kunnen bevragen.
2 Empirisch-analytische, interpretatieve en actiegerichte onderzoeksmethodes adequaat kunnen toepassen.
3 Inzichtelijke kennis hebben van de relevante methoden en technieken voor wetenschappelijk onderzoek en data-analyse.
4 De relevante methoden en technieken voor wetenschappelijk onderzoek en data-analyse correct selecteren in het kader van toegepast en fundamenteel onderzoek.
5 Wetenschappelijke informatie systematisch kunnen verzamelen, opzoeken, interpreteren, integreren en presenteren.

Creditcontractvoorwaarde

Toelating tot dit opleidingsonderdeel via creditcontract is mogelijk mits gunstige beoordeling van de competenties

Examencontractvoorwaarde

De toegang tot dit opleidingsonderdeel via examencontract is open

Didactische werkvormen

Hoorcollege, werkcollege: geleide oefeningen, werkcollege: PC-klasoefeningen

Leermateriaal

Een syllabus is beschikbaar.
Geraamde totaalprijs: 10 EUR
Geen leermateriaal in het Engels beschikbaar ten behoeve van uitwisselingsstudenten

Referenties

- Moore, D.S., McCabe G.P., & Craig, B.A. (2009). Introduction to the practice of statistics. W.H. Freeman, 6th edition.
- Neter, J., Kutner, M.H., Nachtscheim, C.J., Wasserman, W. (2004). Applied linear statistical models. McGraw-Hill, 5th edition.
- Agresti, A. (2007). Introduction to categorical data analysis. Wiley-Interscience.

Vakinhoudelijke studiebegeleiding

Op afspraak.

Evaluatiemomenten

periodegebonden evaluatie

Evaluatievormen bij periodegebonden evaluatie in de eerste examenperiode

Schriftelijk examen met meerkeuzevragen

Evaluatievormen bij periodegebonden evaluatie in de tweede examenperiode

Schriftelijk examen met meerkeuzevragen

Evaluatievormen bij niet-periodegebonden evaluatie

Tweede examenkans in geval van niet-periodegebonden evaluatie

Niet van toepassing

Eindscoreberekening

De periodegebonden evaluatie telt mee voor 100 %.

Bibliografie

R. B. Catell. *Manual for the Catell culture fair intelligence test*. Institute for Personality and Ability Testing, Champaign, IL, 1973.

R. Charafeddine, S. Demarest, S. Drieskens, L. Gisle, J. Tafforeau, and J. Van der Heyden. Highlights of the belgian health interview survey 2008. Technical report, Scientific Institute of Public Health, 2011a.

R. Charafeddine, S. Demarest, S. Drieskens, L. Gisle, J. Tafforeau, and J. Van der Heyden. Highlights of the belgian health interview survey 2008. Technical report, Scientific Institute of Public Health, 2011b.

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.

S. E. Ellis and J. T. Leek. How to share data for collaboration. *PeerJ Preprints*, (5:e3139v5), 2017. URL https://doi.org/10.7287/peerj.preprints.3139v5.

T. Festinger and R. Pratt. Speeding adoptions: an evaluation of the effects of judicial continuity. *Social Work Research*, 26:217—224, 2002.

G. E. Gignac and T. C. Bates. Brain volume and intelligence: The moderating role of intelligence measurement quality. *Intelligence*, 64:18–29, 2017.

S. Kaiser, A. Roth, M. Rentrop, H. C. Friederich, S. Bender, and M. Weisbrod. Intra-individual reaction time variability in schizophrenia, depression and borderline personality disorder. *Brain and Cognition*, 66(1):73 – 82, 2008. ISSN 0278-2626. doi: https://doi.org/10.1016/j.bandc.2007.05.007.

S. Kanazawa. Intelligence and obesity: which way does the causal direction go? *Obesity and nutrition*, 21(5):339–344, 2014.

D. Nolan and T. Speed. *Stat Labs: Mathematical Mathematical Statistics Through Applications*. Springer-Verlag, 2000.

C. K. Sanders. Employment options for low-income women: Microenterprise versus the labor market. *Social Work Research*, 28(2):83–92, 2004.

Ü. Tan, M. Tan, P. Polat, Y. Ceylan, S. Suma, and A. Okur. Magnetic resonance imaging brain size/iq relations in turkish university students. *Intelligence*, 27: 83–92, 1999.

Hieronder een paar interessante boeken. Soms geven ze meer details dan deze cursus, soms minder. Soms leggen ze de nadruk meer op de toepassingen, soms op de theorie. Maar door het feit dat ze een andere aanpak volgen, kunnen ze voor elke student hulpvaardig zijn.

In het Engels, met veel nadruk op R en gratis :

Navarro, D. (2016). *Learning statistics with R: A tutorial for psychology students and other beginners*,
`https://learningstatisticswithr.com`

Om je voorkennis wiskunde op te frissen :

Flohr, R. (2007). *Basiswiskunde voor Statistiek*, Academic Service.

In het Nederlands:

Moore D.S. and McCabe G.P. (2014) *Statistiek in de Praktijk*, Theorieboek en opgavenboek, Academic Service.

Reus, G.-J. and van Buuren, H. (2010). *Basisvaardigheden Toegepaste Statistiek*. Noordhoff Uitgevers bv, Groningen/Houten.

Slotboom A. (1996) *Statistiek in woorden : een gebruiksvriendelijke beschrijving van de meest voorkomende statistische termen en technieken*, Wolters-Noordhoff.

Van den Brink W.P. and Koele P. (1986) *Statistiek*, deel I en II, Boom.

Index