Cursus Onderzoekstechnieken

Academiejaar 2017-2018

Dr. Jens Buysse, Wim De Bruyn, Wim Goedertier, Bert Van Vreckem

HoGent
BEDRIJF
EN
ORGANISATIE

WWW.HOGENT.BE

Gegenereerd op 17 mei 2018

Inhoudsopgave

Voorwoord

Deze cursus werd geschreven in het kader van de lessenreeks Onderzoekstechnieken aan de Hogeschool Gent. Ik wil hierbij gebruik maken om volgende mensen te bedanken bij het nakijken en verbeteren van de cursus.

- Cédric Berlez
- Jürgen Van Meerhaeghe
- Gianni Stubbe
- Jelle Elaut
- Thijs Van Der Burgt
- Lotte Potthé
- Özgür Akin
- Cedric De Vylder

Jens Buysse
08 februari 2016

1. Aan de slag

1.1 Studiewijzer

De studiewijzer geeft een overzicht van de belangrijkste informatie over deze cursus, o.a. leerdoelen, lesmateriaal, weekplanning en leeraanwijzingen. Lees alles aandachtig door!

1.1.1 Doel en plaats van de cursus in het curriculum

Deze cursus is een inleiding op wat tegenwoordig vaak *data science* genoemd wordt. Het doel is om je wegwijs te maken in het correct verzamelen, verwerken en analyseren van numerieke data en daar een onderzoeksverslag over te schrijven.

In de eerste plaats is dit een voorbereiding op de bachelorproef, waar je deze technieken in de praktijk zal moeten omzetten. Maar ook na je afstuderen blijft de kennis die je in deze cursus opdoet waardevol. Succesvolle bedrijven nemen beslissingen, niet op basis van buikgevoel of intuïtie, maar door het verzamelen van data. Aan de hand van de technieken die hier toegelicht worden, heb je voldoende achtergrond om vragen te beantwoorden als:

- Is een (web)applicatie snel genoeg voor de gebruikers? Is de gebruikerservaring consistent, of zit er grote variatie op responstijden?
- Welk van twee systemen (software of hardware) is het meest performant? Is het verschil tussen beide significant, of kunnen verschillen in de metingen te wijten zijn aan het toeval?
- Wanneer moeten aankopen van nieuwe apparatuur (bv. harde schijven, servers, geheugen, enz.) ingepland worden, op basis van historische gebruiksgegevens?

1.1.2 Leerdoelen en competenties

- Kan begrippen, formules, stellingen en de uitwerking ervan uit de beschrijvende en inductieve statistiek benoemen en verklaren
- Kan formules, stellingen uit de beschrijvende en inductieve statistiek in onderzoeksvraagstukken correct toepassen
- Kan data analyseren met statistische software
- Kan een gestructureerd wetenschappelijk document schrijven en voorzien van referenties in LaTeX
- Kan de wetenschappelijke methode vergelijken met niet-wetenschappelijke onderzoeksmethodes en daarbij voor- en nadelen opsommen

Deze vind je ook terug in de studiefiche.

1.1.3 Leerinhoud

Verder in dit hoofdstuk vind je instructies voor het installeren van de nodige software, en een korte inleiding op het werken met R, een programmeertaal voor data-analyse.

Hoofdstuk 2 geeft een inleiding op het verloop van een typisch onderzoeksproces en introduceert enkele basisconcepten van data-analyse.

Hoofdstuk 3 behandelt de analyse van een enkele variabele, meer bepaald centrum- en spreidingsmaten, en ook geschikte grafiektypes voor elk soort variabelen.

Hoofdstuk 4 introduceert het concept van het nemen van steekproeven uit een populatie, en de randvoorwaarden waaronder resultaten binnen een steekproef kunnen veralgemeend worden tot de gehele populatie.

Hoofdstuk 5 gaat hierop verder met de algemene werkwijze voor het voeren van statistische toetsen, en specifiek met toetsen voor uitspraken over het gemiddelde van een populatie: de z-toets en de t-toets.

Waar de vorige hoofdstukken telkens één variabele apart beschouwden, bekijkt Hoofdstuk 6 verschillende technieken om verbanden tussen twee variabelen te leggen, afhankelijk van het variabeletype.

Hoofdstuk 7 introduceert de χ^2-toets, waarmee je kan nagaan of de verdeling van een steekproef relevant is voor een populatie, of in hoeverre twee steekproeven een gelijkaardige verdeling hebben.

Hoofdstuk 8 geeft een inleiding op het analyseren van hoe de waarde van een variabele evolueert in de tijd aan de hand van wiskundige modellen die onder bepaalde voorwaarden ook toelaten om voorspellingen te doen.

1.1.4 Leermateriaal

Het belangrijkste leermateriaal voor dit opleidingsonderdeel is deze cursus, die ook de oefeningenopgaven bevat. Die wordt ter beschikking gesteld via Chamilo als PDF. Op Chamilo vind je ook de PDF's met de slides gebruikt tijdens de lessen.

Daarnaast krijgen studenten toegang tot een GitHub-repository met de broncode voor:

- Deze cursus
- De slides van lessen
- Broncodevoorbeelden in R voor alle technieken die in de cursus aan bod komen.

Errata en wijzigingen aan de cursus worden in GitHub aangebracht. De PDF's op Chamilo zullen niet noodzakelijk bijgewerkt worden. Studenten kunnen zelf de laatste versies van alle documenten met LaTeX genereren.

De software die nodig is voor dit opleidingsonderdeel is gratis/open source. Instructies voor de installatie kan je vinden in Sectie 1.2.

1.1.5 Werkvormen

Studenten afstandsleren kunnen vragen stellen tijdens de contactmomenten. Dit zijn echter geen lesmomenten! Het rooster vind je in de Chamilo-cursus "Informatie voor studenten TILE."

Studenten dagonderwijs krijgen één uur per week hoorcollege en twee uur werkcollege.

1.1.6 Werk- en leeraanwijzingen

Het opleidingsonderdeel *Onderzoekstechnieken* wordt door veel studenten als moeilijk ervaren. Dat is begrijpelijk, want het onderwerp ligt dan ook buiten de comfortzone van de doorsnee informatica-student en we weten allemaal dat wiskundige vakken niet de populairste van onze opleiding zijn.

Er zijn twee manieren om hier mee om te gaan. Je kan de weg van de minste weerstand nemen: je concentreren op de vakken die je graag doet en een dag voor het examen de cursus doornemen in de hoop dat je voldoende punten bij elkaar sprokkelt om een tien te halen. De ervaring leert dat deze strategie niet succesvol is, wat blijkt uit het lage slagingspercentage in de eerste zittijd (in academiejaar 2016-2017 was dat ca. 35% voor het dagonderwijs en 10% voor afstandsleren).

Enkele tips om wél meteen te slagen voor dit vak:

- Kom naar de theorieles en *neem actief nota's*;
- Werk ook voor dit vak *buiten de contactmomenten*. Herhaal de geziene theorie en werk oefeningen af waarmee je nog niet klaar was. Noteer zaken die je niet snapt of waar je vast zit, en stel je vraag bij het eerstvolgende werkcollege.
- Gebruik goede *leertechnieken*. Je vindt een goed overzicht van leertechnieken waarvan het effect wetenschappelijk aangetoond is via de website van *The Learning Scientists*[1].
  - *Spaced practice:* Studeer in meerdere kleine sessies (minstens één keer per week) en niet in grote blokken. Blokkeer een vast moment in je weekagenda/lesplanning.
  - *Retrieval practice:* Neem een leeg blad papier en probeer zoveel mogelijk zaken over een bepaald onderwerp op te schrijven vanuit je herinnering (dus zonder in de cursus te kijken). Controleer dit daarna aan de hand van je lesnota's en in de cursus.
  - *Elaboration:* Stel jezelf vragen over hoe dingen (bv. formules, toetsingsprocedures, ...) in elkaar zitten en waarom dat zo is. Overleg met medestudenten. Vraag je lector

[1] http://www.learningscientists.org/

Week	Theorie	Oefeningen
1	Intro, Onderzoeksproces	Software installeren, LaTeX
2	Analyse van 1 variabele	Wetenschappelijk schrijven
3	Steekproefonderzoek	Analyse van 1 variabele
4	Steekproefonderzoek	Steekproefonderzoek
5	Toetsingsprocedures (z-toets)	Steekproefonderzoek
6	Toetsingsprocedures (t-toets)	Toetsingsprocedures
7	Analyse van 2 variabelen	Toetsingsprocedures
—	**Paasvakantie**	—
8	Analyse van 2 variabelen	Analyse van 2 variabelen
9	χ^2-toets	Analyse van 2 variabelen
10	Tijdreeksen	χ^2-toets
11	Toelichting bachelorproef	Tijdreeksen
12	Herhaling	Herhaling

Tabel 1.1: Weekplanning van de cursus.

om meer uitleg indien nodig. Leg verbanden tussen verschillende onderwerpen in de cursus (bv. vergelijk toetsingsprocedures).
- *Interleaving:* Wissel onderwerpen af tijdens het studeren.
- Gebruik *concrete voorbeelden* om abstracte ideeën te begrijpen. In de cursus worden voorbeelden gegeven, probeer er zelf te bedenken. Overleg met medestudenten en vraag eventueel feedback aan je lector.
- *Dual coding:* Combineer woord en beeld, probeer de leerstof die je instudeert visueel voor te stellen.

Uiteindelijk komt het er op neer dat je voldoende tijd en inspanning investeert om te studeren voor dit vak.

1.1.7 Studiebegeleiding en planning

Studenten **afstandsleren** die vragen hebben over de leerstof kunnen in de eerste plaats terecht op het forum in Chamilo. Wanneer je een oefening gemaakt hebt en twijfelt over de correcte oplossing, kan je de lector per mail contacteren. Zet dan in de onderwerpregel "[OZT][TILE]". Deze mails worden niet dagelijks beantwoord, dus het kan even duren voordat je reactie krijgt.

Studenten **dagonderwijs** kunnen vragen stellen tijdens de werkcolleges, of ook op het forum.

In Tabel 1.1 vind je een overzicht van de lesplanning voor het dagonderwijs die ook als leidraad kan dienen voor de studieplanning van studenten afstandsleren.

1.1.8 Evaluatie

Dagonderwijs

- Eerste examenperiode:
  - 70% periodegebonden evaluatie: schriftelijk examen, bestaande uit een deel gesloten

boek (theorie) en een deel met voorbereiding op pc (oefeningen)
  - 30% niet-periodegebonden evaluatie: het voeren van een mini-onderzoek in groep, bestaande uit een literatuurstudie, opzetten van een reproduceerbaar experiment, verzamelen van meetgegevens en die statistisch analyseren, en er een verslag over schrijven
- Tweede examenperiode:
  - 70% periodegebonden evaluatie: schriftelijk examen, bestaande uit een deel gesloten boek (theorie) en een deel met voorbereiding op pc (oefeningen)
  - 30% niet-periodegebonden evaluatie: er wordt geen tweede examenkans georganiseerd voor dit onderdeel. Wanneer een student in de eerste examenkans niet geslaagd was voor de opdracht blijft de beoordeling voor deze evaluatievorm of de afwezigheid voor deze evaluatievorm geldig voor de tweede examenkans.

Afstandsleren

- Eerste examenperiode:
  - 70% periodegebonden evaluatie: schriftelijk examen, bestaande uit een deel gesloten boek (theorie) en een deel met voorbereiding op pc (oefeningen)
  - 30% niet-periodegebonden evaluatie: individuele opdracht, schrijven van een paper
- Tweede examenperiode:
  - 70% periodegebonden evaluatie: schriftelijk examen, bestaande uit een deel gesloten boek (theorie) en een deel met voorbereiding op pc (oefeningen)
  - 30% niet-periodegebonden evaluatie: er wordt geen tweede examenkans georganiseerd voor dit onderdeel. Wanneer een student in de eerste examenkans niet geslaagd was voor de opdracht blijft de beoordeling voor deze evaluatievorm of de afwezigheid voor deze evaluatievorm geldig voor de tweede examenkans.

1.2 Installatie software

Voor de cursus onderzoekstechnieken maak je gebruik van verschillende softwarepakketten. Hier vind je wat uitleg over de installatie en hoe je er mee aan de slag kan.

- Git client (versiebeheersysteem);
- LaTeX compiler;
- LaTeX editor;
- Jabref (bibliografische databank);
- R (statistische analysesoftware);
- Rstudio (IDE voor R).

Sommige van deze applicaties nemen veel schijfruimte in, dus zorg dat je voldoende ruimte vrij hebt.

In vele andere cursussen rond statistiek of onderzoekstechnieken wordt gebruik gemaakt van commerciële software: SPSS of SAS voor data-analyse, MS Office voor de opmaak van documenten. In deze cursus wordt er expliciet voor gekozen om open source of gratis software te gebruiken. Het grootste voordeel is dat je die ook na je afstuderen nog kan gebruiken zonder dat jij of je bedrijf/organisatie softwarelicenties moet aankopen.

Bovendien zijn de tools die we zullen gebruiken kwalitatief minstens even goed dan hun commerciële tegenhangers. R, een programmeertaal voor statistische analyse, wordt wereldwijd gebruikt in

academische én professionele context. Volgens de TIOBE-index[2] zit R intussen bijna in de top-10 van alle programmeertalen en de taal zit sinds een vijftal jaar in een vrij sterk stijgende trend. De kans is dus niet onbestaande dat je het in je professionele loopbaan nog zal tegenkomen, of het zal kunnen toepassen voor het oplossen van datagerelateerde problemen. Feedback die we kregen van oud-studenten bevestigt dit.

LaTeX is een markuptaal en tekstzetsysteem voor de professionele vormgeving van documenten. De bedoeling is dat de auteur zich vooral moet bezig houden met het logisch structureren van een tekst, en dat het vormgeven op papier wordt overgenomen door de software. Het aanleren van de markuptaal vraagt wat inspanning, maar het is een investering die rendeert wanneer je een lang document (zoals een scriptie) op een professionele, strakke manier wil opmaken. Er zijn in het verleden nog zelden of nooit bachelorproeven ingediend die in MS Word geschreven waren en die een voldoende goede opmaak hadden. Het lijkt veel eenvoudiger om een tekst op te stellen in Word, maar het is zo goed als onmogelijk om in een lang document een consistente en professioneel ogende opmaak te realiseren.

1.2.1 Windows

Omdat het hier toch gaat om een vrij groot aantal applicaties, kunnen Windows-gebruikers beter gebruik maken van de Chocolatey package manager[3] in plaats van alles manueel te downloaden en installeren.

Na installatie van Chocolatey[4], voer je volgende commando's uit als Administrator in een CMD of PowerShell terminal:

```
choco install -y git
choco install -y miktex
choco install -y texstudio
choco install -y JabRef
choco install -y r.project
choco install -y r.studio
```

Wie toch de "klassieke" werkwijze wil hanteren vindt hier de verschillende softwarepakketten:

- Git client: https://git-scm.com/download/win
- LaTeX compiler: https://miktex.org/download
- TeXStudio: http://www.texstudio.org/
- Jabref: https://www.fosshub.com/JabRef.html
- R: https://lib.ugent.be/CRAN/
- Rstudio: https://www.rstudio.com/products/rstudio/download/#download

[2] **\url {https://www.tiobe.com/tiobe-index/}**.
[3] https://chocolatey.org/
[4] https://chocolatey.org/install

1.2.2 macOS

macOS gebruikers installeren de nodige software best via de Homebrew[5] package manager[6]:

```
brew install git
brew cask install mactex
brew cask install texstudio
brew cask install jabref
brew install Caskroom/cask/xquartz
brew install --with-x11 r
brew cask install --appdir=/Applications rstudio
```

Wie toch alles manueel wil installeren kan de applicaties hier downloaden:

- Git client: https://git-scm.com/download/mac
- LaTeX compiler: https://www.tug.org/mactex/mactex-download.html
- TeXStudio: http://www.texstudio.org/
- Jabref: https://www.fosshub.com/JabRef.html
- R: https://lib.ugent.be/CRAN/
- Rstudio: https://www.rstudio.com/products/rstudio/download/#download

1.2.3 Linux

Op RStudio na zijn alle nodige softwarepakketten beschikbaar in de repositories van de meest gebruikte Linux-distributies. We geven hier command-line instructies voor enerzijds Ubuntu (Xenial/16.04) en Debian 9 en anderzijds Fedora.

Ubuntu/Debian

Controleer eerst de link naar de laatste versie van RStudio via de website.

```
sudo aptitude update
sudo aptitude install texlive-latex-base texlive-latex-extra \
  texlive-lang-european texlive-bibtex-extra texlive-extra-utils \
  biber git texstudio jabref r-base
wget https://download1.rstudio.org/rstudio-xenial-1.1.414-amd64.deb
sudo dpkg -i ./rstudio-xenial-1.1.414-amd64.deb
```

Fedora

Controleer eerst de link naar de laatste versie van RStudio via de website. Dit is één lang commando:

```
sudo dnf install git texstudio R \
```

[5] https://brew.sh/

[6] **Let op!** Deze werkwijze is nog niet getest. Feedback van Mac-gebruikers is welkom!

```
java-1.8.0-openjdk-openjfx texlive-collection-latex \
texlive-texliveonfly texlive-babel-dutch \
https://download1.rstudio.org/rstudio-1.1.414-x86_64.rpm
```

Je kan JabRef ook installeren vanuit de Fedora package repository, maar dan krijg je een verouderde versie. Je kan dan beter de "Platform Independent Runnable Jar" downloaden via de projectwebsite[7]. Die kan je dan opstarten vanuit de shell met het commando (hier voorbeeld voor versie 4.1):

```
java -jar JabRef-4.1.jar
```

1.3 Configuratie

1.3.1 Git, GitHub

Wellicht heb je Git al geconfigureerd voor enkele van je andere vakken. Kijk eventueel alles nog eens na! Als alles ok is, kan je deze sectie overslaan.

Wij raden aan om Git via de command line te gebruiken. Zo krijg je het beste inzicht in de werking. Het commando `git status` geeft op elk moment een goed overzicht van de toestand van je lokale repository en geeft aan met welk commando je een stap verder kan zetten of de laatste stap ongedaan kan maken.

Als je nog geen GitHub-account hebt, kies dan een gebruikersnaam die je na je afstuderen nog kan gebruiken (dus bv. niet je HoGent login). De kans is erg groot dat je tijdens je carrière nog van GitHub gebruik zult maken. Koppel ook je HoGent-emailadres aan je GitHub account (je kan meerdere adressen registreren). Op die manier kan je aanspraak maken op het GitHub Student Developer Pack[8], wat je gratis toegang geeft tot een aantal in principe betalende producten en diensten.

Windows-gebruikers voeren volgende instructies uit via Git Bash, macOS- en Linux-gebruikers via de standaard (Bash) terminal.

```
git config --global user.name 'Pieter Stevens'
git config --global user.email 'pieter.stevens.u12345@student.hogent.be'
git config --global push.default simple
```

Maak ook een SSH-sleutel aan om het synchroniseren met GitHub te vereenvoudigen (je moet dan geen wachtwoord meer opgeven bij push/pull van/naar een private repository).

```
ssh-keygen
```

Volg de instructies op de command-line, druk gewoon ENTER als je gevraagd wordt een wachtwoordzin (pass phrase) in te vullen. In de home-directory van je gebruiker (bv. `c:\Users\Pieter`

[7] https://jabref.org/
[8] https://education.github.com/pack

op Windows, `/Users/pieter` op Mac, `/home/pieter` op Linux) is nu een directory met de naam .ssh/ aangemaakt met twee bestanden: `id_rsa` (je private key) en `id_rsa.pub` (je public key). Open dit laatste bestand met een teksteditor en kopieer de volledige inhoud naar het klembord. Ga vervolgens naar je GitHub profiel en kies in het menu links voor SSH and GPG keys. Klik rechtsboven op de groene knop met "New SSH Key" en plak de inhoud van je publieke sleutel in het veld "Key". Bevestig je keuze.

Test nu of je de code van de cursus Onderzoekstechnieken kan downloaden. Ga in de Bash shell naar een directory waar je dit project lokaal wil bijhouden en voer uit:

```
git clone git@github.com:HoGentTIN/onderzoekstechnieken-cursus.git
```

Als dit lukt, is er nu een directory aangemaakt met dezelfde naam als de repository. Doe tijdens het semester regelmatig `git pull` om de laatste wijzigingen in het cursusmateriaal bij te werken. Pas zelf geen bestanden aan binnen deze repository, dit zal leiden tot conflicten.

1.3.2 TeXstudio

Controleer deze instellingen via menu-item *Options* > *Configure TeXstudio*:

- Build
  - Default Compiler: pdflatex
  - Default Bibliography tool: biber
- Editor:
  - Indentation mode: Indent and Unindent Automatically
  - Replace Indentation Tab by Spaces: Aanvinken
  - Replace Tab in Text by spaces: Aanvinken
  - Replace Double Quotes: English Quotes: ‘‘’’

Om te testen of TeXstudio goed werkt, kan je het bestand `cursus/cursus-onderzoekstechnieken.tex` openen. Kies *Tools* > *Build & View* (of druk F5) om de cursus te compileren in een PDF-bestand.

Veel functionaliteiten van LaTeX zitten in aparte packages die niet noodzakelijk standaard geïnstalleerd zijn. De eerste keer dat je een bestand compileert, is het dan ook mogelijk dat er extra packages moeten gedownload worden. MiKTeX zal een pop-up tonen om je toestemming te vragen, bevestig dit. Op Linux is het mogelijk dat je deze packages nog manueel moet installeren. De eerste keer compileren kan enkele minuten duren zonder dat je feedback krijgt over wat er gebeurt. Even geduld, dus!

Indien er zich fouten voordoen bij de compilatie, kan je onderaan in het tabblad Log een overzicht krijgen van de foutboodschappen.

1.3.3 JabRef

JabRef[9] is een GUI voor het bewerken van BibTeX-bestanden, een soort database van bronnen uit de wetenschappelijke- of vakliteratuur voor een LaTeX-document.

[9] http://www.jabref.org/

Kies in het menu voor *File* > *Switch to BibLaTeX* mode. Dit maakt de bestandsindeling van de bibliografische databank compatibel met dat van de cursus en het aangeboden LaTeX-sjabloon voor de bachelorproef.

Kies in het *Preferences*-venster voor de categorie *File* en geef een directory op voor het bijhouden van PDFs van de gevonden bronnen onder *Main file directory*. Het is heel interessant om alle gevonden artikels te downloaden en onder die directory bij te houden. Nog beter is om als naam van het bestand de BibTeX key te nemen (typisch naam van de eerste auteur + jaartal, bv. `Knuth1998.pdf`). Je kan het bestand dan makkelijk openen vanuit JabRef.

Voor meer gedetailleerde informatie over het bijhouden van bibliografische referenties, zie de bachelorproefgids (**VanVreckem2017**).

1.4 Gebruik van R

R is een softwareprogramma voor datamanipulatie, berekening en het grafisch voorstellen van data. Het heeft onder meer:

1. een effectieve gegevensbeheer- en opslagfaciliteit,
2. een reeks operatoren voor berekeningen op arrays, in het bijzonder matrices,
3. een grote verzameling van instrumenten voor data-analyse,
4. grafische faciliteiten voor data-analyse en weergave en
5. een goed ontwikkelde, eenvoudige en effectieve programmeertaal (genaamd 'S').

R heeft een ingebouwde hulpfaciliteit die vergelijkbaar is met die van UNIX man-pages. Voor meer informatie over elke specifieke functie, bijvoorbeeld `solve`, kan je volgende commando oproepen

```
> help(solve)
```

Een alternatief is

```
> ?solve
```

1.4.1 Commando's opslaan en output uitvoeren

Als de commando's in een extern bestand worden opgeslagen, bv. `commands.R` in de werkmap, dan kunnen deze op elk moment uitgevoerd worden in een R-sessie met de opdracht

```
> source("commands.R")
```

De functie `sink`,

```
> sink("record.lis")
```

Zal alle volgende uitvoer van de console naar een extern bestand, `record.lis`, wegschrijven. Het bevel

```
> sink()
```

Herstelt de output opnieuw naar de console.

1.4.2 R omgeving en workspace

De entiteiten die R creëert en manipuleert staan bekend als objecten. Deze kunnen variabelen zijn, arrays van cijfers, reeksen, functies of meer algemene structuren die uit dergelijke componenten zijn gebouwd. Tijdens een R-sessie worden objecten gemaakt en opgeslagen op naam. Het R commando

```
> objects()
```

geeft een overzicht van alle objecten die gemaakt zijn tot op dat moment. De verzameling van objecten die momenteel zijn opgeslagen, heet de werkruimte. Om objecten te verwijderen is de functie `rm` beschikbaar:

```
> rm(x, y, z, inkt, junk, temp, foo, bar)
```

Alle objecten die tijdens een R-sessie zijn aangemaakt, kunnen permanent in een bestand worden opgeslagen voor gebruik in de toekomstige R sessies. Als u aangeeft dat u dit wilt doen, worden de objecten geschreven naar een bestand met extensie `.RData`

In dit hoofdstuk onderzoeken we hoe je een dataset definieert in R. Er worden slechts twee commando's onderzocht. De eerste is voor het eenvoudig toewijzen van gegevens, en de tweede is voor het lezen in een databestand. Er zijn verschillende manieren om gegevens in een R-sessie te lezen, maar we richten ons op slechts twee om het eenvoudig te houden.

1.4.3 Toewijzing

De meest directe manier om een lijst met nummers op te slaan is via een opdracht met behulp van het `c`-commando. (C staat voor "combineren.") Het idee is dat een lijst met nummers onder een bepaalde naam wordt opgeslagen, en de naam wordt gebruikt om te verwijzen naar de gegevens. Een lijst wordt gespecificeerd met de opdracht `c`, en de toewijzing wordt geduid met de symbolen "<-". Een andere term die gebruikt wordt om de lijst met nummers te omschrijven is `vector`.

De cijfers binnen de `c`-opdracht worden gescheiden door komma's. Als voorbeeld kunnen we een nieuwe variabele maken, genaamd "`x`".

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

Wanneer je dit commando invoert, mag je geen uitvoer zien behalve een nieuwe opdrachtregel. Het commando maakt een lijst met nummers genaamd "x". Om te zien welke elementen zijn opgenomen in x, typ zijn naam en druk op de enter-toets.

Als u met één van de nummers wilt werken, kunt u hier toegang krijgen tot de variabele en vervolgens vierkante haakjes noteren die aangeven welk nummer u wilt beschouwen:

```
> x[2]
[1] 5.6
```

1.4.4 Een csv file lezen

We gaan ervan uit dat het gegevensbestand een csv bestand is: "komma-gescheiden waarden"(csv). Dat wil zeggen, elke regel bevat een rij met waarden die getallen of letters kunnen zijn, en elke waarde wordt gescheiden door een komma. We gaan ervan uit dat de eerste rij een lijst met labels bevat. Het idee is dat de labels in de bovenste rij gebruikt worden om te verwijzen naar de verschillende variabelen per rij.

Het commando om het gegevensbestand te lezen is `read.csv`. We moeten tenminste één argument geven aan de opdracht.

> **Oefening 1.1.** *Ga met het help commando na wat de parameters zijn van het commando. Probeer daarna het bestand* `computers.csv` *in te lezen.* ■

Als u niet zeker bent welke bestanden in de huidige werkmap zitten, kunt u het commando `dir` gebruiken om een lijst van bestanden weer te geven. Het `getwd` commando gebruikt u om de huidige werkmap te bepalen.

Het databestand komt uit de publicatie van (**Stengos2005**). Deze dataset bevat data van 1993 tot 1995 over de prijzen van computers. Je kan nagaan wat het effect van de toevoeging van cd-rom-station is op de prijs van de computer of het effect van de kloksnelheid op de prijs.

```
> dir()
[1] "breakingbad.csv"   "Desktop"           "Documents"          "
    Downloads"          "dumps"             "earch php~"         "
    examples.desktop"
[8] "f.r"               "kids.csv"          "kmissles.csv"       "
    kmissles.ods"       "Music"             "out.pdf"            "
    Pictures"
[15] "public"           "Public"            "R"                  "
    Templates"          "test"              "test.php"           "
    Videos"
> getwd()
[1] "/home/eothein"
```

Als u niet zeker weet welke kolommen gedefinieerd zijn, kunt u `names()` gebruiken:

```
> names(computers)
[1] "price"   "speed"   "hd"      "ram"     "screen"  "cd"
    "multi"   "premium" "ads"     "trend"
```

Wanneer u het commando `read.csv` gebruikt, gebruikt R een specifiek soort variabele, dat een dataframe heet. Alle gegevens worden opgeslagen in het dataframe als afzonderlijke kolommen. Als u niet zeker weet wat voor variabele u hebt, dan kunt u de opdracht `attributes` gebruiken. Hiermee worden alle dingen vermeld die R gebruikt om de variabele te beschrijven:

```
attributes(computers)
$names
[1] "price"   "speed"   "hd"      "ram"     "screen"  "cd"
    "multi"   "premium" "ads"     "trend"
```

```
4
5  $class
6  [1] "tbl_df"     "tbl"        "data.frame"
7
8  $row.names
9  [1]   1   2   3   4   5   6   7   8   9  10  11  12
         13  14  15  16  17  18  19  20  21  22  23  24
         25  26  27
10 [28]  28  29  30  31  32  33  34  35  36  37  38  39
         40  41  42  43  44  45  46  47  48  49  50  51
         52  53  54
11 ...
12 [ reached getOption("max.print") -- omitted 5259 entries ]
13
14 $spec
15 cols(
16 price = col_integer(),
17 speed = col_integer(),
18 hd = col_integer(),
19 ram = col_integer(),
20 screen = col_integer(),
21 cd = col_character(),
22 multi = col_character(),
23 premium = col_character(),
24 ads = col_integer(),
25 trend = col_integer()
26 )
```

1.4.5 Data types

We kijken naar enkele manieren waarop R gegevens kan opslaan en organiseren. Dit is echter een inleiding dus beschouwen we maar een kleine subset van de verschillende datatypes die door R worden herkend.

Numbers

De meest eenvoudige manier om een nummer op te slaan is om een variabele van een enkel getal te nemen:

```
1 > a <- 3
2 >
```

Hiermee kunt u allerlei basisoperaties doen en opslaan:

```
1 > b <- sqrt(a*a+3)
2 > b
3 [1] 3.464102
```

Als u een lijst met nummers wilt initialiseren, kan het `numeric` commando worden gebruikt. Om bijvoorbeeld een lijst van 10 nummers te maken, gebruikt u de volgende opdracht. Je kan ook kijken naar het type van de variabele.

```
> a <- numeric(10)
> a
[1] 0 0 0 0 0 0 0 0 0 0
> typeof(a)
[1] "double"
```

Strings

Een tekenreeks wordt gespecificeerd door gebruik te maken van aanhalingstekens. Zowel enkelvoudige als dubbele aanhalingstekens zullen werken:

```
> a <- "hello"
> a
[1] "hello"
> b <- c("hello","there")
> b
[1] "hello" "there"
> b[1]
[1] "hello"
```

Factors

Vaak bevat een experiment proeven voor verschillende niveaus van een verklarende variabele. Bijvoorbeeld een nominale variabele die gecodeerd wordt met een integer. De verschillende niveaus worden ook factoren genoemd.

Je geeft aan dat een variabele een factor is met behulp van het `factor` commando.

Data frames

Data kan worden opgeslagen aan de hand van dataframes. Dit is een manier om verschillende vectoren van verschillende types te nemen en ze op te slaan in dezelfde variabele. De vectoren kunnen van alle soorten zijn. Een dataframe kan bijvoorbeeld verschillende vectoren bevatten en elke lijst kan een vector zijn van factoren, strings of nummers.

Er zijn verschillende manieren om gegevensframes te maken en te manipuleren. De meeste zijn buiten het bereik van deze introductie. Ze worden hier alleen genoemd om een meer volledige beschrijving te geven.

```
> a <- c(1,2,3,4)
> b <- c(2,4,6,8)
> levels <- factor(c("A","B","A","B"))
> bubba <- data.frame(first=a,
                      second=b,
                      f=levels)
```

```
> bubba
  first second f
1     1      2 A
2     2      4 B
3     3      6 A
4     4      8 B
> summary(bubba)
     first           second     f
Min.   :1.00   Min.   :2.0   A:2
1st Qu.:1.75   1st Qu.:3.5   B:2
Median :2.50   Median :5.0
Mean   :2.50   Mean   :5.0
3rd Qu.:3.25   3rd Qu.:6.5
Max.   :4.00   Max.   :8.0
> bubba$first
[1] 1 2 3 4
> bubba$second
[1] 2 4 6 8
> bubba$f
[1] A B A B
Levels: A B
```

Logische variabelen

Een ander belangrijk gegevenstype is het logische type. Er zijn twee vooraf gedefinieerde variabelen, `TRUE` en `FALSE`.

Tables

Een andere manier om informatie op te slaan is in een tabel. We kijken alleen maar naar het maken en definiëren van tabellen.

```
> a <- factor(c("A","A","B","A","B","B","C","A","C"))
> results <- table(a)
> results
a
A B C
4 3 2
> attributes(results)
$dim
[1] 3

$dimnames
$dimnames$a
[1] "A" "B" "C"


$class
[1] "table"
```

```
> summary(results)
Number of cases in table: 9
Number of factors: 1
```

Als je rijen wilt toevoegen aan uw tabel, voeg dan nog een vector toe als argument van de tabelopdracht. In het onderstaande voorbeeld hebben wij twee vragen. In de eerste vraag staan de reacties 'Never', 'Sometimes' of 'Always'. In de tweede vraag staan de reacties 'Yes', 'No' of 'Maybe'. De set van vectoren 'a', en 'b' bevatten het antwoord voor elke meting. Het derde punt in 'a' is hoe de derde persoon op de eerste vraag reageerde en het derde punt in 'b' is hoe de derde persoon op de tweede vraag reageerde.

```
> a <- c("Sometimes","Sometimes","Never","Always","Always","
    Sometimes","Sometimes","Never")
> b <- c("Maybe","Maybe","Yes","Maybe","Maybe","No","Yes","No")
> results <- table(a,b)
> results
           b
a           Maybe No Yes
  Always        2  0   0
  Never         0  1   1
  Sometimes     2  1   1
```

Matrix

Een matrix is een verzameling van gegevens die zijn aangebracht in een tweedimensionale rechthoekige indeling. Een voorbeeld van een matrix is bijvoorbeeld als volgt:

$$\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$$

```
> A = matrix(
+   c(2, 4, 3, 1, 5, 7), # the data elementen
+   nrow=2,              # aantal rijen
+   ncol=3,              # aantal kolommen
+   byrow = TRUE)        # vul de matrix aan per rij

> A                      # print de matrix
     [,1] [,2] [,3]
[1,]    2    4    3
[2,]    1    5    7

> A[2, 3]       # element op 2de rij , 3de kolom
[1] 7

> A[2, ]        # de 2de rij
[1] 1 5 7
```

```
> A[ ,c(1,3)]  # de eerste en de derde kolom
     [,1] [,2]
[1,]    2    3
[2,]    1    7
```

1.5 Oefeningen

Oefening 1.2. *Bekijk de dataset mtcars. Geef de waarde terug voor de eerste rij, tweede kolom. Geef ook het aantal rijen, het aantal kolommen. Geef ook een preview van het volledige data frame. Geef enkel de kolom terug met de definities van de cylinders. Om een data frame te bekomen met de twee kolommen mpg en hp, pakken we de kolomnamen in een indexvector in met single square bracket operator. Probeer ook eens op te zoeken hoe je een rijrecord van de ingebouwde data set mtcars bepaalt.* ■

Oefening 1.3. *Maak zelf een willekeurige datafile aan in Excel en probeer deze in te lezen in R. Zijn er nog dataformaten die ondersteund worden door R?* ■

Oefening 1.4. *Genereer een 4x5 array en noem die x. Geneer daarna een 3x2 array waar de eerste kolom de rij-index kan zijn van x en de tweede kolom een kolomindex voor x. Vervang de elementen gedefinieerd door de index in i in x door 0.* ■

Oefening 1.5. *Genereer een vector waar een voornaam en een achternaam in komen. Benoem ook de naam van de kolommen. Geef daarna ook voornaam terug van het eerste element van de array.* ■

Oefening 1.6. *Probeer voor de datafile `rainforest` in de library `DAAG` te tellen hoeveel rijen er zijn per species die volledig en compleet zijn (dus geen n.a. bevatten). Je kan hiervoor `with`, `table`, `complete.cases` voor gebruiken.* ■

Oefening 1.7. *Genereer een vector met de waarden $e^x cos(x)$ voor $x = 3, 3.1, 3.2, \ldots, 6$* ■

Oefening 1.8. *Bereken: $\sum_i^{100}(i^3 + 4i^2)$* ■

2. Het onderzoeksproces

2.1 De wetenschappelijke methode

Er zijn verschillende manieren om kennis te vergaren:

1. wetenschappelijke methode, maar ook
2. een niet-wetenschappelijke methode

Niet-wetenschappelijk

Er zijn verschillende versies van niet-wetenschappelijk redeneren:

Autoritair hier geldt iemand als autoriteit in een bepaald gebied en wordt als betrouwbaar bestempeld. Alles wat deze persoon beweert wordt aanzien als waarheid.

Deductief gegeven een set van veronderstellingen gaat men op een welbepaalde manier conclusies trekken. Alhoewel hier dus correcte conclusies kunnen behaald worden, hangt dit enkel en alleen af van de waarheid van de veronderstellingen. Maar deze veronderstellingen worden niet-empirisch onderzocht.

Wetenschappelijk

Een kenmerk van de *wetenschappelijke methode* is **empirische validering**: gebaseerd op ervaring en directe observatie. Dus een uitspraak is geldig indien het overeenkomt met wat geobserveerd wordt.

Oefening 2.1. *Probeer nu vertrekkende van de niet-wetenschappelijke en wetenschappelijke manieren aan te tonen dat varkens kunnen vliegen.* ■

Aan de hand van zo'n empirisch onderzoek kunnen we verschillende doelen behalen:

1. Exploratie: bestaat iets of gebeurt er iets?
2. Beschrijving: wat zijn de eigenschappen van deze gebeurtenis?
3. Voorspelling: is een bepaalde gebeurtenis gerelateerd aan een andere en kan ik deze zo voorspellen?
4. Controle: kan ik een gebeurtenis volledig voorspellen aan de hand van andere zaken?

Onderzoeksdoelstellingen

Er zijn twee grote onderzoeksdoelen die we willen behalen:

Generalisatie we gaan vaak maar een onderzoek doen op een bepaalde, beperkte groep van de totale groep (populatie). Indien we correcte conclusies kunnen trekken voor die subgroep, die ook gelden voor de totale groep, dan hebben we een correcte generalisatie gevonden.

Specialisatie Toepassen van algemene kennis op een specifiek domein of probleem. Toegepast onderzoek kan hier meestal onder geclassificeerd worden.

Er zijn twee soorten generalisaties:

1. Over 1 enkel fenomeen.
2. Over verbanden tussen fenomenen.

Er zijn drie redenen waarom verbanden zo belangrijk zijn:

1. Volledig verstaan van een fenomeen.
2. Verbanden kunnen zorgen voor een voorspelling
3. Causale verbanden: één van de fenomenen heeft dat andere fenomeen tot gevolg.

Fundamenteel vs. toegepast onderzoek

Afhankelijk van de onderzoeksdoelstelling spreken we van hetzij fundamenteel, hetzij toegepast onderzoek.

Fundamenteel onderzoek wordt typisch aan universiteiten uitgevoerd. Onderzoekers trachten de bestaande kennis in hun vakgebied uit te breiden. In computerwetenschappen kan het bijvoorbeeld gaan over het ontwikkelen van nieuwe algoritmen. Bij fundamenteel onderzoek wordt er niet in de eerste plaats rekening gehouden met de praktische toepassingen. Je zou kunnen zeggen dat hier in de eerste plaats geprobeerd wordt om oplossingsmethoden te ontwikkelen, en pas dan gekeken wordt welke problemen er efficiënt(er) kunnen mee opgelost worden. Het is moeilijk a priori te voorspellen welke impact (in het bijzonder financiële meerwaarde) fundamentele onderzoeksresultaten kunnen hebben. In het beste geval kunnen ze de wereld veranderen, in het slechtste komt er geen enkele praktische toepassing.

Toegepast onderzoek begint bij een concreet probleem, typisch in een bedrijfscontext. Onderzoekers moeten zich eerst inwerken in het specifieke probleemdomein. Dan kunnen ze op zoek gaan naar de meest geschikte methode om dat probleem op te lossen. Daarom moeten ze ook op de hoogte zijn van de state-of-the-art binnen het relevante fundamentele onderzoek. De meerwaarde van toegepast onderzoek is meestal makkelijker te meten, maar de impact blijft beperkt tot het bedrijf/organisatie in wiens opdracht het onderzoek werd uitgevoerd.

2.2 Basisconcepten in onderzoek

Meetniveaus

In statistiek werken we met variabelen en waarden.

Definitie 2.2.1 (Variabele)**.** *Algemene eigenschap van een object waardoor we objecten van elkaar kunnen onderscheiden. Vb. lengte, gewicht, ...*

Definitie 2.2.2 (Waarde)**.** *Specifieke eigenschap, invulling voor die variabele. Vb. 1.83m, 78 kg, ...*

Er worden meestal vier meetniveaus gebruikt in statistische analyse. Het meetniveau bepaalt welke statische methodes bruikbaar zijn.

Nominaal meetniveau : er is slechts keuze uit een beperkt aantal categorieën, waarbij geen volgorde aanwezig is tussen de antwoorden.

Ordinaal meetniveau : een variabele die is ingedeeld in categorieën, waar er echter wel een logische volgorde is tussen de categoriën.

Intervalniveau : variabelen die niet in categorieën voorkomen, en waarbij berekeningen kunnen mee uitgevoerd worden, maar zonder nulpunt.

Rationiveau : intervalniveau met nulpunt. Je kunt hierdoor verhoudingen berekenen tussen verschillende waarden op de schaal.

Oefening 2.2. *Zoek zelf nu eens voorbeelden voor de verschillende meetniveaus.* ■

Onderzoeksproces

Het onderzoeksproces kan grotendeels opgedeeld worden in 6 grote fasen:

1. Formuleren van de probleemstelling: wat is de onderzoeksvraag?
2. Exacte informatiebehoefte definiëren: welke specifieke vragen moeten we stellen?
3. Uitvoeren van het onderzoek: enquêtes, simulaties, ...
4. Verwerken van de gegevens: statistische software
5. Analyseren van de gegevens: uitvoeren van de statistische methodes
6. Conclusies schrijven: schrijven van onderzoeksverslag

Definitie 2.2.3 (Oorzakelijk verband)**.** *Een variabele veroorzaakt een oorzakelijk verband wanneer een verandering in die variabele op een betrouwbare manier een geassocieerde verandering van een andere variabele tot gevolg heeft, op voorwaarde dat alle andere potentiële oorzaken geëlimineerd zijn.*

Er is niet altijd een verband zichtbaar, en we moeten soms verder kijken dan naar de absolute waarden van de variabelen alvorens conclusies te trekken.

Voorbeeld 2.1. *Bij het voobeeld van Pepsi versus cola zou je initeel kunnen denken dat Pepsi lekkerder is omdat er meer mensen ervan geproefd hebben (70 ten opzichte van 30). Maar dit zou een verkeerde manier van redeneren zijn. We moeten relatief ten op zichte van de marginale totalen kijken. Hier zien we dan dat 56 van de 70 ($\frac{56}{70} = 0.8$) mensen die Pepsi gedronken hebben het lekker vonden, en 24 van de 30 ($\frac{24}{30}$=0.8) mensen vonden cola lekker. Dus is er geen verschil in waarden*

voor cola en Pepsi voor de gemiddelde waarde van smaak. ∎

2.3 Oefeningen

In de Bachelorproef van **Akin2016** wordt een vergelijkende studie verricht rond verschillende persistentiemogelijkheden in Android. In de Abstract kunnen we het volgende lezen:

> Vandaag de dag bestaan er veel applicaties, maar hoeveel daarvan blijven werken zonder internetverbinding? Tegenwoordig is het ondersteunen van offline werking in een applicatie geen luxe meer, maar een must-have. Om offline-support te voorzien binnen een applicatie, is er nood aan het gebruik van een database. Hierdoor zijn databases belangrijk binnen de IT-sector.
>
> Er bestaan verschillende soorten databases, maar welke moet men gebruiken? Welke is het meest geschikt bij een bepaalde soort applicatie? De keuze van de database kan een grote invloed hebben op verschillende eigenschappen: performantie, opstartsnelheid, CPU-gebruik,.. Als de database deze eigenschappen op een negatieve manier beïnvloedt, kan dit tot gevolg hebben dat het aantal gebruikers van de mobiele applicatie zal verminderen. Ter beantwoording van de probleemstelling zijn volgende deelvragen geformuleerd met betrekking op de applicatie:
>
> - Wat is de invloed van de gekozen database op de opstartsnelheid? Vertraagt het gebruik van de gekozen database de opstartsnelheid van de applicatie, of heeft het helemaal geen invloed (in vergelijking met gebruik van andere databases)?
> - Wat is de invloed van de gekozen database op het CPU-gebruik? Een hoger CPU-gebruik zal zorgen voor meer batterijverbruik. Zal de applicatie bij gebruik van de gekozen database meer of juist minder CPU gebruiken (in vergelijking met gebruik van andere databases)?
> - Wat is de gemiddelde snelheid van de gekozen database bij het toevoegen van records aan de database?
>
> Het onderzoek werd uitgevoerd op drie verschillende applicatieprofielen: weinig data (profiel 1), gemiddelde hoeveelheid data (profiel 2), veel data (profiel 3). De verwachtingen waren dat Realm altijd de beste keuze zou zijn, behalve bij applicatieprofiel 1. Daar zou SharedPreferences de beste keuze moeten zijn, aangezien het speciaal ontwikkeld is voor kleine hoeveelheden simpele data. Het onderzoek heeft echter volgend resultaat opgeleverd:
>
> 1. Weinig data : Realm
> 2. Gemiddelde hoeveelheid data : Realm
> 3. Veel data : SQLite
>
> De details van het onderzoek zijn te vinden in het volgende deel van dit scriptie.

We gaan dit onderzoek eens onder de loep nemen, kijken wat er goed aan was en wat de eventuele verbeterpunten kunnen zijn.

Oefening 2.3. *Probeer volgende vragen zo goed mogelijk te beantwoorden.*

1. *Wat is de doelstelling van het onderzoek?*
2. *Wie is het publiek?*

3. *Worden de conclusies expliciet gemaakt?*
4. *Schets kort hoe de structuur van het document in elkaar zit. Komt dit overeen met wat er gezien is in de les?*

■

Oefening 2.4. *Schrijf voor jezelf hoe de volgende componenten van het onderzoek ingevuld zijn.*

- *Context*
- *Nood*
- *Taak*
- *Object*
- *Resultaat*
- *Conclusie*
- *Perspectief*

Indien je op vorige componenten geen antwoord vindt uit de tekst, probeer dan zelf een antwoord te formuleren indien jij dit onderzoek zou uitvoeren. ■

3. Analyse op 1 variabele

Definitie 3.0.1 (Beschrijvende statistiek)**.** *Met beschrijvende statistiek bedoelen we een verzameling van technieken om data synthetisch voor te stellen en samen te vatten.*

3.1 Voorbeeld met superhelden

3.2 Gemiddelde

Definitie 3.2.1 (Gemiddelde)**.** *Het gemiddelde (symbool μ) van een set waarden is de som van al deze waarden gedeeld door het aantal waarden. De formule staat beschreven in 3.1.*

$$\mu = \frac{1}{n} \times \sum_{i=1}^{n} x_i \tag{3.1}$$

Waarbij:

- *x_i de waarden zijn vanuit tabel 3.1.*
- *n het aantal waarden is. In het voorbeeld van de superhelden zou dit 5 zijn, want we hebben 5 lengtes van superhelden.*

Oefening 3.1. *Wat is de gemiddelde lengte van de superhelden?* ■

x_1	x_2	x_3	x_4	x_5
141	198	143	201	184

Tabel 3.1: Voorbeeldtabel superhelden vanuit slides

Oefening 3.2. *Vraag: het gemiddelde van 15 cijfers is 12. Welk nummer moeten we aan de rij van cijfers toevoegen om een gemiddelde van 13 te bekomen?* ■

Het rekenkundig gemiddelde is gevoelig aan outliers: een extreme waarde kan het rekenkundig gemiddelde zwaar beïnvloeden.

3.3 Mediaan

Definitie 3.3.1 (Mediaan)**.** *Indien we alle cijfers sorteren van klein naar groot, is de mediaan het middelste cijfer, of het gemiddelde van de twee middelste cijfers indien het aantal cijfers even is.*

De mediaan is niet gevoelig aan outliers.

3.4 Modus

Definitie 3.4.1 (Modus)**.** *De modus is het cijfer dat het meest voorkomt in een set van cijfers.*

- Heeft niet veel zin als alle cijfers even veel voorkomen. (Zoals bij onze superhelden). Misschien is het nuttig om ze dan te groeperen.
- Er kunnen twee modi zijn: dit noemen we bimodaal;
- Er kunnen meerdere modi zijn: dit noemen we multimodaal.

Voorbeeld 3.1. *Het groeperen kunnen we tonen bijvoorbeeld bij het aantal mensen gered door Batman de laatste acht jaar.*

- $[0-9]$ *mensen : 4, 7*
- $[10-19]$ *mensen: 11, 16*
- $[20-29]$ *mensen : 20, 22, 25, 26*
- $[30-39]$ *mensen: 33*

Dus categorie $[20-29]$ *komt het meest voor. We kunnen dus bv. kiezen om 25 als modus te gebruiken. Zo'n klasse noemen we dan een modale klasse.* ■

3.5 Range / Bereik

Definitie 3.5.1 (Bereik)**.** *Het bereik in een set van getallen is de absolute waarde van het verschil tussen het laagste en grootste getal.*

3.6 Kwartielen & kwartielafstand

Definitie 3.6.1 (Kwartielen & Kwartielafstand)**.** *De kwartielen zijn de waarden die een gesorteerde lijst van nummers in 4 gelijke delen deelt. Elk deel vormt dus een kwart van de dataset. Men spreekt van een eerste, tweede en derde kwartiel (*Q_1*,* Q_2*,* Q_3*).*

Dus:

- eerste kwartiel Q_1 is de getalswaarde die de laagste 25 % van de reeks afscheidt.
- tweede kwartiel Q_2 is de getalwaarde die de laagste 50% van de reeks afscheidt.
- derde kwartiel Q_3 is de getalwaarde die de laagste 75% van de reeks afscheidt.

Definitie 3.6.2. *Kwartielafstand is het verschil tussen Q_3 en Q_1 (dus $Q_3 - Q_1$).*

Methode om te berekenen (volgens **Moore2002**) (met n oneven aantal getallen):

- Q_1 komt overeen met cijfer $\frac{n+1}{4}$
- Q_3 komt overeen met cijfer $\frac{3n+3}{4}$

Methode om te berekenen (met n even aantal getallen):

- Q_1 komt overeen met cijfer $\frac{n+2}{4}$
- Q_3 komt overeen met cijfer $\frac{3n+2}{4}$

Oefening 3.3. *Met welke voorgaande statistiek komt Q_2 overeen?* ■

3.7 Variantie en standaardafwijking

Definitie 3.7.1 (Variantie)**.** *De variantie (symbool σ^2 - lees sigma kwadraat) is het gemiddelde van de kwadraten van de verschillen tussen de waarde van de dataset en het gemiddelde.*

$$\sigma^2 = \frac{1}{n} \times \sum_{i=1}^{n} (\mu - x_i)^2 \tag{3.2}$$

Voorbeeld 3.2. *De variantie bij de lengtes van onze superhelden wordt als volgt berekend:*

$$\begin{aligned}
\sigma^2 &= \frac{(173.4-141)^2 + (173.4-198)^2 + (173.4-143)^2 + (173.4-201)^2 + (173.4-184)^2}{5} \\
&= \frac{(-32.4)^2 + (24.6)^2 + (-30.4)^2 + (27.6)^2 + (10.6)^2}{5} \\
&= \frac{1049.76 + 605.16 + 924.16 + 761.76 + 112.36}{5} \\
&= \frac{3453.2}{5} = 690.64
\end{aligned} \tag{3.3}$$

■

Definitie 3.7.2 (Standaardafwijking)**.** *De standaardafwijking wordt dan gedefinieerd als de vierkantswortel van de variantie.*

$$\sigma = \sqrt{\sigma^2} \tag{3.4}$$

Dit geeft ons dus inzicht in wat normaal is en wat abnormaal is: een kleine standaardafwijking wijst erop dat de waarden dicht bij de centrummaat (μ) liggen, terwijl een grote standaardafwijking duidt dat de waarden verspreid liggen over een groot bereik van waarden. In sommige gevallen wil men een grote standaardafwijking, in andere gevallen niet zoals hieronder beschreven.

Analyse	**Nominaal**	**Ordinaal**	**Interval** of **Ratio**
Centrum	Modus Modale klasse	Mediaan Modus Modale klasse	Gemiddelde Mediaan Modale klasse
Spreiding		Range Interkwartielafstand	Range Interkwartielafstand Standaarddeviatie

Tabel 3.2: Meetniveaus en mogelijkheden op variabelen

Voorbeeld 3.3. *Bij het vervaardigen van een schroevendraaier is de grootte van de kop belangrijk voor het goed functioneren van de schroevendraaier. Als we dus van 100 verschillende schroevendraaiers de kopgrootte meten, is het beter dat die grootte redelijk constant is en wensen we dus een kleine standaardafwijking.* ∎

Voorbeeld 3.4. *Bij het onderzoek naar onze superhelden, wensen we te weten hoeveel ze ongeveer verdienen in hun normale job. We hebben een aantal rijke superhelden (bv. Batman) en een aantal minder rijke superhelden (bv. Spiderman). De spreiding op hun inkomen is dus groot, maar dat is niet per definitie slecht.* ∎

Een aangename eigenschap van de standaardafwijking is dat het uitgedrukt kan worden in dezelfde metriek als de gemeten data. Bij ons voorbeeld van de superhelden, wil dat zeggen dat de standaardafwijking 26.28 cm is.

Zoals het gemiddelde is de variantie en de standaarddeviatie gevoelig aan outliers (uitschieters). De variantie is eigenlijk gevoeliger dan het gemiddelde. Inderdaad, voor een outlier is de afstand tot het gemiddelde kleiner dan het kwadraat van deze afstand.

3.8 Toepassing spreidingsmaten en maten centraliteit op verschillende soorten variabelen

3.9 Grafieken

3.9.1 Boxplot

De boxplot wordt gevormd door een rechthoek begrensd door de kwartielwaarden (25% en 75%). In deze rechthoek wordt ook de mediaan getekend. De stelen, die aan de rechthoek zitten, bevatten de rest van de waarnemingen op de uitschieters en extremen na.

- Een uitschieter is een waarde die meer dan 1.5 keer de interkwartielafstand boven/onder het derde/eerste kwartiel ligt. Wordt aangeduid met een cirkeltje.
- Een extremum is een waarde die meer dan 3 keer de interkwartielafstand boven/onder het derde/eerste kwartiel ligt. Wordt aangeduid met een sterretje.

3.10 R

Zodra u een vector (of een lijst met nummers) in het geheugen hebt, zijn de meeste basiswerkingen beschikbaar. De meeste basiswerkzaamheden werken op een hele vector en kunnen gebruikt worden om snel een groot aantal berekeningen uit te voeren met een enkele opdracht. Indien je een operatie uitvoert op meerdere vectoren, is het vaak nodig dat de vectoren allemaal hetzelfde aantal elementen bevatten.

Hieronder zie je een set van eenvoudige operaties die je met vectoren kan doen. Let op dat de operaties allemaal op een element per element basis uitgevoerd worden.

```
> a <- c(1,2,3,4)
> a
[1] 1 2 3 4
> a + 5
[1] 6 7 8 9
> a - 10
[1] -9 -8 -7 -6
> a*4
[1]  4  8 12 16
> a/5
[1] 0.2 0.4 0.6 0.8

> b <- a - 10
> b
[1] -9 -8 -7 -6

> sqrt(a)
[1] 1.000000 1.414214 1.732051 2.000000
> exp(a)
[1]  2.718282  7.389056 20.085537 54.598150
> log(a)
[1] 0.0000000 0.6931472 1.0986123 1.3862944
> exp(log(a))
[1] 1 2 3 4

> c <- (a + sqrt(a))/(exp(2)+1)
> c
[1] 0.2384058 0.4069842 0.5640743 0.7152175

> a + b
[1] -8 -6 -4 -2

> a*b
[1]  -9 -16 -21 -24
> a/b
[1] -0.1111111 -0.2500000 -0.4285714 -0.6666667
> (a+3)/(sqrt(1-b)*2-1)
[1] 0.7512364 1.0000000 1.2884234 1.6311303
```

De volgende commando's kunnen worden gebruikt om het gemiddelde,de kwartielen, het minimum, het maximum, de variantie en de standaardafwijking van een reeks getallen te verkrijgen.

```
> attach(computers)
> mean(price)
[1] 2219.577
> median(price)
[1] 2144
> quantile(price)
  0%  25%  50%  75% 100%
 949 1794 2144 2595 5399
> min(price)
[1] 949
> max(price)
[1] 5399
> var(price)
[1] 337333.2
> sd(price)
[1] 580.804

> summary(computers)
      price            speed             hd               ram
                       screen            cd               multi
                       premium
 Min.   : 949   Min.   : 25.00   Min.   :  80.0   Min.   : 2.000
       Min.   :14.00   Length:6259        Length:6259
   Length:6259
 1st Qu.:1794   1st Qu.: 33.00   1st Qu.: 214.0   1st Qu.: 4.000
       1st Qu.:14.00   Class :character   Class :character
   Class :character
 Median :2144   Median : 50.00   Median : 340.0   Median : 8.000
       Median :14.00   Mode  :character   Mode  :character
   Mode  :character
 Mean   :2220   Mean   : 52.01   Mean   : 416.6   Mean   : 8.287
       Mean   :14.61
 3rd Qu.:2595   3rd Qu.: 66.00   3rd Qu.: 528.0   3rd Qu.: 8.000
       3rd Qu.:15.00
 Max.   :5399   Max.   :100.00   Max.   :2100.0   Max.   :32.000
       Max.   :17.00
      ads              trend
 Min.   : 39.0   Min.   : 1.00
 1st Qu.:162.5   1st Qu.:10.00
 Median :246.0   Median :16.00
 Mean   :221.3   Mean   :15.93
 3rd Qu.:275.0   3rd Qu.:21.50
 Max.   :339.0   Max.   :35.00
```

Pinnen x	Frequentie f_x
0	2
1	1
2	2
3	0
4	2
5	4
6	9
7	11
8	13
9	8
10	8

Tabel 3.3: Tijdens het spelen van een kegelspel is bijgehouden hoeveel pinnen telkens omver gegooid werden. Voor elke mogelijke score x is bijgehouden hoeveel keer

3.11 Oefeningen

3.11.1 Centrum- en spreidingsmaten

Definitie 3.11.1. *Een frequentietabel is tabel waarin opgesomd staat hoeveel keer een waarde voorkomt in de volledige dataset (= frequentie). Meestal zijn de tabellen verticaal georiënteerd.*

Oefening 3.4. *De formules voor gemiddelde μ en variantie σ^2 staan beschreven in secties 3.2 en 3.7, resp. Hoe moeten deze formules aangepast worden om μ en σ^2 te berekenen wanneer we te maken hebben met een frequentietabel? Doe dit voor de data in tabel 3.3.* ■

Oefening 3.5. *In de formule voor de variantie wordt het verschil tussen de meetpunten en het gemiddelde gekwadrateerd. Waarom? Zouden we geen eenvoudiger formule kunnen bedenken die een even goede maatstaf is voor de spreiding van een dataset? Hieronder vind je drie voorstellen (de derde is de "echte" formule).*

$$\sigma_1^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu - x) \tag{3.5}$$

$$\sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n}|\mu - x| \tag{3.6}$$

$$\sigma_3^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu - x)^2 \tag{3.7}$$

Pas elke formule toe op de twee datasets hieronder. Door het resultaat te vergelijken zou je moeten kunnen besluiten of de formules geschikt zijn als een spreidingsmaat.

$$X = \{4,4,-4,-4\}$$
$$Y = \{7,1,-6,-2\}$$

■

Oefening 3.6. *Zoek eens zelfstandig op wat de variatiecoëfficiënt is. Hoe wordt die gedefinieerd voor een volledige populatie en wat zou je ermee kunnen doen?* ■

Oefening 3.7. *Beschouw de volgende subsets uit het data frame* `ais` *(uit de library DAAG):*

1. *Ontleed de gegevens voor de roeiers.*
2. *Ontleed de gegevens voor de roeiers, de netballers en de tennissers.*
3. *Ontleed de gegevens voor de de vrouwelijke basketballers en roeiers.*

■

Oefening 3.8. *Gebruik de functies* `mean` *en* `range` *om het gemiddelde en bereik van:*

- *de cijfers 1, 2, . . . , 21*
- *50 willekeurige normale waarden, die worden gegenereerd vanuit een normale distributie met gemiddelde 0 en variantie 1 (functie* `rnorm`*)*
- *de kolommen* `height` *en* `weight` *in de data frame* `women` *(standaard in R).*

■

In vorige oefeningen hebben we de verschillende spreidingsmaten en centrummaten besproken. Zoals je merkt worden deze metrieken ook gebruikt in het onderzoek van **Akin2016**. In de volgende oefeningen gaan we trachten de resultaten te reproduceren.

Hiervoor hebben we het bestand `android_persistence_cpu.csv` nodig. Je vindt die in de Github-repository van deze cursus, onder directory `oefeningen/data/oef_3_1variabele`

Oefening 3.9. *Open de file met excel en bekijk de structuur van het document. Hoe ziet die er uit? Kan je de variabelen identificeren en hun type benoemen.* ■

We gaan het programma `R` gebruiken samen met `RStudio`. Open de file in `RStudio`.

```
1 android_cpu <-  read.csv("android_persistence_cpu.csv", sep=";",
      dec=",")
2 attach(android_cpu)
```

We hebben nu de data ingeladen. We kunnen eens kijken wat de gemiddelde tijd, de standaarddeviatie, de kwartielen e.a. zijn. Gebruik hiervoor de commando's `mean`, `median`, `quantile`, `min`, `max`, `var`, `sd`. Je kan ook makkelijk gebruik maken van de methode `summary`.

Oefening 3.10. *Als je de vorige metrieken berekend hebt, wat kan je daar dan over zeggen. Kan je zinnige conclusies trekken uit de vorige resultaten. Zo ja vermeld ze, zo nee beschrijf waarom je dat denkt.* ■

3.11.2 Grafieken in R

Een histogram is een eenvoudige plot. het toont de frequenties van de data die in een bepaald bereik voorkomen.

```
1 hist(android_cpu$Tijd,main="Verdeling van de tijd",xlab="De
    gemeten cpu tijd");
2 hist(android_cpu$Tijd,main="Verdeling van de tijd",xlab="De
    gemeten cpu tijd",breaks=2);
```

Oefening 3.11. *Wat concludeer je als je bovenstaande grafiek*[a] *genereert? Is dit een zinnig resultaat? Wat gebeurt er als je de variabele breaks verhoogt?* ■

[a] Heb je wat problemen met het genereren van grafieken, volgende link https://www.datacamp.com/community/tutorials/15-questions-about-r-plots#gs.RK_ORsI bevat een aantal goede tips and tricks om je op weg te helpen.

Een boxplot toont de mediaan, de kwartielen, het maximum en het minimum van een dataset. Dit geeft ons een duidelijk impressie van hoe de data er uitziet.

```
1 boxplot(x = android_cpu$Tijd);
2 boxplot(android_cpu$Tijd,main='Spreiding van de CPU tijd',ylab='
    Tijd in ms');
```

Oefening 3.12. *De boxplot wordt standaard verticaal getekend. Gebruik het commando* `help(boxplot)` *om uit te zoeken hoe we de tekening horizontaal krijgen.* ■

Als je goed geantwoord hebt op de volgende vragen merk je natuurlijk dat het weinig zin heeft de volledige dataset te analyseren, aangezien de dataset verdeeld is over verschillende categorieën. We willen dus wel deze statistieken weten, maar per categorie. We kunnen zus een boxplot maken voor elke categorie.

```
1 boxplot(android_cpu$Tijd~android_cpu$Datahoeveelheid,main='
    Spreiding van de CPU tijd t.o.v. datahoeveelheid',ylab='Tijd
    in ms');
```

Oefening 3.13. *Interpreteer de resultaten die je behaalt uit deze grafiek. Zijn deze al wat zinniger?* ■

We kunnen hetzelfde doen voor de verschillende soorten dataopslagmogelijkheden in android.

Oefening 3.14. *Zelfde vraag als 3.13 Interpreteer de resultaten die je behaalt uit deze grafiek. Zijn deze al wat zinniger?* ■

We kunnen eens kijken hoe de data eruit ziet over alle categorieën heen.

```
boxplot(android_cpu$Tijd~android_cpu$PersistentieType*android_
    cpu$Datahoeveelheid,main='Spreiding van de CPU tijd',ylab='
    Tijd in ms');
```

Het blijkt dat we wel al een duidelijker zicht krijgen over de data over de categorieën heen, maar de figuur is op dit moment te druk.

We moeten de data dus onderverdelen in categorieën namelijk onder `PersistentieType` en `Datahoeveelheid`. We gaan hiervoor de functie `which`[1] gebruiken en kijken hoe de verschillende datahoeveelheden verschillen per datahoeveelheidcategorie.

```
greenDOA <- android_cpu[which(android_cpu$PersistentieType=='
    GreenDAO'),];
boxplot(greenDOA$Tijd~greenDOA$Datahoeveelheid);
```

Oefening 3.15. *Wat concludeer je uit de vorige grafiek?* ■

Oefening 3.16. *Ga nu zelf na welke boxplots er interessant zijn om te maken, en kijken of jouw resultaten overeen met die van* **Akin2016**. *Welke conclusies trek je?* ■

3.11.3 Antwoorden op geselecteerde oefeningen

Oefening 3.4

- $\mu = 7$
- $\sigma^2 \approx 5.7333$
- $\sigma \approx 2.3944$

Oefening 3.7

De opgave van deze oefening is zeer algemeen gesteld. Onder "ontleed de gegevens" wordt concreet bedoeld alle technieken voor de analyse van een variabele uit te proberen op de dataset. We geven hier een voorbeeld van de resultaten voor enerzijds kwantitatieve en anderzijds kwalitatieve variabelen.

Tabel 3.4 geeft een overzicht met de belangrijkste centrum- en spreidingsmaten voor de variabele `ht` (height, lengte) over de gevraagde groepen.

In deeloefening 1 en 2 nemen we de variabele `sex` als voorbeeld. Zie tabel 3.5 voor een overzicht. Over kwalitatieve variabelen valt minder te zeggen, we geven hier een frequentietabel waaruit we de modus kunnen afleiden.

In deeloefening 3 zijn enkel vrouwen geselecteerd, en voor deze oefening tonen we in tabel 3.6 de frequenties van variabele `sport`.

[1] Je kan ook gebruik maken van de functie `subset`, wat misschien zelfs eenvoudiger is

	(1)	(2)			(3)	
	Row	**hele groep**	**Row**	**Netball**	**Tennis**	**B_ball**	**Row**
gemiddelde	182.376	179.066	182.376	176.087	174.164	182.269	178.859
stdev	7.798	7.936	7.798	4.124	9.858	8.621	5.970
min	156.000	156.000	156.000	168.600	157.900	169.100	156.000
Q1	179.300	174.200	179.300	173.450	167.300	174.000	177.600
mediaan	181.800	179.500	181.800	176.000	175.000	184.600	179.650
Q3	186.300	183.400	186.300	179.150	180.750	188.700	181.200
max	198.000	198.000	198.000	183.300	190.800	195.900	186.300
IQR	7.000	9.150	7.000	5.700	13.450	14.700	3.600

Tabel 3.4: Overzicht resultaten in oefening 3.7 voor de variabele `ht` (height/lengte), met drie cijfers na de komma. In deeloefening 2 zijn de resultaten zowel gegeven voor de hele groep (roeiers, netballers én tennissers) als opgesplitst (via de functie `aggregate`).

	(1)	(2)			
	Row	**hele groep**	**Row**	**Netball**	**Tennis**
f	22	52	22	23	7
m	15	19	15	0	4
modus	f	f	f	f	f

Tabel 3.5: Overzicht resultaten in oefening 3.7 (1) en (2) voor de variabele `sex`. Meer bepaald zijn hier de frequenties van de waarden opgegeven, en ook telkens de modus.

	Frequenties
B_ball	13
Row	22
modus	Row

Tabel 3.6: Overzicht resultaten in oefening 3.7 (3) voor de variabele `sport`.

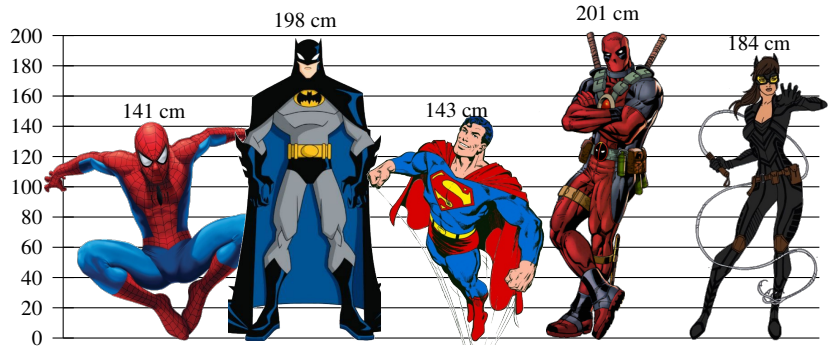
4. Steekproefonderzoek

Een reden om kwantitatief onderzoek uit te voeren is het kunnen doen van uitspraken die een representatief beeld van de werkelijkheid geven. Hierbij wordt vaak gebruikgemaakt van een steekproef. Een steekproef is een selectie uit een totale populatie ten behoeve van een meting van bepaalde eigenschappen van die populatie.

4.1 Populatie en Steekproeven

Definitie 4.1.1 (Populatie)**.** *De verzameling van **alle** objecten of personen waar men in geïnteresseerd is en onderzoek naar wil doen noemt men de populatie. Een ander woord is ook wel onderzoeksgroep of doelgroep.*

Definitie 4.1.2 (Steekproef)**.** *Wanneer met een subgroep uit een populatie gaat onderzoeken, dan noemen we die groep een steekproef.*



Figuur 4.1: De superhelden die we onderzoeken

Figuur 4.2: Onze superhelden die we onderzoeken vormen een steekproef uit de populatie van alle superhelden.

Definitie 4.1.3 (Steekproefkader)**.** *Een steekproefkader is een lijst van alle leden van een te onderzoeken populatie.*

Er zijn een aantal redenen waarom een steekproef genomen wordt:

- Populatie is te groot om een volledig onderzoek te doen.
- Kostbare metingen waardoor het onderzoek te duur wordt.
- Wanneer snelheid belangrijk is, is het vaak sneller een subgroep te onderzoeken;
- Gemakkelijker ...

Om een steekproef op te zetten volg je volgende stappen:

1. **Definitie populatie**: Wie is er deel van de populatie? Dit hangt nauw samen met de probleemstelling van het onderzoek. Dit is een zeer belangrijke stap waar je niet licht over mag gaan. Elementen die van belang zijn, zijn bijvoorbeeld sociale, demografische of fysieke kenmerken zoals geslacht, leeftijd, woonplaats ...
2. **Bepalen van steekproefkader**: Een populatie heeft verschillende segmenten zoals bijvoorbeeld rijke superhelden, arme superhelden, bekende en onbekende superhelden, superhelden met en zonder oudercomplex ... In de praktijk is het meestal onmogelijk om de populatie als geheel te onderzoeken. Daarom beperken we ons vaak tot enkele redelijk homogene subpopulaties of segmenten. De populatiesegmenten die daadwerkelijk onderzocht worden noemen we de operationele populatie.
3. **Budget en Tijd**: het aantal te onderzoeken objecten of personen zal ook afhankelijk zijn van budget en tijd.

4.2 Kiezen van steekproefmethode

Soms is de populatie die men wenst te bestuderen erg verschillend op een aantal belangrijke kenmerken. Daartoe wordt de populatie als geheel in een aantal elkaar niet-overlappende en

	Leeftijd				
Geslacht	≤ 18	$]18,25]$	$]25,40]$	>40	Totaal
Vrouw	500	1500	1000	250	3250
Man	400	1200	800	160	2560
Totaal	900	2700	1800	410	5810

Tabel 4.1: Frequenties van de superhelden in de populatie volgens geslacht en leeftijdscategorie

	Leeftijd				
Geslacht	≤ 18	$]18,25]$	$]25,40]$	>40	Totaal
Vrouw	50	150	100	25	325
Man	40	120	80	16	256
Totaal	90	270	180	41	581

Tabel 4.2: Steekproef van superhelden gestratificeerd volgens geslacht en leeftijdscategorie.

homogene strata of klassen ingedeeld.

Definitie 4.2.1 (Gestratificeerde steekproef)**.** *Een **gestratificeerde** steekproef is proportioneel als het aandeel van de subpopulatie in de steekproef gelijk is aan het aandeel van de subpopulatie in de populatie als geheel.*

Voorbeeld 4.1. *Indien we uit een populatie van de superhelden kijken naar de leeftijd van mannen en vrouwen, zien we in tabel 4.1 de absolute waarden. We kunnen niet alle superhelden ondervragen, maar indien we een steekproef nemen waarbij de mannen en leeftijdscategorieën relatief equivalent zijn met de populatie, hebben we een gestratificeerde steekproef genomen (zie tabel 4.2).* ∎

Nadat gestratificieerd is, moet bepaald worden op welke wijze binnen ieder stratum het aantal benodigde objecten of respondenten gekozen moet worden. Bij de toevals- of **aselecte** steekproeven heeft elk element van de populatie een even mogelijke kans om in de steekproef te worden opgenomen. Dit heeft als gevolg dat je op basis van de data van een aselecte steekproef conclusies kan trekken ten aanzien van de kenmerken van een populatie, en dit in tegenstelling met een **niet-aselecte** steekproef. In een niet-aselecte steekproef kent men de kans niet dat elk lid van de populatie heeft om in de steekproef terecht te komen, met als gevolg dat je gegevens enkel gelden voor je onderzochte groep.

4.2.1 Fouten bij steekproeven

Toevallige steekproeffouten

Wanneer er puur door het toeval een verschil is in een waarde voor de populatie en de steekproef.

Systematische steekproeffouten

Een procedure in de steekproef die een fout oplevert die een systematische oorzaak heeft en dus niet te wijten is aan toevallige effecten. Bijvoorbeeld door systematisch een bevoordeeld deel van de populatie te ondervragen. Als we onze superhelden zouden ondervragen via het internet, sluiten

we alle superhelden uit die geen internetverbinding hebben.

Toevallige niet-steekproeffouten

Hieronder vallen bijvoorbeeld verkeerd aangekruiste antwoorden of verschil in interpretatie van de vragen.

Systematische niet-steekproeffouten

Wanneer bijvoorbeeld respondenten met een sterke band met het onderzoek eerder geneigd zijn om een vragenlijst in te vullen, ga je positievere antwoorden krijgen - terwijl ze niet representatief zijn voor de gehele populatie.

4.2.2 Aanpassing formules standaarddeviatie

We noemen het gemiddelde van de steekproef het steekproefgemiddelde en gebruiken hiervoor het symbool $\bar{x}$ (dit hebben we stilzwijgend al een aantal keer gedaan in de vorige hoofdstukken).

Als we de standaardafwijking van een steekproef willen bepalen dan moeten we niet meer delen door n (aantal metingen) maar door $(n-1)$. Waarom?

Aangezien de som van de afwijkingen $x_i - \bar{x}$ steeds 0 oplevert (zie hieronder in vergelijking 4.1), kan de laatste afwijking gevonden worden uit de eerste $n-1$ afwijkingen. We berekenen dus niet het gemiddelde van n getallen zonder verwantschap. Slecht $n-1$ van de gekwadrateerde afwijkingen kunnen vrij bewegen, daarom berekenen we het gemiddelde door het totaal te delen door $n-1$. Het getal $n-1$ noemt men het aantal vrijheidsgraden van de variantie of van de standaardafwijking.

$$\sum_i^n (x_i - \bar{x}) = \sum_i^n x_i - \sum_i^n \bar{x} = \sum_i^n x_i - n(\frac{1}{n}\sum_i^n x_i) \tag{4.1}$$

4.3 Kansverdeling van een steekproef

4.3.1 Stochastisch experiment

Een random (of stochastisch) experiment heeft volgende elementen nodig:

Definitie 4.3.1 (Universum of Uitkomstenruimte)**.** *Het universum of uitkomstenruimte van een experiment is de verzameling van alle mogelijke uitkomsten van dit experiment en wordt genoteerd met* Ω.

Opmerkingen

- De uitkomstenruimte moet *volledig* zijn: elke mogelijke uitkomst van een experiment moet tot Ω behoren.

- Bovendien moet elke uitkomst van een experiment overeenkomen met *juist één* element van Ω.
- Samengevat: na het uitvoeren van een experiment is het mogelijk om eenduidig aan te geven welk element van Ω zich heeft voorgedaan.

Definitie 4.3.2 (Gebeurtenis)**.** *Een gebeurtenis is een deelverzameling van de uitkomstenruimte. Een enkelvoudige of elementaire gebeurtenis is een singleton; een samengestelde gebeurtenis heeft cardinaliteit groter dan 1.*

Gebeurtenissen die geen gemeenschappelijke uitkomsten hebben noemt men disjunct.
Disjuncte gebeurtenissen kunnen dus nooit samen voorkomen.
Wanneer de gebeurtenissen A en B disjunct zijn dan geldt $A \cap B = \emptyset$. Startend met de gebeurtenissen A en B kan men de volgende gebeurtenissen vormen:

- A **of** B, of wiskundig genoteerd $A \cup B$;
- A **en** B, of wiskundig genoteerd $A \cap B$;
- **niet** A, of wiskundig genoteerd $\overline{A}$.

Opmerkingen

- Door inductie leidt men gemakkelijk af dat de unie van n gebeurtenissen A_1 t.e.m. A_n eveneens een gebeurtenis is.
- Idem voor de doorsnede van gebeurtenissen.
- Voor sommige toepassingen is het nodig om ook (aftelbaar) oneindige unies en doorsnedes te beschouwen.

Definitie 4.3.3 (Kansruimte)**.** *Het toekennen van kansen aan gebeurtenissen dient aan de volgende drie regels te voldoen.*

1. *Kansen zijn steeds positief:* $P(A) \geq 0$ *voor elke* A.
2. *De uitkomstenruimte heeft kans 1:* $P(\Omega) = 1$.
3. *Wanneer* A *en* B disjuncte *gebeurtenissen zijn dan is*

$$P(A \cup B) = P(A) + P(B).$$

Dit noemt men de somregel.

Wanneer de functie P *aan de bovenstaande eigenschappen (axioma's) voldoet dan noemt men het drietal* $(\Omega, \mathscr{P}(\Omega), P)$ *een kansruimte (met* $\mathscr{P}(\Omega)$ *de* machtsverzameling *van* Ω*, d.w.z. de verzameling van alle deelverzamelingen van* Ω*).*

Voorbeeld 4.2. *Beschouw een uitkomstenruimte* $\Omega = \{1,2,3,4,5,6\}$ *en een kansfunctie* $P(\omega) = \frac{1}{|\Omega|}$*, dan zou dit een dobbelsteen kunnen voorstellen met uitkomsten 1 tot en met 6 met een kans* $P(\omega) = \frac{1}{6}$ *om een van de nummers te werpen.* ∎

In dit onderdeel van de cursus gaan we ons bezig houden met **inductieve statistiek**: op basis van een getrokken steekproef uitspraken doen over de populatie.

4.3.2 Kansverdeling

Discrete kansverdeling

Als we het voorbeeld nemen van het gooien van een dobbelsteen, dan kunnen we de kans dat we een van de getallen $\Omega = \{1,2,3,4,5,6\}$ in een tabel zetten of kunnen er een histogram van maken. Er zijn een aantal belangrijke opmerkingen hierbij:

1. De kansen zijn allemaal groter of gelijk aan nul.
2. De kans op een getal is gelijk aan de bijbehorende oppervlakte van de staaf.
3. De totale oppervlakte van alle staven is 1.

Een ander voorbeeld is het gooien van twee dobbelstenen met de mogelijke uitkomst. Je hebt volgens de productregel $6 \times 6 = 36$ mogelijke uitkomsten. Om bijvoorbeeld drie te gooien heb je twee mogelijkheden (kans $P(X = 3) = \frac{2}{36}$). Zie voor de andere getallen tot en met 7 de tabel ($P[X = n] = \frac{n-1}{36}$).

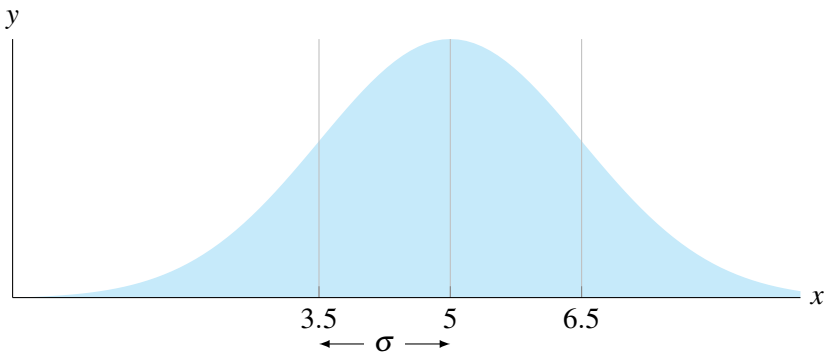
Als we dit nu in een histogram steken bekomen we een mooie trap naar boven tot 7 en dan weer naar beneden. Nu kan je makkelijk zien dat:

- Voor de kans om 10 of meer te gooien moet je bijvoorbeeld die blauwe oppervlakte hebben.
- Voor de kans op een aantal meer dan 2 maar minder dan 7 moet je de rode oppervlakte hebben.
- Voor de kans op een aantal meer dan 7 maar minder dan 10 moet je de groene oppervlakte hebben.
- Dan is het ook logisch dat de totale oppervlakte 1 is: de kans dat 1 van al die mogelijkheden voorkomt is natuurlijk 100%.

Continue kansverdeling

Continue kansverdelingen zijn verdelingen waarbij hetgeen we meten niet alleen een beperkt aantal waarden kan aannemen (nominaal en ordinaal meetniveau), maar ook alle er tussenliggende waarden (ratio- en intervalniveau). Neem bijvoorbeeld het gewicht van onze superhelden. Dat is continu, immers dat kan niet alleen 60 of 70 kilo zijn, maar ook (bij benadering) $66,8735485653$ kilo. In principe zijn alle tussenliggende waarden mogelijk (al is dat in praktijk vaak niet te meten). Dat heeft een belangrijk gevolg voor de kansverdeling. Die bestaat nu (in theorie) niet meer uit losse staafjes, maar is een vloeiende kromme geworden. Dat betekent dat de kans op bijvoorbeeld precies 70 kilogram een kans nul heeft. Bij precies 70 kg hoort een verticaal lijntje, en een lijntje heeft oppervlakte nul. Nu is die kans natuurlijk ook nul. Als we zeggen 70 kg, dan bedoelen we meestal tussen $69,5$ en $70,5$, of preciezer het interval $[69,5;70,5[$. Als we zeggen $70,00000$ kg, dan bedoelen we iets als binnen $[70,000005;69,999995[$ kg.

De twee regels voor kansverdelingen hierboven blijven gewoon geldig. Als zo'n kromme een goede kansverdeling is, dan moet de totale oppervlakte ervan 1 zijn, en dan kun je de kans op een gewicht dat bijvoorbeeld tussen de 60 en 70 kg ligt uitrekenen door de oppervlakte hiernaast te bepalen (merk op dat het uiteindelijk niet belangrijk is of die 60 en 70 zelf ook nog tot het interval behoren, die hebben toch kans nul!).



Figuur 4.3: De kansverdeling van de reactiesnelheid van superman. Deze grafiek noemen we de normaalverdeling met gemiddelde $\mu = 5$ ms en standaarddeviatie $\sigma = 1,5ms$.

4.4 De normale verdeling

In figuur 4.3 tonen we de kansverdeling van de reactiesnelheid X van superman. Deze grafiek noemen we de normaalverdeling met gemiddelde 5 ms en standaarddeviatie 1,5 ms. Symbolisch:

$$X \sim Nor(\mu = 5; \sigma = 1,5)$$

De functie die hiermee gepaard gaat is de volgende:

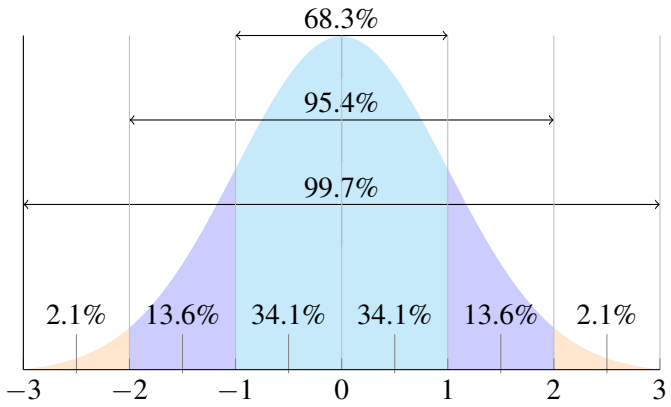
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{4.2}$$

De normale verdeling kent volgende eigenschappen:

- Normale verdeling is klokvormig
- De normale verdeling is symmetrisch
- Vanwege symmetrie is gemiddelde, mediaan en modus aan elkaar gelijk
- De totale oppervlakte onder de klokvormige figuur is 1
- In gebied σ onder μ en σ boven μ (het zogenoemde sigma gebied) ligt ongeveer 68% van de waarnemingen.
- In het gebied 2σ boven en onder μ ligt ongeveer 95% van alle waarnemingen.
- Voor de verschillende gebieden zie figuur 4.4

4.4.1 De standaardnormale verdeling

Indien de toevalsveranderlijke $X \sim N(\mu, \sigma)$ verdeeld is dan is de toevalsvariabele $Z = \frac{X-\mu}{\sigma}$ normaal verdeeld: $Z \sim N(0,1)$. Dit noemen we de standaardnormale verdeling.



Figuur 4.4: De standaardnormale verdeling met opdeling in zones

Functie	Betekenis
`pnorm(x, m, s)`	Linkerstaartkans, $P(X < \texttt{x})$
`dnorm(x, m, s)`	Hoogte van de Gausscurve op punt `x`
`qnorm(p, m, s)`	Onder welke grens zal `p`% van de waarnemingen liggen?
`rnorm(n, m, s)`	Genereer `n` normaal verdeelde random getallen

Tabel 4.3: Kansberekeningsfuncties in R voor een normale verdeling met gemiddelde `m` en standaardafwijking `s`. Indien argumenten `m` en `s` weglaten worden, wordt de standaardnormaalverdeling verondersteld.

In het algemeen kan dus bij een waarneming x de zogenaamde z-score bepalen als volgt:

$$z = \frac{x - \mu}{\sigma} \tag{4.3}$$

Deze score geeft dus aan hoe extreem een waarneming is of anders gezegd, hoeveel standaarddeviaties is de waarneming x van het gemiddelde μ verwijderd. Voor een willekeurige x-waarde kunnen we met formule 4.3 de bijhorende z-score bepalen. Voor deze z-scores heeft men tabellen opgesteld met de kansen dat een waarde kleiner dan z getrokken wordt uit Z, de zgn. linkerstaartkans[1]: $P(Z < z)$.

R heeft eveneens functies voor het rekenen met kansen van normaal verdeelde variabelen. Deze worden samengevat in Tabel 4.3.

We komen dan tot de volgende methode voor het berekenen van kansen met de normale verdeling:

1. Bepaal de kansvariabele met de bijbehorende normale verdeling
2. Bereken de z-score bij de bijhorende x-waarde.
3. Schets de plaats van de gevraagde kans
4. Herleid de gevraagde kans met behulp van de schets tot een linkerstaartkans en gebruik de z-tabel van de standaardnormale verdeling om deze te bepalen. Gebruik indien nodig de symmetrieregel en de regel van 100% kans.

[1] Er bestaan ook tabellen met de rechterstaartkans

Voorbeeld 4.3. *Hoe groot is de kans dat superman in minder dan 4 ms reageert?*

$$P(X < 4) = P(Z < -0,67) = 0,2514$$

■

Voorbeeld 4.4. *Hoe groot is de kans dat hij in minder dan 7 ms reageert?*

$$P(X < 7) = P(Z < 1,33) = 0,9082$$

■

Voorbeeld 4.5. *Hoe groot is de kans dat superman in minder dan 3 ms reageert?*

$$P(X < 3) = P(Z < -1,33) = 0,0918$$

■

Voorbeeld 4.6. *Hoe groot is de kans dat hij reageert tussen de 2 en 6,5 ms*

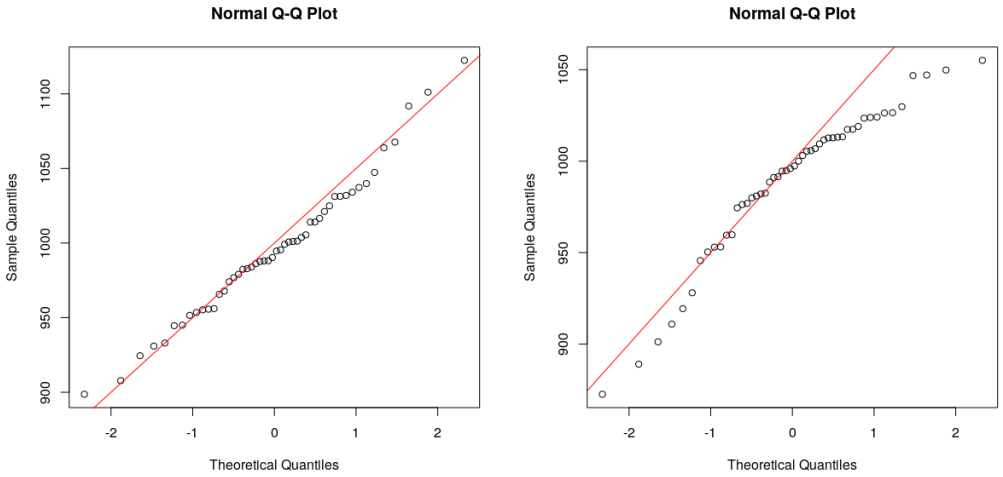
$$P(2 < X < 6,5) = P(X < 6,5) - P(X < 2) = P(Z < 1) - P(Z < -2) = 0,8186$$

■

4.4.2 Testen op normaliteit

Er zijn verschillende methoden die kunnen gebruikt worden om na te gaan of een steekproef uit een normale verdeling komt.

1. Construeer een histogram voor de gegevens en bekijk de vorm van de grafiek. Als de gegevens bij benadering een normale verdeling hebben, zal de vorm van het histogram een klokcurve vormen.
2. Bereken de intervallen $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$ en bepaal het percentage meetwaarden dat binnen elk van deze intervallen valt. Als de gegevens ongeveer normaal verdeeld zijn, zullen de percentages ongeveer gelijk zijn aan respectievelijk 68%, 95% en 99,7%.
3. Construeer een QQ-plot (normaliteitsplot, zie Definitie 4.4.1) voor de gegevens. Als de gegevens ongeveer normaal verdeeld zijn, zullen de punten ongeveer op een rechte lijn liggen.
4. Bereken de *kurtosis* ("welving" of "platheid"): duidt aan hoe scherp de "piek" van de verdeling is.
   - Een normale verdeling heeft een kurtosis = 0
   - Een vlakke distributie heeft een negatieve kurtosis
   - Een eerder piekvormige distributie heeft een positieve kurtosis
   - Let op: bij de originele definitie van kurtosis (zoek die eens op!) heeft de normale verdeling een kurtosis van 3. Wij gebruiken hier een alternatieve definitie, meestal de "excess kurtosis" genoemd, waar men 3 aftrekt van de originele waarde, zodat je op 0 uitkomt.
5. Berken de *Skewness* (scheefheid): duidt aan hoe symmetrisch de data is.
   - Een symmetrische distributie heeft een skewness = 0
   - Bijgevolg: een normale verdeling heeft een skewness = 0.
   - Een distributie met een lange linkerstaart heeft een negatieve skewness
   - Een distributie met een lange rechterstaart heeft een positieve skewness
   - Vuistregel: absolute waarde van skewness > 1, geen symmetrische distributie.



Figuur 4.5: De QQ-plot links is gebaseerd op een steekproef van 50 observaties uit een normale distributie met gemiddelde 1000 en standaardafwijking 50. De rechterplot is gebaseerd op een Student-t distributie met 15 vrijheidsgraden. Het aantal observaties, gemiddelde en standaardafwijking zijn hetzelfde als links. De lijnen in het rood duiden aan waar zich in theorie de observaties zouden moeten bevinden. Links is dat min of meer zo, maar rechts wijken de observaties af, vooral in de extremen.

Definitie 4.4.1 (QQ-plot of normaliteitsplot)**.** *Een normaliteitsplot of QQ-plot*[a] *voor een gegevensverzameling is een spreidingsdiagram met de gesorteerde gegevenswaarden op de ene as en de bijbehorende verwachte z-waarden van een standaardnormale verdeling op de andere as. Zie figuur 4.5 voor enkele voorbeelden. De R-code voor het genereren van deze afbeeldingen is hieronder gegeven.*

[a] Q staat hier voor *quantile*, kwantiel

```r
# Example of a Q-Q plot

m <- 1000
s <- 50
n <- 50

# Normal distribution
observations <- rnorm(n, m, s)
# Student's t distribution
#observations <- m + rt(n, df = 15) * s

# Q-Q plot of observations compared to normal distribution
qqnorm(observations)

# Plots the line of the expected position of observations
x <- seq(-3, +3, length = n)
lines(x, m+s*x, col = 'red')
```

4.5 Centrale limietstelling

Definitie 4.5.1 (Lineaire combinatie van onafhankelijke, gelijk verdeelde stochasten)**.** *Formeel: Een lineaire combinatie van onafhankelijke, gelijk verdeelde stochasten is steeds normaal verdeeld.*

$$X_i \sim Nor(\mu_i, \sigma_i) \Rightarrow Y = \sum_i \alpha_i X_i \text{ ook normaal verdeeld}$$

Bijgevolg zal ook het steekproefgemiddelde van een steekproef uit een populatie met een willekeurige verdeling, nagenoeg normaal verdeeld zijn voor een voldoende grote n.

Wanneer men dus een aselecte steekproef neemt van onafhankelijke variabelen met een normale verdeling, dan zegt de centrale limietstelling dat het gemiddelde van deze steekproef bij benadering normaal verdeeld zal zijn. Dus als men steeds opnieuw een steekproef neemt met dezelfde grootte, en telkens het gemiddelde optekent, bekomt men bij benadering de grafiek van een normale verdeling. Hoe groter de steekproef, hoe beter de benadering. Het steekproefgemiddelde is dus normaal verdeeld, onafhankelijk van de onderliggende verdeling van de grootheid waarvan men een steekproef neemt. Algemeen kunnen we volgende stelling poneren:

Definitie 4.5.2 (Centrale limietstelling)**.** *Beschouw een aselecte steekproef van n waarnemingen die uit een populatie met verwachtingswaarde μ en standaardafwijking σ wordt genomen. Als n groot genoeg is zal de kansverdeling van het steekproefgemiddelde $\bar{x}$ een normale verdeling benaderen met verwachting $\mu_{\bar{x}} = \mu$ en standaardafwijking $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Hoe groter de steekproef is, des te beter zal de kansverdeling van $\bar{x}$ de verwachtingswaarde van de populatie benaderen.*

Bij het afnemen van een steekproef is zelden de onderliggende verdeling gekend, en toch kan men uitspraken doen over de gemiddelde waarde. Dit is volledig te danken aan de centrale limietstelling, die dit gemiddelde een regel oplegt los van de onderliggende kansverdeling. De centrale limietstelling houdt het steekproefgemiddelde in bedwang, sluit het op in de Gaussische kooi waaruit het nooit kan ontsnappen. Dit, en alleen dit, laat wetenschappers toe het nauwkeurig te bestuderen, te observeren en stelt hen in staat te concluderen.

Want, mocht de verdeling van het steekproefgemiddelde afhankelijk zijn van de onderliggende verdeling, een resultaat dat men tot op zekere hoogte zelfs zou verwachten, zou het onmogelijk zijn om concrete uitspraken te doen over vele wetenschappelijke resultaten. In de theoretische statistiek duiken vrijwel constant limieten van steekproefgemiddeldes op, en deze kunnen dankzij de centrale limietstelling zonder verpinken vervangen worden door een normale verdeling. Zou dit niet mogelijk zijn, dan zou de ganse theorie rond het schatten van parameters in elkaar storten wat dan weer rampzalig zou zijn voor de praktijk. Onderzoeken vergelijken zou herleid worden tot een quasi onmogelijke opgave, en de statistiek in het algemeen zou veel lastiger en ingewikkelder worden.

4.5.1 Toepassing van de centrale limietstelling

Bij het trekken van een aselecte steekproef van omvang n uit een populatie met (onbekend) gemiddelde μ en standaarddeviatie σ is de kansverdeling van het steekproefgemiddelde een

kansvariabele $M \sim N(\bar{x}, \frac{\sigma}{\sqrt{n}})$, op voorwaarde dat de steekproefomvang voldoende groot is.

Voorbeeld 4.7. *We bekijken nu de reactiesnelheid van al onze superhelden en uit onze steekproef met* $n = 100$ *en* $\bar{x} = 90, \sigma = 60$ *(miliseconden). Dan kunnen we ons de vraag stellen: wat is de kans dat de gemiddelde reactiesnelheid van een superheld minder is dan* $104ms$*?*

1. *De kansvariabele hier is de gemiddelde reactiesnelheid* $\bar{x}$ *in een steekproef van* $n = 100$ *superhelden. Daarom geldt wegens de centrale limietstelling:*

$$\bar{x} \sim Nor(\mu = 90, \sigma_{\bar{x}} = \frac{60}{\sqrt{100}} = 6)$$

2. *We kunnen hierbij de passende z-score bepalen:*

$$z = \frac{104 - 90}{\frac{60}{\sqrt{100}}} = \frac{104 - 90}{6} = 2,33$$

Dus geldt : $P(\bar{x} < 104) = P(Z < 2,33) = 1 - 0,0099 \approx 0,99$

■

4.5.2 Schatten van een parameter

Indien we nu een steekproef onderzoeken, willen we uit de berekening op de steekproef een aantal conclusies kunnen trekken met betrekking tot de populatie. We willen bijvoorbeeld de gemiddelde kracht kennen van een superheld of de fractie superhelden die rijk zijn. Als we een schatting geven voor dergelijke onbekende parameter, noemen we dat ook een puntschatter. We gebruiken bijvoorbeeld $\bar{x}$ als schatter om μ te schatten.

Definitie 4.5.3 (puntschatter)**.** *Een puntschatter voor een populatieparameter is een regel of een formule die ons zegt hoe we uit de steekproef een getal moeten berekenen om de populatieparameter te schatten. Een puntschatter is dus een steekproefgrootheid.*

4.5.3 Betrouwbaarheidsinterval populatiegemiddelde bij grote steekproef

In het geval het schatten van een gemiddelde van een populatie uit een steekproef hebben we totaal geen idee over hoe correct onze schatting is. Daarvoor gaan we op zoek naar een interval waarvan we met een bepaalde zekerheid, bv. 95%, kunnen zeggen dat het de te schatten karakteristiek bevat.

Definitie 4.5.4 (Betrouwbaarheidsinterval)**.** *Een betrouwbaarheidsinterval is een regel of een formule die ons zegt hoe we uit de steekproef een interval moeten berekenen dat de waarde van de parameter met een bepaalde hoge waarschijnlijkheid bevat.*

Een eerste goede schatting voor populatiegemiddelde zou het steekproefgemiddelde zijn:

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

Natuurlijk is deze schatting niet de werkelijke waarde van de populatie. Daarom wordt vaak rondom $\bar{x}$ een interval geconstrueerd dat de waarden bevat die aannemelijk zijn voor μ. Hiervoor kunnen

we gebruik maken van de centrale limietstelling: het gemiddelde in een te trekken steekproef van omvang n is normaal verdeeld met karakteristieken μ en $\frac{\sigma}{\sqrt{n}}$. Als we nu het gemiddelde standaardiseren krijgen we:

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Deze uitdrukking hangt van μ af maar we weten wel dat deze standaardnormaal verdeeld is. We kunnen daarom getallen $-z$ en z vinden, onafhankelijk van μ, waartussen Z met een gekozen kans $1-\alpha$ ligt. Deze kans $1-\alpha$ wordt het *betrouwbaarheidsniveau* genoemd. We nemen hier $1-\alpha = 0,95$.

$$P(-z < Z < z) = 1 - \alpha = 0,95$$

Hieruit halen we dat $\alpha = 0,05$. Door het toepassen van de symmetrieregel weten we dus dat we volgende term moeten berekenen:

$$P(Z < z) = 0,025$$

Kijken we in de Z-tabel dan vinden we voor de rechterstaartkans $0,025$ de z-score van $1,96$.

Dus vinden we :

$$P(-1,96 < \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1,96)$$

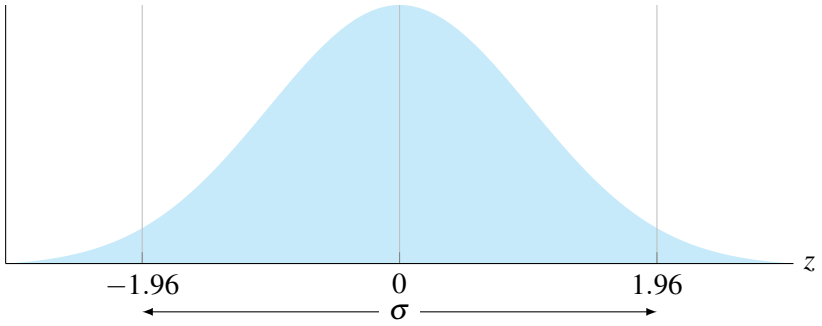
en dus

$$P(\overline{x} - 1,96\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1,96\frac{\sigma}{\sqrt{n}})$$

Op die manier kunnen we dus grenzen bepalen die een interval aanduidt waar 95% kans is dat μ gevonden wordt. Formeel: als je herhaalde steekproeven zou nemen en telkens op basis van het gerealiseerde steekproefgemiddelde $\overline{x}$ een betrouwbaarheidsinterval zou maken, dan zal bij 95% van de intervallen μ binnen de intervalgrenzen liggen.

Opgelet, we gaan er hier van uit dat we de standaarddeviatie van de populatie kennen, wat meestal niet zo is. Indien de steekproef groot genoeg is, kunnen we de steekproefstandaarddeviatie nemen als schatter voor de standaarddeviatie voor de populatie.

$$P(\overline{x} - 1,96\frac{\sigma_{\overline{x}}}{\sqrt{n}} < \mu < \overline{x} + 1,96\frac{\sigma_{\overline{x}}}{\sqrt{n}})$$



Figuur 4.6: Standaardnormale verdeling die 95% betrouwbaarheidsinterval aanduidt.

4.5.4 Betrouwbaarheidsinterval populatiegemiddelde bij een kleine steekproef

Bij kleine steekproeven kunnen we niet langer veronderstellen dat de kansverdeling van $\bar{x}$ bij benadering normaal verdeel is, omdat de centrale limietstelling alleen normaliteit garandeert voor grote steekproeven ($n > 30$). De vorm van de kansverdeling van het steekproefgemiddelde $\bar{x}$ hangt nu af van de vorm van de verdeling van de populatie waaruit de steekproef genomen wordt. Alhoewel nog steeds geldt dat $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ kan de standaardafwijking s een slechte benadering zijn voor σ als de steekproef klein is.

Als oplossing kunnen we een nieuwe grootheid bepalen. In plaats van

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

construeren we

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Deze heeft een kansverdeling die beschreven wordt door een Student-t verdeling. Deze lijkt zeer goed op de normale verdeling: klokvormig, symmetrisch en met verwachtingswaarde 0.

De precieze vorm van de kansverdeling t hang af van de steekproefomvang n. We zeggen dat de t-verdeling $(n-1)$ vrijheidsgraden heeft (afgekort df). Merk op dat:

- $(n-1)$ ook gebruikt werd om s^2 te berekenen
- als $n \to \infty$ we de standaardnormale verdeling verkrijgen.

Indien we nu een betrouwbaarheidsinterval willen bepalen voor een steekproef met een klein aantal waarden moeten we het volgende doen:

Definitie 4.5.5 (Betrouwbaarheidsinterval kleine steekproef)**.** *Om een betrouwbaarheidsinterval*

Functie	Betekenis
`pt(x, df)`	Linkerstaartkans, $P(X < \mathtt{x})$
`dt(x, df)`	Hoogte van de curve op punt `x`
`qt(p, df)`	Onder welke grens zal `p`% van de waarnemingen liggen?
`rt(n, df)`	Genereer `n` random getallen volgens deze verdeling

Tabel 4.4: Kansberekeningsfuncties in R voor de Student-t verdeling met `df` vrijheidsgraden, verwachte waarde 0 en standaardafwijking 1.

voor het gemiddelde te bepalen op basis van een klein steekproef bepalen we:

$$\overline{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right)$$

waarbij $t_{\frac{\alpha}{2}}$ gebaseerd is op $(n-1)$ vrijheidsgraden. We veronderstellen wel dat we een aselecte steekproef genomen hebben uit een populatie die bij benadering normaal verdeeld is.

4.5.5 Betrouwbaarheidsinterval voor populatiefractie bij een grote steekproef

Indien je een variabele wil meten als een fractie, bijvoorbeeld % mensen die ja geantwoord heeft op een bepaalde vraag, dan willen we in feite de kans p op succes in een bernouilli experiment schatten, waarbij p de kans is dat een willekeurig geselecteerde respondent (of element van de populatie) een succes is (succes in termen van binomiaal experiment). We kunnen p dan schatten door bijvoorbeeld:

$$\overline{p} = \frac{\text{aantal successen}}{n}$$

Om nu de betrouwbaarheid van de schatter $\overline{p}$ te bepalen moeten we de kansverdeling kennen van $\overline{p}$. Dit kunnen we beredeneren door toepassing van de centrale limietstelling op het gemiddelde aantal successen in de steekproef van omvang n. Indien succes = 1 en faling = 0, dan hebben we een steekproef van n elementen, ieder met dezelfde verdeling (kans op 1 is p en kans op 0 is $q = 1 - p$). Het gemiddelde $\overline{p}$ heeft dan bij benadering een normale verdeling. Of dus:

- Verwachting van kansverdeling van $\overline{p}$ is p.
- De standaardafwijking van kansverdeling $\overline{p} = \sqrt{\frac{pq}{n}}$
- Voor grote steekproeven is $\overline{p}$ bij benadering normaal verdeeld.

Aangezien $\overline{p}$ een steekproefgemiddelde is van het aantal successen, stelt dit ons in staat een betrouwbaarheidsinterval te berekenen analoog als die voor de intervalschatting van μ voor grote steekproeven.

Definitie 4.5.6 (Betrouwbaarheidsinterval voor p gebaseerd op grote steekproef)**.**

$$\overline{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{pq}}{n}}$$

met $\overline{p} = \frac{x}{n}$ *en* $\overline{q} = 1 - \overline{p}$

4.6 R

We kijken naar enkele basisoperaties die verband houden met enkele distributies. Er zijn een groot aantal verdelingen beschikbaar, maar we kijken maar naar een paar. Als u wilt weten welke distributies beschikbaar zijn, kunt u een zoekopdracht uitvoeren met behulp van de opdracht

```
> help.search ("distribution").
```

Hier geven we details over de commando's die verband houden met de normale distributie en vermelden kort de commando's voor andere distributies. De functies voor verschillende verdelingen zijn zeer vergelijkbaar.

De prefixen zijn als volgt:

d geeft de hoogte van de respectievelijke kansdichtheidsfunctie
p geeft de cumulatieve kansdichtheidsfunctie
q geeft de omgekeerde cumulatieve dichtheidsfunctie
r geeft een willekeurige waarde

4.6.1 De normale verdeling

Er zijn vier functies die kunnen worden gebruikt om de waarden geassocieerd met de normale distributie te genereren.

dnorm

De eerste functie waarnaar we kijken, is `dnorm`. Gegeven een waarde geeft het de hoogte van de kansverdeling op elk punt terug. Als u alleen de punten zonder gemiddelde en standaardafwijking ingeeft wordt een gemiddelde van nul en standaardafwijking van 1 beschouwd. Er zijn opties om verschillende waarden voor de gemiddelde en standaardafwijking te gebruiken.

```
> dnorm(0)
[1] 0.3989423
> dnorm(0)*sqrt(2*pi)
[1] 1
> dnorm(0,mean=4)
[1] 0.0001338302
> dnorm(0,mean=4,sd=10)
[1] 0.03682701
>v <- c(0,1,2)
> dnorm(v)
[1] 0.39894228 0.24197072 0.05399097
> x <- seq(-20,20,by=.1)
> y <- dnorm(x)
> plot(x,y)
> y <- dnorm(x,mean=2.5,sd=0.1)
```

```
> plot(x,y)
```

pnorm

Dit is de cumulatieve kansdichtheidsfunctie, of anders gezegd de linkerstaartkans: `pnorm(x)` is $P(Z < x)$.

qnorm

De volgende functie die we bekijken is `qnorm`, die de inverse van `pnorm` is. Het idee achter `qnorm` is dat je het een kans α geeft, en het geeft het getal weer waarvan de cumulatieve distributie overeenkomt met de waarschijnlijkheid α.

rnorm

```
> qnorm(0.5)
[1] 0
> qnorm(0.5,mean=1)
[1] 1
> qnorm(0.5,mean=1,sd=2)
[1] 1
> qnorm(0.5,mean=2,sd=2)
[1] 2
> qnorm(0.5,mean=2,sd=4)
[1] 2
> qnorm(0.25,mean=2,sd=2)
[1] 0.6510205
> qnorm(0.333)
[1] -0.4316442
> qnorm(0.333,sd=3)
[1] -1.294933
> qnorm(0.75,mean=5,sd=2)
[1] 6.34898
> v = c(0.1,0.3,0.75)
> qnorm(v)
[1] -1.2815516 -0.5244005  0.6744898
> x <- seq(0,1,by=.05)
> y <- qnorm(x)
> plot(x,y)
> y <- qnorm(x,mean=3,sd=2)
> plot(x,y)
> y <- qnorm(x,mean=3,sd=0.1)
> plot(x,y)
```

De laatste functie die we onderzoeken is de `rnorm` functie die willekeurige getallen kan genereren waarvan de distributie normaal is. Het argument dat je ingeeft is het aantal willekeurige getallen dat u wilt, met optionele argumenten om de gemiddelde en standaardafwijking op te geven:

```
> rnorm(4)
[1]  1.2387271 -0.2323259 -1.2003081 -1.6718483
> rnorm(4,mean=3)
[1] 2.633080 3.617486 2.038861 2.601933
> rnorm(4,mean=3,sd=3)
[1] 4.580556 2.974903 4.756097 6.395894
> rnorm(4,mean=3,sd=3)
[1]  3.000852  3.714180 10.032021  3.295667
> y <- rnorm(200)
> hist(y)
> y <- rnorm(200,mean=-2)
> hist(y)
> y <- rnorm(200,mean=-2,sd=4)
> hist(y)
> qqnorm(y)
> qqline(y)
```

4.7 Oefeningen

Oefening 4.1. *Een onderzoeker wil zo correct mogelijk de consumptiegewoontes van de inwoners van 18 jaar en ouder in een bepaalde gemeente, met 3 woonkernen, onderzoeken. Hij onderscheidt 4 leeftijdsgroepen zodat hij uiteindelijk aan 12 deelgroepen komt. Hij vraagt de procentuele samenstelling van de bevolking op in de gemeente en berekent daaruit hoeveel bevragingen hij per deelgroep moet uitvoeren. Dit noemen we een* quotasteekproef.

Vragen:

a. Wat zijn de voor- en nadelen?
b. Welke soort fouten kunnen hier gemaakt worden?
c. Welke andere parameters zouden kunnen gebruikt worden bij het opsplitsen in deelgroepen?

■

Oefening 4.2. *Een onderzoeksbureau wil het aankoopgedrag van wasproducten nagaan. Men beslist een aantal vragen te stellen aan vrouwen tussen de 25 en 55 jaar omdat men ervan uitgaat dat de relevante populatie uit deze categorie consumenten bestaat.*

Vraag:

a. Welke fout wordt hier gemaakt?
b. Hoe groot is de impact van deze fout?

■

Oefening 4.3. *De vakbonden willen een onderzoek doen naar de werkomstandigheden van de werknemers van een IT-bedrijf. Dat bedrijf heeft in totaal 3200 werknemers die verdeeld zijn*

over 12 vestigingen. Omdat het aantal werknemers groot is worden aselect 40 werknemers gekozen per vestiging. De steekproefomvang is dus $n = 480$.

a. Welk bezwaar kan tegen deze steekproefprocedure worden gebracht?
b. Wanneer zou dit geen bezwaar zijn?

■

Oefening 4.4. *We willen een onderzoek voeren naar onze studenten aan de Hogeschool Gent, faculteit Bedrijf en Organisatie. Hiervoor worden de aanwezige studenten in een bepaald opleidingsonderdeel bevraagd.*

a. Welke kritiek kan je op deze methode geven?
b. Stel dat de aanwezige docent een kernvak geeft, zeer streng is en tijdens de bevraging rondloopt. Welk bezwaar kan hier gegeven worden?
c. Stel dat de bevraging niet tijdens een les, maar na een examen gehouden wordt. Welke kritiek kan je op deze methode geven?

■

Oefening 4.5. *Bereken ook elke keer het gevraagde gebied.*

a. $P(Z < 1.33)$
b. $P(Z > 1.33)$
c. $P(Z < -1.33)$
d. $P(Z > -1.33)$
e. $P(Z < 0.45)$
f. $P(Z > -1.05)$
g. $P(Z < 0.65)$
h. $P(-0.45 < Z < 1.20)$
i. $P(-1.35 < Z < -0.10)$
j. $P(-2.10 < Z < -0.90)$

■

Oefening 4.6. *Bepaal de dichtheid en de cumulatieve waarschijnlijkheidscurve voor een normale verdeling met een gemiddelde μ van 2,5 en $\sigma = 1,5$. Bepaal de oppervlakte voor het gebied onder de dichtheidscurve tussen $x = 0.5$ en $x = 4$. Controleer uw antwoord door de berekening te doen.* ■

Oefening 4.7. *Bepaal de dichtheid en de cumulatieve waarschijnlijkheidscurve voor een t-verdeling met $df = 3$. Teken ook een normale verdeling met een $\mu = 0$ en $\sigma = 1$.* ■

Oefening 4.8. *Gebruik de functie* `rnorm()` *een willekeurige steekproef van 25 waarden uit een normale verdeling te tekenen met een gemiddelde van 0 en een standaardafwijking gelijk aan 1,0. Gebruik een histogram, met* `probability = TRUE`.

Maak een overlay over het histogram met: (a) de theoretische dichtheidscurve voor een normale verdeling met gemiddelde 0 en standaardafwijking gelijk aan 1,0; (b) een "geschatte" dichtheidscurve op basis van het gemeten steekproefgemiddelde en -standaardafwijking.

Herhaal dit voor een steekproef van 100 en 500 waarden. ■

Oefening 4.9. *In de Hogeschool zijn er twee klassen voor het vak onderzoekstechnieken. De studenten werden willekeurig over de klassen verdeeld, zodat we mogen veronderstellen dat de ene klas niet slimmer is dan de andere. In de A-klas geeft mevr. X les, in de B-klas geeft mr. Y les. X is nogal streng en op het einde van het schooljaar behaalt haar klas een gemiddelde van 54 op 100 met een standaardafwijking van 11.*

Y is iets losser en stimuleert de leerlingen al gauw met een puntje meer. Op het einde van het schooljaar behaalt zijn klas een gemiddelde van 62 op 100 en een standaardafwijking van 7.

Wouter zit in de A-klas en heeft $\frac{63}{100}$ voor wiskunde. Stijn zit in de B-klas en behaalt $\frac{67}{100}$. Wie heeft volgens jou het beste gescoord binnen de eigen klas? ■

Oefening 4.10. *Een gezondheidsonderzoek tussen 1988 en 1994 gaf aan dat de gemiddelde cholesterolwaarde bij vrouwen tussen 20 en 29 jaar 183 mg/dl bedroeg, met een standaardafwijking gelijk aan 36. We nemen nu een aselecte steekproef van 81 vrouwen. Los volgende vragen op:*

a. *Schets de kansdichtheidsfunctie voor de populatie en de kansverdeling van het steekproefgemiddelde $\bar{x}$.*
b. *Bepaald de kans dat $\bar{x}$ kleiner is dan 185.*
c. *Bepaal de kans dat $\bar{x}$ tussen 175 en 185 ligt.*
d. *Bepaal de kans dat $\bar{x}$ groter is dan 190.*

■

Oefening 4.11. *Een aselecte steekproef van 64 stuks wordt getrokken uit een populatie met onbekende verdeling. De verwachting en de standaardafwijking van de populatie zijn wel gekend: $\mu = 20$ en $\sigma = 16$. Los volgende vragen op:*

a. *Bepaal de verwachting en standaardafwijking van het steekproefgemiddelde.*
b. *Beschrijf de vorm van de verdeling van het steekproefgemiddelde. In hoeverre hangt je antwoord af van de grootte van de steekproef?*
c. *Bereken de z score bij $\overline{x_1} = 15.5$ en $\overline{x_2} = 23$.*
d. *Bepaal kans dat $\bar{x} < 16$.*
e. *Bepaal kans dat $\bar{x} > 23$.*
f. *Bepaal kans dat $16 < \bar{x} < 22$.*

■

Oefening 4.12. *Verkeersdrempels zijn bedoeld om de snelheid van automobilisten te beïnvloeden. Afhankelijk van de gewenste snelheid in een straat worden de drempels steiler of minder steil gemaakt. Drempel A is zo ontworpen dat 85 % van de automobilisten de drempel passeert met*

11.5	16.5	11	17.3	10.8	5.6	13.1	11.5	14.2	12.9
8.7	9.2	15	14.4	10	10.3	18.3	12.9	14.2	8.7

Tabel 4.5: Examenresultaten

een snelheid van minder dan 50 km per uur. In de praktijk blijkt dat de passeersnelheid bij een drempel normaal verdeeld is. Bij drempel A werd een gemiddelde passeersnelheid van 43,1 km/h gevonden met standaardafwijking 6,6 km/h.

a. *Toon aan dat 85% van de automobilisten niet harder dan 50 km/h rijdt.*
b. *Bij hoeveel van de 1200 metingen kan, op grond van eerdere ervaringen, een snelheid van meer dan 55 km/h worden verwacht?*

■

Oefening 4.13. *Gegeven 20 examenresultaten in Tabel 4.5. Uit resultaten van de laatste jaren blijkt dat* $\sigma = 2.45$.

a. *Wat is* $\sigma_{\bar{x}}$ *, de standaardafwijking van* $\bar{x}$*?*
b. *Geef het 92% betrouwbaarheidsinterval voor* μ.
c. *Kunnen we er zeker van zijn dat het gemiddeld resultaat minder dan 12.5 bedraagt?*

■

Oefening 4.14. *Een schoenhandelaar voert een marktonderzoek uit bij 500 klanten. Daaruit blijkt dat 30% van hen minstens éénmaal per jaar sportschoenen koopt. Op basis van secundaire informatie weet hij dat het nationaal gemiddelde op 26% ligt. Hij vraagt zich nu af in hoeverre zijn zaak in dat opzicht afwijkt van de nationale norm? (We werken met* $\alpha = 5\%$*, tweezijdig.)* ■

Oefening 4.15. *Een conservenfabrikant krijgt de laatste tijd klachten over de netto inhoud van zijn conserven met wortelen en erwtjes, die volgens de verpakking netto 1 liter zouden moeten bevatten. Daarom laat hij een steekproef nemen waarin de netto inhoud van 40 willekeurig gekozen blikjes wordt gecontroleerd. De resultaten worden samengevat in Tabel 4.6.*

Vraag A:

- *Vul de tabel aan met de cumulatieve absolute frequentie*
- *Vul de tabel aan met de relatieve frequentie*
- *Vul de tabel aan met de cumulatieve relatieve frequentie.*

Vraag B:

- *Bereken het gemiddelde*
- *Bereken de standaardafwijking*
- *Hoeveel procent van de blikken bevatten te weinig wortelen en erwtjes.*
- *Teken een histogram van de absolute frequentie.*
- *Zijn de gegevens normaal verdeeld? Hoe zie je dat?*

■

Inhoud	n_i
$[970, 980[$	3
$[980, 990[$	5
$[990, 1000[$	13
$[1000, 1010[$	11
$[1010, 1020[$	5
$[1020, 1030[$	3

Tabel 4.6: Steekproefwaarden

Oefening 4.16. *Een webhostingfirma heeft een Service Level Agreement met een klant voor een gegarandeerde uptime van "five nines" (99,999%). Die wordt aan het einde van elk jaar gecontroleerd en als de minimale uptime niet gehaald wordt, moet de hostingfirma een boete betalen.*

Om de uptime te meten, voert een monitoringsysteem elke minuut een `HTTP GET /` *uit en controleert het resultaat a.h.v. de HTTP return code. In de maand januari is er één enkele HTTP request onsuccesvol geweest.*

- *Als deze trend zich voortzet, wat is de kans dat de SLA niet gehaald wordt aan het einde van het jaar? Gebruik de formule voor de kansverdeling van een fractie.*
- *De gebruikte formule is eigenlijk niet geschikt in dit specifieke geval en geeft een vertekend beeld. Wat zou de reden kunnen zijn?*

■

4.8 Antwoorden op geselecteerde oefeningen

Oefening 4.5

a. 0,908
b. 0,092
c. 0,092
d. 0,908
e. 0,674
f. 0,853
g. 0,742
h. 0,559
i. 0,372
j. 0,166

5. Toetsingsprocedures

In de voorbije hoofdstukken hebben we gezien hoe we aan de hand van steekproefonderzoek bepaalde kerngetallen over een populatie kunnen berekenen, bijvoorbeeld aan de hand van puntschatters of betrouwbaarheidsintervallen. We kunnen deze informatie ook gebruiken om bepaalde hypothesen over een populatie te toetsen. Een hypothese is een veronderstelling waarvan nog bewezen moet worden dat ze correct is. Het doel van een toetsingsprocedure is het testen van een hypothese omtrent de waarden van 1 of meerdere populatieparameters.

Definitie 5.0.1 (Statistische hypothese.)**.** *Een statistische hypothese is een uitspraak over de numerieke waarde van een populatieparameter.*

Voorbeelden van hypothesen:

- Gemiddeld redt een superheld minstens 3,3 mensen per dag.
- De gemiddelde lengte van een superheld is minstens 120 cm.
- ...

In dit hoofdstuk gaan we de algemene theorie over toetsen formuleren aan de hand van het testen van hypothesen over het populatiegemiddelde μ, de z-toets. Naast de z-toets bestaan er echter nog vele andere statistische hypothesetoetsen die in specifieke situaties gebruikt kunnen worden. De meest geschikte statistische toets hang o.a. af van de populatiegrootheid in kwestie (gemiddelde, standaardafwijking, enz.), en veronderstellingen over de onderliggende stochastische verdeling van de populatie (normaal verdeeld of niet, enz,).

5.1 Elementen van een hypothesetoets

Algemeen gezien bestaat een toetsingsprocedure uit 4 elementen:

1. **Nulhypothese** H_0: Deze hypothese proberen we te ontkrachten door een redenering in het ongerijmde. We gaan deze hypothese accepteren, tenzij de observaties uit de steekproef overtuigend wijzen op het tegendeel.
2. **Alternatieve hypothese** H_1: Dit is meestal de hypothese die de onderzoeker wil bewijzen. Deze hypothese zal echter alleen worden geaccepteerd als de observaties uit de steekproef overtuigend wijzen op de juistheid ervan.
3. **Teststatistiek**: De veranderlijke die berekend wordt uit de steekproef
4. Aanvaardings- en kritiek gebied:
   - **Aanvaardingsgebied**: Het gebied van waarden die de nulhypothese H_0 ondersteunt
   - **Verwerpingsgbied**: gebied dat waarden bevat die de nulhypothese verwerpen. Ook kritiek gebied genoemd.

Een alternatief voor de laatste stap is het berekenen van de *overschrijdingskans* (zie verder).

De beslissing om de nulhypothese H_0 te verwerpen of te aanvaarden is gebaseerd op informatie uit een steekproef, getrokken uit de populatie waarover de hypothese is geformuleerd. De steekproefwaarden worden gebruikt om 1 enkele waarde van een teststatistiek te berekenen die de beslissing zal bepalen. Daartoe worden alle waarden die de teststatistiek kan aannemen, verdeeld in twee gebieden, (i) het aanvaardingsgebied en (ii) het verwerpingsgebied. Indien de waarde van de teststatistiek ligt in het verwerpingsgebied, dan wordt de nulhypothese verworpen en de alternatieve hypothese aanvaard. Indien de waarde van de teststatistiek in het aanvaardingsgebied valt dan wordt de nulhypothese aanvaard.

5.2 Toetsingsprocedure voor de z-toets

In de eerste toetsingsprocedure die we in deze cursus uitwerken, gaan we een uitspraak over het populatiegemiddelde μ verifiëren. Deze is algemeen bekend als de z-toets.

1. De vermoedens over de populatie worden vastgelegd in twee hypothesen H_0 en H_1. Voor de (rechtszijdige) z-toets is de nulhypothese dat het populatiegemiddelde μ een bepaalde waarde heeft, en de alternatieve hypotehese dat μ *groter* is.
2. Het significantieniveau α en steekproefomvang n worden vastelegd. Je kan α in principe zelf kiezen (bv. 0,05)[1]. Hoe dichter het significantieniveau bij 0 ligt, hoe minder twijfel er is over het resultaat van de toets. Maar langs de andere kant wordt het ook moeilijker om de nulhypothese te verwerpen.
3. De waarde van de toetsingsgrootheid in de steekproef wordt berekend. De uitkomst is bepalend voor de beslissing of we de nulhypothese H_0 kunnen verwerpen of niet. We weten dankzij de centrale limietstelling dat de kansverdeling van het steekproefgemiddelde $M \sim Nor(\mu, \frac{\sigma}{\sqrt{n}})$.
4. Het kritieke gebied bepalen, of meer bepaald de grens tussen het aanvaardings- en het verwerpingsgebied. We zoeken een grenswaarde g zodat:

$$\begin{aligned} P(M > g) = \alpha \Leftrightarrow P\left(Z > z = \frac{g-\mu}{\sigma/\sqrt{n}}\right) &= \alpha \qquad (normalisatie) \\ \Leftrightarrow P\left(Z < z = \frac{g-\mu}{\sigma/\sqrt{n}}\right) &= 1-\alpha \quad (100\% - regel) \end{aligned} \tag{5.1}$$

[1] Merk op dat het significantieniveau gerelateerd is aan het betrouwbaarheidsniveau $1-\alpha$. Zie Sectie 4.5.3

De z-waarde hangt af van het gekozen significantieniveau en kan worden opgezocht in een z-tabel of berekend worden met de R-functie `qnorm(1-alpha)`. Daaruit kunnen we dan g afleiden:

$$z = \frac{g-\mu}{\sigma/\sqrt{n}} \Leftrightarrow g = \mu + z\frac{\sigma}{\sqrt{n}} \tag{5.2}$$

Alle waarden *links* van g vormen het aanvaardingsgebied. Waarden rechts, die dus ver van het H_0 veronderstelde populatiegemiddelde liggen, zijn het verwerpingsgebied. Zie ook Sectie 5.3.

Voorbeeld 5.1. *Algemeen wordt aangenomen dat superhelden gemiddeld 3,3 mensen per dag redden. De onderzoekers krijgen echter gevoel dat dat niet zo is: ze hebben de indruk dat een superheld* meer *dan* $3,3$ *mensen per dag redt.*

Ze gaan dit onderzoeken en voeren een steekproef uit bij $n = 30$ *superhelden. In deze steekproef is het gemiddelde* $\bar{x} = 3,483$ *is. De standaardafwijking in de populatie is verondersteld gekend en is* $\sigma = 0,55$.

Kan hieruit besloten worden dat superhelden gemiddeld meer dan 3,3 mensen per dag redt?

1. *We veronderstellen dat het aantal mensen dat een superheld redt normaal verdeeld is en formuleren twee hypothesen omtrent de parameter* μ.
   - H_0 = *de nulhypothese (hetgeen we willen weerleggen). In dit geval*

     $$H_0 : \mu = 3,3$$

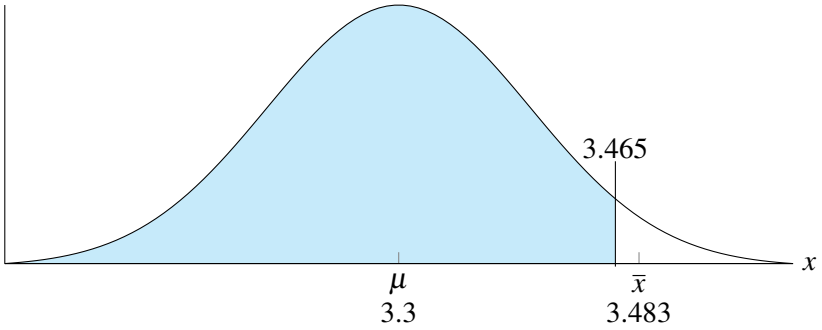
   - H_1 = *alternatieve hypothese (vermoeden dat we willen aantonen). In dit geval*

     $$H_1 = \mu > 3,3$$

   We veronderstellen in de redenering initieel dat de nulhypothese H_0 *waar is. Indien het gemiddelde aantal mensen gered per dag* $\bar{x}$ *van de steekproef sterk afwijkt van de veronderstelde waarde, verwerpen we de nulhypothese* H_0 *en aanvaarden we de alternatieve hypothese* H_1. *Wat betekent nu "sterk afwijken"? Zou je uit een populatie met gemiddelde van* $3,3$ *gemakkelijk een steekproef kunnen trekken met gemiddelde* $3,483$*? De centrale limietstelling (zie Sectie 4.5) laat ons toe de kans hiertoe te berekenen.*
2. *Vastleggen significantieniveau* α *en steekproefomvang n. We willen een significantieniveau van 5% kiezen, dus* $\alpha = 0,05$. *De steekproefomgang is gegeven en is hier* $n = 30$.
3. *De waarde van de toetsingsgrootheid in de steekproef bepalen. We nemen hier het steekproefgemiddelde:* $\bar{x} = 3,483$
   We veronderstellen in de redenering dat de nulhypothese H_0 *waar is en dat we* σ *goed kunnen schatten hebben (*$\sigma = 0,55$*). Dan geldt voor het steekproefgemiddelde M volgens de centrale limietstelling dat:*

   $$M \sim Nor(\mu = 3,3; \sigma = \frac{0,55}{\sqrt{30}})$$

   De waarde $\bar{x} = 3,483$ *bevindt zich erg rechts (zie Figuur 5.1).* $\bar{x}$ *ligt zelfs zo ver naar rechts dat de kans (indien* H_0 *waar is) om dergelijke geobserveerde waarde te krijgen of groter, zeer klein is. Een dergelijke geobserveerde waarde onder de nulhypothese kan dus moeilijk verklaard worden door louter toeval. Intuïtief voelen we dus aan dat hoe verder de geobserveerde waarde* $\bar{x}$ *zich bevindt in de rechtse richting, hoe meer we geneigd zijn om de nulhypothese te verwerpen. Maar wat is te ver en wat niet?*



Figuur 5.1: Verdeling van het aantal mensen dat gemiddeld per dag gered wordt door een superheld (Voorbeeld 5.1). De kansverdeling voor het steekproefgemiddelde is normaal verdeeld met $\mu = 3,3$ en $\sigma = 0,55$. Het steekproefgemiddelde $\bar{x} = 3,483$. De grens voor aanvaarding/verwerping van H_0 is 3.465.

4. *De kritieke grenswaarde berekenen. De z-waarde voor een significantieniveau van* $0,05$ *is 1,645*[2].

$$g = \mu + z \times \frac{\sigma}{\sqrt{n}} = 3,3 + 1,645 \times \frac{0,55}{\sqrt{30}} \approx 3,45$$

Het steekproefgemiddelde $\bar{x} = 3,483$ *ligt nog verder van* $\mu = 3,3$ *dan de grenswaarde* $g = 3,45$. *De kans is heel klein dat zo'n steekproef getrokken wordt uit een populatie met dit gemiddelde. Slechts in 34 steekproeven op 1000 zal een dergelijke gebeurtenis optreden. Met andere woorden, de steekproefwaarde ligt in het verwerpingsgebied. We kunnen dus* H_0 *verwerpen en besluiten met dat superhelden inderdaad* meer *dan 3,3 mensen per dag redden.*

■

Oefening 5.1. *Kunnen we in Voorbeeld 5.1 zomaar veronderstellen dat het gemiddelde normaal verdeeld is? Waarom (niet)?* ■

5.3 Kritieke gebied

De formule voor de berekening van de grenswaarde (zie Formule 5.2) is gebaseerd op de centrale limietstelling, en meer bepaald betrouwbaarheidsintervallen.

De kritieke grenswaarde vormt een betrouwbaarheidsinterval rond μ met een gekozen zekerheidsniveau. Als we bijvoorbeeld stellen dat $\alpha = 0.05$, weten we vanuit de centrale limietstelling dat we kunnen verwachten dat als we herhaaldelijk voldoende steekproeven uit deze populatie nemen, in 95% van de gevallen het steekproefgemiddelde binnen dit betrouwbaarheidsgeval zal liggen.

Als we de redenering omkeren, en een steekproef genomen hebben waar het gemiddelde $\bar{x}$ *niet* binnen dit betrouwbaarheidsinterval ligt, dan is de kans heel klein (kleiner dan 5%) dat deze uit een populatie getrokken is met het veronderstelde gemiddelde μ. In dat geval kunnen we de nulhypothese dus verwerpen.

[2] In R kan je dit berekenen met `qnorm(1 - 0.05)`

In Voorbeeld 5.1 is de kritieke grenswaarde het getal g waarvoor geldt dat

$$P(M > g) = \alpha$$

wat dan wordt verder uitgewerkt als:

$$P(Z > \frac{g-\mu}{\frac{\sigma}{\sqrt{n}}}) = \alpha$$

waaruit volgt dat:

$$g = \mu + z \times \frac{\sigma}{\sqrt{n}} \tag{5.3}$$

5.4 Overschrijdingskans

Een karakteristiek die gebruikt wordt om weer te geven hoe sterk de geobserveerde waarde afwijkt van H_0, is de overschrijdingskans (probability value of ook p-waarde). Dit vormt een alternatieve manier om te bepalen of de nulhypothese al dan niet verworpen kan worden.

Definitie 5.4.1 (overschrijdingskans)**.** *De* overschrijdingskans *is de kans, indien de nulhypothese waar is, om een waarde te verkrijgen van de toetsingsgrootheid die minstens even extreem is als de geobserveerde waarde.*

Definitie 5.4.2 (statistische significantie)**.** *In een statistische hypothesetoets heeft men een* statistisch significant *resultaat behaald wanneer de geobserveerde overschrijdingskans p van de teststatistiek lager is dan het significantieniveau α. De p-waarde wordt onder het gekozen significantieniveau beschouwd als te extreem om de veronderstelling dat de nulhypothese waar is aan te houden.*

Voorbeeld 5.2. *In het onderzoek naar het aantal dagelijkse reddingen door superhelden (Voorbeeld 5.1) kan de overschrijdingskans als volgt berekend worden:*

$$P(M > 3,483) = P\left(Z > \frac{3,483-3,3}{\frac{\sigma}{\sqrt{n}}}\right) = P(Z > 1,822) = 0,0344$$

■

Als de overschrijdingskans of p-waarde kleiner is dan de onbetrouwbaarheidsdrempel dan moet H_0 verworpen worden, is de p-waarde gelijk of groter dan α dan mag je H_0 niet verwerpen. In ons geval is de p-waarde $0,0344$ en die is kleiner dan $\alpha = 0,05$ dus moeten we H_0 verwerpen.

- p-waarde $< \alpha \Rightarrow H_0$, verwerpen want de gevonden waarde voor $\bar{x}$ is te extreem;
- p-waarde $\geq \alpha \Rightarrow H_0$ niet verwerpen, want de gevonden waarde voor $\bar{x}$ kan nog verklaard worden door toeval.

5.5 Eenzijdig of tweezijdig toetsen

In Voorbeeld 5.1 gaat het om een hypothese waar we vermoeden dat het populatiegemiddelde *hoger* ligt dan een bepaalde waarde. We twijfelen dus aan de de nulhypothese als ons steekproefgemiddelde significant boven het vooropgestelde gemiddelde $\mu = 3,3; \alpha = 0,05$ ligt. Het kritieke gebied om H_0 te verwerpen ligt dus aan de rechterzijde van de curve en we noemen deze toets dan ook rechtszijdig.

We zouden ook een toets kunnen maken waar we denken dat de superhelden gemiddeld *minder* mensen redden per dag. Dan ligt het kritieke gebied aan de linkerzijde en noemen we de toets linkszijdig.

Oefening 5.2. *Wat zou je in vergelijking 5.3 moeten veranderen opdat je de correcte kritieke waarde zou berekenen voor een linkszijdige z-toets?* ■

Soms kan het ook zijn dat er tweezijdig moet getoetst worden. De alternatieve hypothese wordt dan geformuleerd als zijnde dat het populatiegemiddelde verschillend is van de opgegeven waarde. Er moeten dan twee kritieke grenswaarden berekend worden namelijk de linker- en de rechter grenswaarden.

$$g = \mu \pm z \times \frac{\sigma}{\sqrt{n}} \tag{5.4}$$

De totale oppervlakte van het kritieke gebied moet $1 - \alpha$ zijn, en je moet er rekening mee houden dat zowel links als rechts een gebied met telkens oppervlakte $\alpha/2$ samen het aanvaardingsgebied vormen. Je moet dan ook de overeenkomstige z-waarde kiezen. Als we opnieuw significantieniveau $\alpha = 0,05$ nemen, zoeken we dus de z waarde waarvoor geldt dat;

$$P(Z < -z) + P(Z > z) = \alpha \Leftrightarrow 2P(Z > z) = \alpha \Leftrightarrow P(Z < z) = 1 - \frac{\alpha}{2} = 0,975$$

De overeenkomstige z-waarde is dan ongeveer 1.96 (op te zoeken in de z-tabel of in R met `qnorm(.975)`).

De drie vormen van de z-toets worden samengevat in tabel 5.1.

5.6 De z-toets in R

Het codevoorbeeld hieronder is de uitwerking van Voorbeeld 5.1 in R.

```
# De z-toets

# We hebben een steekproef met
n <- 30       # steekproefgrootte
sm <- 3.483   # steekproefgemiddelde
s <- 0.55     # standaardafwijking (verondersteld gekend)
```

Doel	Test op gemiddelde waarde μ van de populatie aan de hand van een steekproef van n onafhankelijke steekproefwaarden		
Voorwaarde	De populatie is willekeurig verdeeld, n voldoende groot		
Type test	Tweezijdig	Eenzijdig links	Eenzijdig rechts
H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
Verwerpingsgebied	$\lvert z \rvert > g$	$z < -g$	$z > g$
Teststatistiek		$z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	

Tabel 5.1: Samenvatting verschillende vormen van de z-toets

```r
a <- 0.05     # significantieniveau (gekozen door de onderzoeker)
m0 <- 3.3     # hypothetisch populatiegemiddelde (H0)

# Kunnen we vanuit deze steekproef besluiten dat mu > 3.3?
# H0: mu = 3.3     -> nulhypothese, willen we ontkrachten
# H1: mu > 3.3     -> alternatieve hypothese, willen we aantonen

#
# Methode 1. Overschrijdingskans
#
# Wat is de kans dat je in een steekproef het gegeven steekproefgemiddelde
# ziet? P(M > sm) in een verdeling M ~ Nor(m0, s/sqrt(n))
p <- 1 - pnorm(sm, m0, s/sqrt(n))  # => 0.03419546

# De gevonden kans is bijzonder klein, kleiner dan het significantieniveau
if(p < a) {
  print("H0 verwerpen")
} else {
  print("H0 niet verwerpen")
}

#
# Methode 2. Kritieke grensgebied
#
# Onder welke waarde kan je H0 niet verwerpen?
g <- m0 + qnorm(1-a) * s / sqrt(n)

# Als het gevonden steekproefgemiddelde onder g ligt, kan je H0 niet verwerpen
if (sm < g) {
  print("H0 niet verwerpen")
} else {
  print("H0 verwerpen")
```

```
}

#
# Plot van deze casus
#

# grenzen van de plot (x-waarden)
x <- seq(m0-4*s/sqrt(n), m0+4*s/sqrt(n), length=200)
# y-waarden (volgen de Gauss-curve)
dist <- dnorm(x, m0, s/sqrt(n))
plot (x, dist, type = 'l', xlab = '', ylab = '')

# Toon het gevonden steekproefgemiddelde ahv rode vertikale lijn
abline(v=sm, col='red')
text(sm, 2, sm)

# Het aanvaardingsgebied plotten
i <- x <= g                    # Waarden van x links van g
polygon(                       # Plot deze waarden op de grafiek
  c(x[i],     g,                          g),
  c(dist[i], dnorm(g, m0, s/sqrt(n)), 0),
  col = 'lightgreen')
text(g,.5,signif(g, digits=4)) # Toon grenswaarde

text(m0, 0.1, m0)              # Hypothetisch
   populatiegemiddelde
abline(v=m0)                   # Trek daar een vertikale lijn

text(m0, 1.5, 'aanvaardingsgebied (H0)')
```

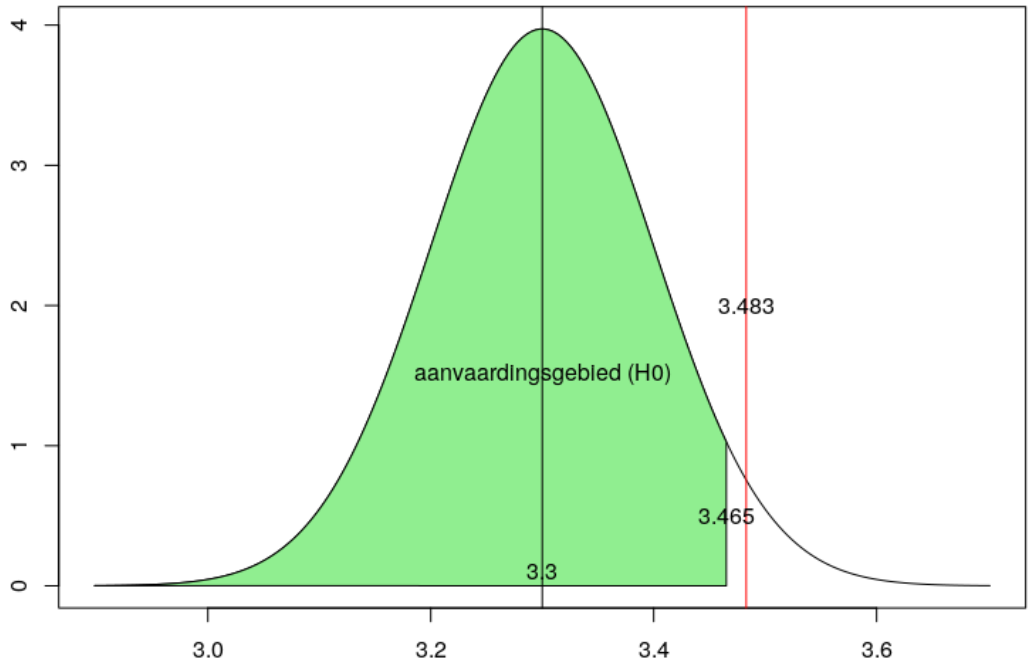
5.7 Voorbeelden

Voorbeeld 5.3. *Bij een aselecte steekproef van 50 waarnemingen vinden we we volgende grootheden:*

- $\bar{x} = 25$
- $s = \sqrt{55} = 7{,}41$

We willen weten of er reden is om aan te nemen dat μ van de populatie kleiner is dan 27.

1. *Bepalen van de hypothesen:*
   $H_0 : \mu = 27$ *en* $H_1 : \mu < 27$.
2. *Vastleggen significantieniveau α en steekproefomvang n:*
   $\alpha = 0{,}05$ *en* $n = 50$.
3. *Waarde toetsingsgrootheid bepalen.*
   We kiezen hiervoor het steekproefgemiddelde M. Volgens de centrale limietstelling geldt:



Figuur 5.2: Plot in R van de situatie van Voorbeeld 5.1

$$M \sim Nor(\mu = 27, \frac{\sigma}{\sqrt{n}})$$

De toetsingsgrootheid is

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{25 - 27}{\sqrt{\frac{55}{50}}} \approx -1,91$$

4. *Overschrijdingskans berekenen.*
   We vinden een overschrijdingskans van het gemiddelde van `pnorm(-1.91)` *of ongeveer* $0,0281$. *Bij een significantieniveau van 0,05 duidt dit er op dat we* H_0 *mogen verwerpen.*
5. *Bereken en teken kritiek gebied.*

$$g = \mu - z \times \frac{\sigma}{\sqrt{n}}$$

en dus

$$g = 27 - 1,645 \times \sqrt{\frac{55}{50}}$$

$$g = 25,27470944$$

We vinden dus dat $\bar{x} < g$ *komen tot hetzelfde besluit, nl. dat we* H_0 *kunnen verwerpen.*

■

Voorbeeld 5.4. *In een onderzoek naar het kleingeld dat in de zakken van van onze superhelden zit, stellen de onderzoekers dat zij gemiddeld 25 euro op zak hebben. Ze gaan ervan uit de spreiding* $\sigma = 7$ *is. Verder zijn de gegevens van de aselecte steekproef van omvang* $n = 64$ *beschikbaar met gemiddeld zakgeld* $\bar{x}$ *van 23 euro. Voor het significantieniveau kiezen ze* $\alpha = 0,05$.

1. *Bepalen van de hypothesen.*
   $H_0 : \mu = 25$ *en* $H_1 : \mu \neq 25$.
2. *Vastleggen significantieniveau* α *en steekproefomvang n.*
   $\alpha = 0,05$ *en* $n = 64$.
3. *Bepalen van de kritieke grenzen.*

$$g_1 = \mu - z \times \frac{\sigma}{\sqrt{n}} = 23,28$$

$$g_2 = \mu + z \times \frac{\sigma}{\sqrt{n}} = 26,72$$

4. *Kritiek gebied.*
   We vinden dat $\bar{x}$ *in het kritieke gebied ligt (want* $\bar{x} = 23 < g_1 = 23,28$*), dus mogen we* H_0 *verwerpen.*

■

5.8 De t-toets

Bij de z-toets gaan we uit van een aantal veronderstellingen waar we rekening moeten mee houden:

- De steekproef moet voldoende groot zijn ($n \geq 30$);
- De variatie van de toetsingsgrootheid moet normaal verdeeld zijn;
- We veronderstellen dat de standaardafwijking van de populatie, σ, gekend is.

De eerste drie voorwaarden maken dat de centrale limietstelling kan toegepast worden.

Soms zijn deze veronderstellingen niet geldig en mogen we dan ook de z-toets *niet* gebruiken! In deze gevallen kunnen we wel gebruik maken van de Student-t verdeling. In de t-toets wordt er wel van uit gegaan dat de onderzochte variabele normaal verdeeld is.

De formule voor de kritieke grenswaarde wordt dan aangepast als:

$$g = \mu \pm t \times \frac{s}{\sqrt{n}} \tag{5.5}$$

Voor het bepalen van de t-waarde hebben we het aantal vrijheidsgraden nodig, $n-1$. Om de standaardafwijking te schatten, gebruiken we de steekproefstandaardafwijking, s.

Voorbeeld 5.5. *Stel dat de onderzoekers van de superhelden uit Voorbeeld 5.1 door tijdsdruk niet in staat waren om een voldoende grote steekproef te nemen en slechts $n = 25$ observaties gedaan hebben, met hetzelfde steekproefgemiddelde $\bar{x} = 3{,}483$. De standaardafwijking in deze steekproef bleek $s = 0{,}55$.*

Kunnen we in deze omstandigheden, met eenzelfde significantieniveau $\alpha = 0{,}05$, het besluit dat superhelden dagelijks meer *dan 3,3 mensen redden aanhouden?*

1. *Bepalen van de hypothesen.*
   $H_0 : \mu = 3{,}3$ *en* $H_1 : \mu > 3{,}3$.
2. *Vastleggen significantieniveau α en steekproefomvang n.*
   $\alpha = 0{,}05$ *en* $n = 25$.
3. *Bepalen van de kritieke grenswaarde.*

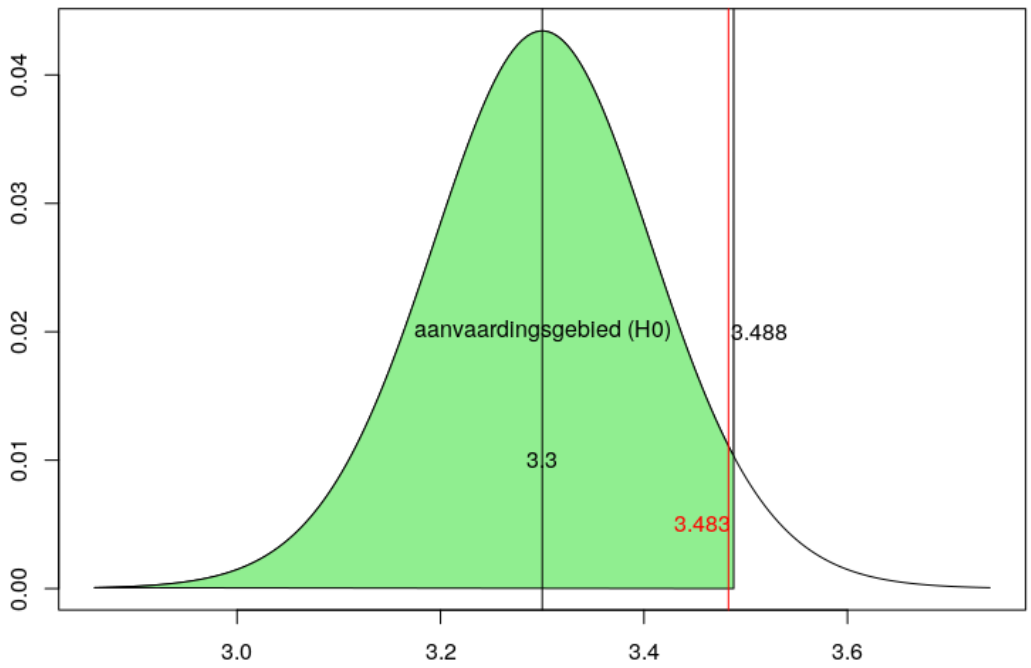
   $$g_2 = \mu + t \times \frac{s}{\sqrt{n}} \approx 3{,}3 + 1{,}711 \times \frac{0{,}55}{\sqrt{25}} \approx 3.488$$

   De waarde voor t wordt in R berekend met `qt(1-a, df = n - 1)` *(met* `a` *het significantieniveau en* `df` *het aantal vrijheidsgraden.)*
4. *Conclusie.*
   We vinden dat $\bar{x} = 3{,}483$ kleiner is dan de kritieke grenswaarde en dus in het aanvaardingsgebied ligt. Met andere woorden, we mogen H_0 niet *verwerpen.*

Met andere woorden, ook al krijgen we gelijkaardige resultaten in onze steekproef, kunnen we niet hetzelfde besluit trekken. Omdat onze steekproef te klein is, is er grotere onzekerheid of de waarde van het steekproefgemiddelde extreem genoeg is om de nulhypothese te verwerpen.

Hieronder vind je de uitwerking van dit voorbeeld in R. ∎

```r
# De t-toets

# We hebben een steekproef met
n <- 25       # steekproefgrootte
sm <- 3.483   # steekproefgemiddelde
ss <- 0.55   # standaardafwijking van de steekproef
a <- 0.05     # significantieniveau (gekozen door de onderzoeker)
m0 <- 3.3     # hypothetisch populatiegemiddelde (H0)

# Kunnen we vanuit deze steekproef besluiten dat mu > 3.3?
# H0: mu = 3.3     -> nulhypothese, willen we ontkrachten
# H1: mu > 3.3     -> alternatieve hypothese, willen we aantonen

#
# Methode 1. Kritieke grensgebied
#
# Onder welke waarde kan je H0 niet verwerpen?
g <- m0 + qt(1-a, df = n-1) * ss / sqrt(n)

# Als het gevonden steekproefgemiddelde onder g ligt, kan je H0
    niet verwerpen
if (sm < g) {
  print("H0 niet verwerpen")
} else {
  print("H0 verwerpen")
}

#
# Methode 2. Overschrijdingskans
#
# Wat is de kans dat je in een steekproef het gegeven
    steekproefgemiddelde
# ziet? P(M > sm) in een verdeling M ~ T(m0, ss/sqrt(n), df=n-1)
p <- 1 - pt((sm - m0) / (ss/sqrt(n)), df = n-1)

# De gevonden kans is bijzonder klein, kleiner dan het
    significantieniveau
if(p < a) {
  print("H0 verwerpen")
} else {
  print("H0 niet verwerpen")
}

#
# Plot van deze casus
#

# grenzen van de plot (x-waarden)
x <- seq(m0-4*ss/sqrt(n), m0+4*ss/sqrt(n), length=200)
```



Figuur 5.3: Plot in R van de situatie van Voorbeeld 5.5

```
# y-waarden (volgen de Gauss-curve van de t-verdeling)
dist <- dt((x-m0)/(ss/sqrt(n)), df = n-1) * ss/sqrt(n)
plot (x, dist, type = 'l', xlab = '', ylab = '')

# Het aanvaardingsgebied plotten
i <- x <= g                    # Waarden van x links van g
polygon(                       # Plot deze waarden op de grafiek
  c(x[i],     g,                                  g),
  c(dist[i], dt((g-m0)/(ss/sqrt(n)),df=n-1), 0),
  col = 'lightgreen')

text(m0, 0.01, m0) # Hypothetisch populatiegemiddelde
abline(v=m0)       # Trek daar een vertikale lijn

text(g+.025,.02,signif(g, digits=4)) # Toon grenswaarde

# Toon het gevonden steekproefgemiddelde ahv rode vertikale lijn
abline(v=sm, col='red')
text(sm-.025, .005, sm, col = 'red')

text(m0, 0.02, 'aanvaardingsgebied (H0)')
```

Voorbeeld 5.6. *Een uitbraak van een door Salmonella veroorzaakte ziekte werd toegeschreven*

*aan vanille-ijs van een bepaalde fabriek (***Lindquist***). Wetenschappers hebben het niveau van Salmonella gemeten in 9 willekeurig genomen steekproeven.*

De niveaus (in MPN/g[3]*) zijn de volgende:*

0,593	*0,142*	*0,329*	*0,691*	*0,231*
0,793	*0,519*	*0,392*	*0,418*	

Is er reden om aan te nemen dat het Salmonella-niveau in het ijs significant groter is dan 0,3 MPN/g? We zullen gebruik maken van de R-functie `t.test` *om deze vraag te beantwoorden. Lees zelf de help-pagina van deze functie om de mogelijke opties te leren kennen.*

1. *Bepalen van de hypothesen*
   $H_0 : \mu = 0.3, H_1 : \mu > 0,3$
2. *Vastleggen significantieniveau* $\alpha = 0.05$ *(in R moet je het betrouwbaarheidsniveau* $1 - \alpha$ *opgeven, dus 0,95) en steekproefomvang* $n = 9$
3. *Bepalen overschrijdingskans. Het gaat hier over een rechtszijdige toets, wat aangegeven wordt met de optie* `alternative="greater"`*. Het gekozen betrouwbaarheidsniveau is de standaardwaarde voor deze functie en moet niet expliciet meegegeven worden.*

```
x <- c(0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519,
    0.392, 0.418)
t.test(x, alternative = "greater", mu = 0.3)
```

Het resultaat is:

```
One Sample t-test

data:  x
t = 2.2051, df = 8, p-value = 0.02927
alternative hypothesis: true mean is greater than 0.3
95 percent confidence interval:
0.3245133       Inf
sample estimates:
mean of x
0.4564444
```

4. *Conclusie. De overschrijdingskans* $p = 0,029 < \alpha = 0,05$*. We kunnen dus de nulhypothese verwerpen; er is met ander worden en vrij sterke aanwijzing dat het gemiddelde Salmonella-niveau in het ijs groter is dan 0,3 MPN/g.*

Opmerking*: "*`confidence interval:  0.3245133 Inf`*" heeft niet rechtstreeks met het aanvaardingsgebied of het kritieke gebied te maken. Het staat ook los van de waarde* $\mu_0 = 0,3$*. Het zegt enkel dat vermits* $\bar{x} = 0,45644$*, dat we met 95% procent zekerheid kunnen zeggen dat het échte gemiddelde (*μ*) van de populatie tussen* $0,3245133$ *en* $+\infty$ *ligt. Zie paragraaf 4.5.3 (p. 58) en 4.5.4 (p. 60) over betrouwbaarheidsintervallen.* ∎

[3] Most Probable Number. Zie bv. http://www.microbiologie.info/mpn-methode.html voor meer uitleg over deze methode.

5.9 De t-toets voor twee steekproeven

De t-toets kan ook gebruikt worden om twee steekproeven met elkaar te vergelijken. Je kan er dan mee nagaan of het steekproefgemiddelde van beide steekproeven *significant* verschillend is.

Men maakt onderscheid tussen twee gevallen:

- Beide steekproeven zijn onafhankelijk, zijn afzonderlijk genomen. Een voorbeeld is een onderzoek naar een medische behandelingsmethode waar een contolegroep de behandeling *niet* krijgt en een testgroep de behandeling wel krijgt.
- De steekproeven zijn afhankelijk, of gepaard. Een voorbeeld is twee metingen uitvoeren op hetzelfde lid van de populatie, zoals de koorts nemen voor en na het innemen van een medicijn om het effect ervan te meten.

In R kan je eveneens de functie `t.test` gebruiken voor het uitvoeren van een toets met twee steekproeven. We geven hieronder twee voorbeelden, één voor elk geval.

Voorbeeld 5.7. *In een klinisch onderzoek wil men nagaan of een nieuw medicijn als bijwerking een verminderde reactiesnelheid heeft (***Lindquist***).*

Zes deelnemers kregen een medicijn toegekend (interventiegroep) en zes anderen een placebo (controlegroep). Vervolgens werd hun reactietijd op een stimulus gemeten (in ms). We willen nagaan of er significante verschillen zijn tussen de interventie- en controlegroep.

***Opmerking**: De interventiegroep en de controle groep zijn hier toevallig even groot (elk 6 proefpersonen). Dit is niet noodzakelijk. Bij een onafhankelijke (niet-gepaarde) steekproeven mogen de 2 groepen een verschillende grootte hebben.*

- *Controlegroep: 91, 87, 99, 77, 88, 91* $(\bar{x} = 88,83)$
- *Interventiegroep: 101, 110, 103, 93, 99, 104* $(\bar{y} = 101,67)$

We noteren μ_1 voor het gemiddelde van de niet behandelde populatie (controlegroep) en μ_2 voor het populatiegemiddelde van de patiënten die het medicijn nemen (interventiegroep).

De hypothesen worden formeel als volgt genoteerd:

$H_0 : \mu_1 - \mu_2 = 0$ *en* $H_1 : \mu_1 - \mu_2 < 0$

Als teststatistiek (steekproefgrootheid) gebruiken we $\bar{x} - \bar{y}$, met $\bar{x}$ en $\bar{y}$ schattingen voor de échte waarden μ_1 en μ_2 .

Het gaat hier dus over een linkszijdige test, wat weergegeven wordt door de optie `alternative = "less"`. In de nulhypothese veronderstellen we dat het verschil tussen de populatiegemiddelden 0 is, wat met de optie `mu = 0` wordt aangeduid. Merk op dat dit de standaardwaarde is voor deze parameter en dus in principe niet moet worden opgegeven.

```
controle <- c(91, 87, 99, 77, 88, 91)
interventie <- c(101, 110, 103, 93, 99, 104)
t.test(controle, interventie, alternative="less", mu=0)
```

Het resultaat van de toets:

Auto	1	2	3	4	5	6	7	8	9	10
Gewone benzine	16	20	21	22	23	22	27	25	27	28
Additieven	19	22	24	24	25	25	25	26	28	32

Tabel 5.2: Verbruik in mijl per gallon met 2 soorten benzine.

```
t.test(controle, interventie, alternative="less")

Welch Two Sample t-test

data:  controle and interventie
t = -3.4456, df = 9.4797, p-value = 0.003391
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -6.044949
sample estimates:
mean of x mean of y
88.83333 101.66667
```

De teststatistiek $\bar{x}-\bar{y}=-12,833$ komt overeen met een t-waarde $t=-3,4456$. De parameter $df=9,48$ wordt bepaald door `t.test()` op basis van het aantal elementen in de reeksen x en y. De berekening hiervan is niet triviaal.

De p-waarde, 0,003391, ligt duidelijk onder het significantieniveau (niet expliciet opgegeven, dus werd de standaardwaarde $\alpha=0,05$ gebruikt.)

We mogen dus de nulhypothese verwerpen en besluiten dat volgens de resultaten van deze steekproef het medicijn inderdaad een significant effect heeft op de reactiesnelheid van patiënten.

***Opmerking**: Vermits $\bar{x}-\bar{y}=-12,833$ kunnen we met 95% procent zekerheid zeggen dat het verschil van de échte gemiddelden ($\mu_1-\mu_2$) van een grotere control- en interventiepopulatie tussen $-\infty$ en -6.044949 zal liggen. Zie paragraaf 4.5.3 (p. 58) en 4.5.4 (p. 60) over betrouwbaarheidsintervallen.* ∎

Voorbeeld 5.8. *In een studie werd nagegaan of auto's die rijden op benzine met additieven ook een lager verbruik hebben. Tien auto's werden eerst volgetankt met ofwel gewone benzine, ofwel benzine met additieven (bepaald door opgooien van een munt), waarna het verbruik werd gemeten (uitgedrukt in mijl per gallon). Vervolgens werden de auto's opnieuw volgetankt met de andere soort benzine en werd opnieuw het verbruik gemeten. De resultaten worden gegeven in Tabel 5.2.*

We gaan door middel van een gepaarde *t*-test *na of auto's significant zuiniger rijden met benzine met additieven.*

We kiezen x voor benzine met `additieven` ($\bar{x}=25,1$ mijl per gallon), en we kiezen y voor `gewone` bezine ($\bar{y}=23,1$ mijl per gallon).

De nulhypothese H_0 is dat je met beiden even veel mijl per gallon kunt rijden ($\mu_{x-y}=0$). De alternatieve hypothese H_1 dat je verder kunt rijden op benzine met additieven ($\mu_{x-y}>0$).

De optie `paired=TRUE` geeft aan dat het hier om een gepaarde t-toets gaat.

```
gewone      <- c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28)
additieven <-c(19, 22, 24, 24, 25, 25, 26, 26, 28, 32)
t.test(additieven, gewone, alternative="greater", paired=TRUE)
```

Resultaat:

```
 Paired t-test

data:  additieven and gewone
t = 4.4721, df = 9, p-value = 0.0007749
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.180207      Inf
sample estimates:
mean of the differences
                      2
```

De teststatistiek $\overline{x-y}=2$. *Dit komt overeen met een t-waarde* $t=4{,}4721$. *De p-waarde, 0,0007749, ligt onder het significantieniveau* ($\alpha=0{,}05$), *dus we kunnen de nulhypothese verwerpen. Volgens deze steekproef rijden auto's inderdaad zuiniger met benzine met additieven.*

***Ter info**: Bij 95% van de "paren" van een grotere populatie auto's, zal het verschil* $x-y$ *tussen* 1.180207 *en* $+\infty$ *liggen. Dit is het betrouwbaarheidsinterval waarvan sprake in paragraaf 4.5.3 (p. 58) en 4.5.4 (p. 60).* ∎

5.10 Fouten in hypothesetoetsen

Bij het uitvoeren van een hypothesetoets kunnen altijd nog fouten optreden. Indien we H_0 verwerpen wanneer ze in werkelijkheid juist is, spreken we van een fout van type I en wanneer we H_0 ten onrechte aanvaarden van een fout van type II.

Het significatieniveau α bepaalt bij het uitvoeren van een hypothesetoets wanneer de nulhypothese precies verworpen kan worden. Stel dat we een significatieniveau van 5% kiezen. Als de nulhypothese waar is, dan is de kans dat we een steekproef trekken met een toetsingswaarde die in het verwerpingsgebied terecht komt 5%. M.a.w. de kans om de nulhypothese te verwerpen terwijl ze waar is, is 5 % of in het algemeen: het significantieniveau van een toets is gelijk aan de kans op het maken van een fout van type I.

Het is vanzelfsprekend dat we de kans op een fout van type I zo klein mogelijk willen houden. Jammer genoeg is dit ten koste van de kans op een type II fout, aangeduid met β, die hierdoor groter wordt. Het verband tussen α en β is niet triviaal en we gaan hier in deze cursus niet verder op in.

In vele gevallen is het maken van een fout van type I erger dan een van type II. Denk maar aan een rechtszaak waarbij de nulhypothese is dat de persoon onschuldig is. Indien we toetsen op een 5% significantieniveau is de kans op een type I fout 5 op 100. M.a.w. er is een betrouwbaarheid van 95% dat de juiste beslissing wordt genomen indien H_0 correct is. Daarom vermijden we liever de

Conclusies	Werkelijke stand van zaken	
	H_0 **correct**	H_1 **correct**
H_0 **geaccepteerd**	Juist	Fout van type II
H_0 **verworpen**	Fout van type I	Juist

Tabel 5.3: Conclusies en consequenties bij toetsen van een hypothese; types van fouten.

conclusie dat H_0 geaccepteerd wordt, maar eerder dat de steekproef onvoldoende bewijs bevat om H_0 bij een bepaald significantieniveau te verwerpen.

5.11 Oefeningen

Oefening 5.3. *Betrouwbaarheidsintervallen.*

1. *Wat is de onder- en bovengrens van een betrouwbaarheidsinterval van 99%?*
2. *Een betrouwbaarheidsinterval van 99% is breder dan een van 95%. Waarom is dit zo?*
3. *Hoe zou het betrouwbaarheidsinterval voor 100% er uit zien?*

■

Oefening 5.4. *Er wordt gezegd dat het invoeren van een bindend studieadvies (BSA) een rendementsverhoging tot gevolg heeft in slaagkans. Voor het invoeren van het BSA was in de studentenpopulatie het gemiddelde aantal behaalde studiepunten per jaar per student gelijk aan 44 met een standaardafwijking van 6,2. Na invoering van het BSA wijst een onderzoek uit onder 72 studenten dat deze een gemiddeld aantal studiepunten haalden van 46,2.*

1. *Toets of er bewijs is dat het invoeren van een BSA leidt tot een rendementsverhoging. Gebruik methode van kritieke grenswaarde. ($\sigma = 6,2, \alpha = 2,5\%$).*
2. *Toon hetzelfde aan met de methode van de overschrijdingskans.*
3. *Geef een interpretatie wat de betekenis is van $\alpha = 2,5\%$.*

■

Oefening 5.5. *Eén van de motieven voor het kiezen van een garage is de inruilprijs voor de oude auto. De importeur van Ford wil graag dat de verschillende dealers een gelijk prijsbeleid voeren. De importeur vindt dat het gemiddelde prijsverschil tussen de dichtstbijzijnde Ford-dealer en de dealer waar men de auto gekocht heeft hoogstens €300 mag bedragen. De veronderstelling is dat als het verschil groter is, potentiële klanten eerder geneigd zullen zijn om bij hun vorige dealer te blijven.*

In een steekproef worden volgende verschillen genoteerd:

400	*350*	*400*	*500*	*300*	*350*	*200*
500	*200*	*250*	*250*	*500*	*350*	*100*

Toets of er reden is om aan te nemen dat het gemiddelde prijsverschil in werkelijkheid significant groter is dan €300. Gebruik een significantieniveau van 5%.

■

Oefening 5.6. *In Oefening 3.9 en volgende hebben we de resultaten van performantiemetingen voor persistentiemogelijkheden in Android geanalyseerd (***Akin2016***). Er werden experimenten uitgevoerd voor verschillende combinaties van hoeveelheid data (klein, gemiddeld, groot) en persistentietype (GreenDAO, Realm, SharedPreferences, SQLite). Voor elke hoeveelheid data hebben we kunnen bepalen welk persistentietype het beste resultaat gaf.*

Nu gaan we uitzoeken of het op het eerste zicht beste persistentietype ook significant *beter is dan de concurrentie.*

Concreet: ga aan de hand van een toets voor twee steekproeven voor elke datahoeveelheid na of het gemiddelde van het best scorende persistentietype significant lager *is dan het gemiddelde van (i) het* tweede *beste en (ii) het slechtst scorende type .*

Kunnen we de conclusie aanhouden dat voor een gegeven datahoeveelheid één persistentietype het beste is, d.w.z. significant beter is dan gelijk welk ander persistentietype? ■

Oefening 5.7. *Een groot aantal studenten heeft deelgenomen aan een test die in verschillende opeenvolgende sessies werd georganiseerd. Omdat het opstellen van een aparte opgave voor elke sessie praktisch onhaalbaar was, is telkens dezelfde opgave gebruikt. Eigenlijk bestaat er dus het gevaar dat studenten na hun sessie info konden doorspelen aan de groepen die nog moesten komen. De latere groepen hebben dan een voordeel ten opzichte van de eerste. Blijkt dit ook uit de cijfers?*

Het bestand `puntenlijst.csv` *bevat alle resultaten van de test. Elke groep wordt aangeduid met een letter, in de volgorde van de sessie.*

- *Dag 1: sessies A, B*
- *Dag 2: sessies C, D, E*
- *Dag 3: sessies F, G, H*

Sessies A en B zijn doorgegaan op een andere campus, dus er zou kunnen verondersteld worden dat er weinig tot geen communicatie is met de studenten van de andere sessies.

Als er info met succes doorgespeeld werd, dan verwachten we dat de scores van de groepen die later komen significant beter zijn dan de eerste.

Merk op dat de omgekeerde redenering niet noodzakelijk geldt: als blijkt dat het resultaat van de latere sessies inderdaad significant beter blijkt, dan betekent dat niet noodzakelijk dat de oorzaak (enkel) het doorspelen van informatie is. Er kunnen ook andere oorzaken zijn (bv. "zwakkere" klasgroepen zijn toevallig eerder geroosterd).

1. *Ga op verkenning in de data. Bereken de gepaste centrum- en spreidingsmaten voor de dataset als geheel en voor elke sessie afzonderlijk.*
2. *Maak een staafgrafiek van de gemiddelde score per sessie. Is dit voldoende om een beeld*

te vormen van de resultaten? Waarom (niet)?

3. *Maak een boxplot van de scores opgedeeld per groep. Vergelijk onderling de hieronder opgesomde sessies. Denk je dat er een significant verschil is tussen de resultaten? Wordt ons vermoeden dat er informatie doorgespeeld wordt bevestigd?*
   - *A en B*
   - *C, D en E*
   - *F, G en H*
   - *C en H*
   - *A en H*
4. *Ga door middel van een geschikte statistische toets voor na of de verschillen tussen die hierboven opgesomde groepen ook* significant *is. Kunnen we concluderen dat de latere groepen beter scoren of niet?*

■

5.12 Antwoorden op geselecteerde oefeningen

Oefening 5.2:

$$g = \mu - z \times \frac{\sigma}{\sqrt{n}} \tag{5.6}$$

want

$$P(M < g) = P\left(Z < \frac{g-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = 0,05$$

Wegens de symmetrieregel kunnen we zeggen

$$P\left(Z > -\left(\frac{g-\mu}{\frac{\sigma}{\sqrt{n}}}\right)\right) = 0,05$$

De z-waarde die ermee overeen komt is 1,645 dus hebben we

$$z = \frac{-g+\mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\Leftrightarrow -g = \frac{\sigma}{\sqrt{n}}z - \mu$$

$$\Leftrightarrow g = -\frac{\sigma}{\sqrt{n}}z + \mu$$

Oefening 5.4

1. $g \approx 45,4 < \bar{x} = 46,2$.
   $\bar{x}$ ligt in het kritieke gebied, dus we mogen de nulhypothese verwerpen. We hebben dus redenen om aan te nemen dat bindend studieadvies inderdaad het studierendement significant verhoogt.

Grootte	Beste	2e beste	p-waarde
Small	Realm	SharedPreferences	0.1699
Medium	Realm	GreenDAO	0.0002506
Large	Realm	SQLite	0.0017

Tabel 5.4: Resultaten t-toets voor de beste en 2e beste persistentietype op basis van steekproefgemiddelde (**Akin2016**).

2. $P(M > 46.2) \approx 0,0013 < \alpha = 0,025$. De overschrijdingskans is kleiner dan het significantieniveau, dus we mogen de nulhypothese verwerpen.
3. α is de kans dat je H_0 ten onrechte verwerpt. Er is m.a.w. een kans van 2,5% dat je ten onrechte de conclusie trekt dat het studierendement hoger is geworden.

Oefening 5.5

In deze situatie ($n = 14 < 30$) mogen we geen z-toets gebruiken, maar vallen we terug op de t-toets.

- $\overline{x} \approx 332,143$
- $s \approx 123,424$
- $g \approx 358,42$. Het steekproefgemiddelde ligt niet in het kritieke gebied, dus we kunnen H_0 *niet* verwerpen.
- $p \approx 0,1738$. $p \not< \alpha$, dus we kunnen H_0 *niet* verwerpen.

Er is op basis van deze steekproef dus geen reden om aan te nemen dat het gemiddelde prijsverschil op de inruilprijs van oude wagens significant groter is dan door de importeur aanbevolen.

Oefening 5.6

Tabel 5.4 geeft een overzicht met voor elke datasetgrootte het beste en tweede beste persistentietype (op basis van het steekproefgemiddelde). De conclusie van **Akin2016**, dat *Realm* het performantste persistentietype is, blijft overeind, maar voor de kleine datasets is het verschil niet significant.

Merk op dat we hier niet expliciet vooraf een significantieniveau gekozen hebben. Voor $\alpha = 0,1$, $0,05$ of zelfs $0,01$, kunnen we echter dezelfde conclusie trekken.

6. Analyse op 2 variabelen

In de vorige hoofdstukken hebben we telkens één variabele tegelijkertijd onderzocht. Vaak hebben onderzoeksvragen echter te maken met *verbanden* (en dan vooral oorzakelijke) tussen variabelen. In dit hoofstuk gaan we hier verder op in.

Wanneer we een verband beschrijven tussen variabelen, onderscheiden we:

- De *afhankelijke variabele*, waarover we een voorspelling willen doen;
- De *onafhankelijke variabele*, op basis van dewelke we de voorspelling doen.

Als de onafhankelijke variabele op een bepaalde manier verandert, verwachten we dat de waarde van de afhankelijke variabele op een voorspelbare manier mee verandert.

Voorbeeld 6.1. *Een voorbeeld waarbij verbanden kunnen gevonden tussen variabelen vind je bijvoorbeeld bij Ant Colony Optimization (ACO). Dit is een techniek die gebruikt wordt in verschillende computationele problemen. Men baseert zich hier op hoe mieren voedsel zoeken en vinden en dat communiceren aan de groep. Mieren verspreiden feromonen als ze op pad gaan op zoek naar eten. Hoe langer het pad, hoe minder feromonen het pad zal bevatten, hoe korter het pad, hoe groter de kans dat er een grote concentratie aan feromonen te vinden is. Mieren worden aangetrokken door deze feromonen en zullen dus proberen de meest bewandelde paden te gebruiken om naar een bepaalde voedselbron te gaan. Nu kan je onderzoeken of de tijd voor het vinden van een pad, afhangt van een aantal variabelen:*

- *De mate waarin feromonen verspreid worden*
- *De mate waarin een feromoon verdwijnt*
- *Het aantal obstakels tussen het nest en de voedselbron*
- *De vorm van de obstakels tussen nest en voedselbron (vinden ze sneller het pad indien er geen hoeken aan de obstakels zijn bv.)*

■

Om een vergelijking te maken kunnen we (i) de bekende statistieken zoals gemiddelde e.a. berekenen en analyseren of (ii) grafische voorstellingen maken van deze statistieken.

Welke soort van grafieken we kunnen gebruiken hangt af van het meetniveau: @

- Interval of ratio:
  - Staafdiagram van de gemiddelden
  - Boxplot per groep
- Ordinaal of nominaal
  - Kruistabel
  - Geclusterd staafdiagram
  - Rependiagram

Bij de vraag of er samenhang is tussen twee variabelen kunnen we volgende grafieken/statistieken gebruiken:

- Nominaal x Nominaal:
  - Kruistabel met Cramér's V
- Ordinaal x Ordinaal
  - Geclusterd staafdiagram
  - Rependiagram
- Ratio x Ratio
  - Spreidingsdiagram
  - Regressie en correlatie met correlatiecoëfficiënt.

6.1 Kruistabellen en Cramér's V

Definitie 6.1.1 (Kruistabel)**.** *In een kruistabel (zie bv. Figuur 6.1) worden de frequenties van twee variabelen samengevat.*

Elke cel van de laatste kolom bevat de som van de overeenkomstige rij en elke cel van de laatste rij bevat de som van de overeenkomstige kolom. Dit worden de marginale totalen *genoemd.*

In R kan je een kruistabel (Eng.: *contingency table* of *cross table*) opstellen met de functie `table`. Een voorbeeld i.v.m. de oefening over Android persistentietypes (zie Oefening 3.9):

```
> table(DataHoeveelheid, PeristentieType)
             PersistentieType
DataHoeveelheid GreenDAO Realm Sharedpreferences SQLLite
Large                30    30                 0       30
Medium               30    30                 0       30
Small                30    30                30       30
```

In een gewone kruistabel kunnen we geen directe conclusies trekken, aangezien het analyseren of er samenhang bestaat tussen variabelen niet goed gaat op basis van de celfrequenties. Niet alle metingen zijn even groot! Daarom moeten we percenteren. Nog even snel de regel van percenteren:

- Om te weten hoeveel percent x is van y, deel je x door y en vermenigvuldig je met 100: $perc = 100 \times \frac{x}{y}$.
- Om te weten hoeveel x % is van y : $\frac{x \times y}{100}$

	Vrouw	Man	Totaal
Goed	9	8	17
Voldoende	8	10	18
Onvoldoende	5	5	10
Slecht	0	4	4
Totaal	22	27	49

Tabel 6.1: Een kruistabel voor de waardering door mannen en vrouwen van een bepaald assortiment producten.

	Vrouw	Man	Totaal	Vrouw %	Man%	Totaal
Goed	9	8	17	41%	30%	35%
Voldoende	8	10	18	36%	37%	37%
Onvoldoende	5	5	10	23%	18%	20%
Slecht	0	4	4	0%	15%	8%
Totaal	22	27	49	100%	100%	100%

Tabel 6.2: De kruistabel waarbij we de waarden gepercenteerd hebben.

Voorbeeld 6.2. *In Tabel 6.1 vinden we de data waar er gekeken wordt naar het verschil in waardering van een assortiment tussen mannen en vrouwen. We percenteren per geslacht en vinden bijvoorbeeld dat 41% van de vrouwen een waardering goed heeft (zie Tabel 6.2). Nu kunnen we ons de vraag stellen of de waarderingskeuze afhangt van het geslacht van de persoon.* ∎

In ons voorbeeld kunnen we besluiten dat 30% van de mannen tevreden is en 15% van de mannen ontevreden. Maar hoe goed is die samenhang tussen de verschillende variabelen (geslacht en tevredenheid)? Dat kunnen we bepalen aan de hand van Cramér's V. Voordat we die definitie kunnen geven, moeten we echter eerst de waarde χ^2 introduceren.

6.2 χ^2 test voor associatie

De χ^2 (*chi-kwadraat*) waarde is een grootheid die gebruikt wordt om te bepalen of er een significant verband bestaat tussen twee variabelen. Meer hierover volgt later in Hoofdstuk 7. De berekening ervan leggen we alvast hier uit.

1. Stel de kruistabel op samen met marginale totalen (zie tabel 6.2).
2. Stel voor elke cel een schatter op voor de theoretische kans om in die cel te geraken. Deze schatter kan je bereken als volgt: (kans op in de rij van deze cel te komen) $\times$ (kans om in de kolom van de cel te komen). In het voorbeeld is dit dus voor $\text{cel}_{1,2}$:

$$P[rij_1] \times P[kolom_2] = \frac{17}{49} \times \frac{27}{49} = 0.191170346$$

Algemeen kan je dus stellen dat de verwachte theoretische waarde e als volgt kan berekend worden:

$$e = (\frac{rijtotaal}{n} \times \frac{kolomtotaal}{n}) \times n = \frac{rijtotaal \times kolomtotaal}{n} \tag{6.1}$$

	Vrouw	Man	Totaal	Vrouw %	Man%	Totaal
Goed	$9-7.63$	$8-9.36$	17	41%	30%	35%
Voldoende	$8-8.08$	$10-9.91$	18	36%	37%	37%
Onvoldoende	$5-4.48$	$5-5.51$	10	23%	18%	20%
Slecht	$0-1.79$	$4-2.20$	4	0%	15%	8%
Totaal	22	27	49	100%	100%	100%

Tabel 6.3: De kruistabel waarbij we de schatter e (hetgeen we verwachten bij geen samenhang) bepaald hebben voor elke cel en die aftrekken van de geobserveerde waarde.

	Vrouw	Man	Totaal	Vrouw %	Man%	Totaal
Goed	0.2	0.2	17	41%	30%	35%
Voldoende	0	0	18	36%	37%	37%
Onvoldoende	0.1	0	10	23%	18%	20%
Slecht	1.8	1.5	4	0%	15%	8%
Totaal	22	27	49	100%	100%	100%

Tabel 6.4: De kruistabel waarbij we het verschil gekwadrateerd en genormeerd hebben.

Met:

- e verwachte waarde bij onafhankelijkheid
- $rijtotaal$ totaal van de rij van de betreffende cel
- $kolomtotaal$ totaal van de kolom van de betreffende cel

Voor $\text{cel}_{1,2}$ is dit dus 9.36.

3. Dan bereken je het verschil tussen geobserveerde (notatie a) en verwachte frequentie (e). (Zie tabel 6.3)
4. De laatste stap houdt in dat we een berekening gaan maken voor de maat van afwijking voor elke cel. Opnieuw gaan we hier een kwadraat nemen om het teken kwijt te spelen. We gaan ook de afwijking delen door de verwachte theoretische waarde om hen relatief even belangrijk te maken. Bijvoorbeeld: een afwijking van 5 op een verwachte frequentie van 20 is groter dan bv. een afwijking op een verwachte waarde van 200. Dit geeft dan volgende berekening (zie Tabel 6.4):

$$\frac{(a-e)^2}{e} \tag{6.2}$$

5. Deze gekwadrateerde deviaties gaan we dan optellen en vormt de χ^2 [1]

$$\chi^2 = \sum \frac{(a-e)^2}{e} \tag{6.3}$$

Met deze statistiek kunnen de waarde Cramér's V berekenen:

[1] Let op dat er afgerond wordt.

$V = 0$	geen samenhang
$V \approx 0,1$	zwakke samenhang
$V \approx 0,25$	redelijk sterke samenhang
$V \approx 0,50$	sterke samenhang
$V \approx 0,75$	zeer sterke samenhang
$V = 1$	volledige samenhang

Tabel 6.5: Interpretatie van de waarde van Cramérs'V

Definitie 6.2.1 (Cramér's V)**.**

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} \tag{6.4}$$

met

χ^2 *de berekende chi-kwadraatwaarde.*

- *n het aantal waarnemingen.*
- *k = de kleinste waarde van het aantal kolommen of het aantal rijen van de tabel.*

Cramér's V is de χ^2, gecorrigeerd voor steekproefomvang en het aantal categorieën in de variabelen. Het resultaat is altijd een getal tussen 0 en 1. Tabel 6.5 geeft aan hoe je het resultaat kan interpreteren.

Voor ons voorbeeld waarbij gekeken wordt naar de samenhang tussen geslacht en waardering van het assortiment vinden we een $\chi^2 = 3.811$ en dus een Cramér's V van 0.279 (want $n = 49$ en $k = 2$), wat duidt op redelijk sterke samenhang tussen de variabelen. Met andere woorden, de resultaten van de bevraging geven aan dat er een verschil is in de waardering die vrouwen en mannen geven over het assortiment.

Hieronder vind je de uitwerking van Voorbeeld 6.2 in R.

```
# Voorbeeld kruistabellen: waardering v/e product tussen
# vrouwen en mannen. Vergelijk de uitkomsten van elke stap
# met het voorbeeld in de cursus!
#
# Bron: http://www.cyclismo.org/tutorial/R/tables.html#creating-a-table-directly

# Kruistabel opmaken. Normaal zou je deze berekenen uit
# de frequenties van ordinale/nominale variabelen bij
# observaties met bv. table(waardering, geslacht)
waarderingen_m <- matrix(c(9,8,5,0,8,10,5,4), ncol = 2)
rownames(waarderingen_m) <-
  c("Goed", "Voldoende", "Onvoldoende", "Slecht")
colnames(waarderingen_m) <- c("Vrouw", "Man")
waarderingen <- as.table(waarderingen_m)

# Marginale totalen berekenen:
margin.table(waarderingen, 1)  # Rijtotalen
```

```r
margin.table(waarderingen, 2)  # Kolomtotalen
margin.table(waarderingen)     # Algemeen totaal (# observaties)

# Gepercenteerde waarden, over de rijtotalen
waarderingen_pct <- prop.table(waarderingen, 2)

# Berekening chi-kwadraat, de moeilijke manier
# Verwachte waarden (ahv matrix-vermenigvuldiging)
verwacht <- as.array(margin.table(waarderingen,1)) %*%
  t(as.array(margin.table(waarderingen,2))) /
  margin.table(waarderingen)
# Afwijkingen, gekwadrateerd en genormeerd
afwijkingen <- (waarderingen - verwacht) ^ 2 / verwacht
# Chi-kwadraat:
sum(afwijkingen)

# Rechtstreekse berekening chi-kwadraatwaarde
summ <- summary(waarderingen)
chi_sq <- summ$statistic

# Cramér's V
k <- min(nrow(waarderingen), ncol(waarderingen))
V <- sqrt(chi_sq/
            (margin.table(waarderingen) *
               (k - 1)))

# Plot: mosaic plot
plot(t(waarderingen))

# Clustered bar chart
barplot(waarderingen, beside = TRUE)

# Stacked percentage chart
barplot(waarderingen_pct, horiz = TRUE)
```

Voorbeeld 6.3. *In Tabel 6.6 worden de voorkeuren van vrouwen en mannen voor de gegeven automerken opgesomd. We zien dat nog steeds dertig van de honderd respondenten een voorkeur hebben voor de Mercedes, maar dat tweederde van deze dertig vrouwen zijn. We zouden ook kunnen zeggen dat de helft van de vrouwen een voorkeur heeft voor de Mercedes. Evenzo blijkt dat een derde van de mannen een voorkeur heeft voor een Alfa Romeo, tegenover geen van de vrouwen. Het lijkt alsof de onderscheiden automerken niet gelijkelijk gewaardeerd worden door mannen en vrouwen. Om dit te staven bepalen we χ^2 en Cramér's V. Probeer dit zelf, hetzij in R, hetzij met een rekenblad (Excel, Numbers, LibreOffice Calc)! We vinden:*

$$\chi^2 = 22.619$$

$$V = \sqrt{\frac{22.169}{100.(2-1)}} = 0.476$$

We vinden dus tussen een redelijk sterke tot sterke samenhang. ∎

	Mercedes	BMW	Porsche	Alfa Romeo	Totaal
Mannen	10	10	20	20	60
Vrouwen	20	5	15	0	40
Totaal	30	15	35	20	100

Tabel 6.6: Tabel die uitdrukt hoeveel vrouwen en hoeveel mannen een voorkeur voor een bepaald automerk hebben.

6.3 Regressie

Bij regressie gaan we proberen een consistente en systematische koppeling tussen de variabelen te vinden. Dat betekent concreet: "als we de waarde van de onafhankelijke variabele kennen, kunnen we dan ook de waarde van de afhankelijke variabele voorspellen?" We kennen twee soorten verbanden:

Monotoon: een monotoon verband is een verband waarbij de onderzoeker de algemene richting van de samenhang tussen de twee variabelen kan aanduiden, hetzij stijgend, hetzij dalend. De richting van het verband verandert nooit.

Niet-monotoon: bij een niet-monotoon verband wordt de aanwezigheid (of afwezigheid) van de ene variabele systematisch gerelateerd aan de aanwezigheid (of afwezigheid) van een andere variabele. De richting van het verband kan echter niet aangeduid worden.

Bij lineaire regressie gaan we ons beperken tot een lineair verband: een rechtlijnige samenhang tussen een onafhankelijke en afhankelijke variabele, waarbij kennis van de onafhankelijke variabele kennis over de afhankelijke variabele geeft.

Bij een lineair verband zijn er drie karakteristieken:

1. Aanwezigheid: is er wel een verband tussen de twee variabelen?
2. Richting: is er een dalend of een stijgend verband?
3. Wat is de sterkte van het verband: sterk, gematigd of niet-bestaand?

Een voorbeeld van een linear verband $y = \beta_0 + \beta_1 x$ vind je bijvoorbeeld in figuur 6.1.

Zo'n verband kunnen we vinden aan de hand van de kleinste kwadraten methode van Gauss. Dit wordt als volgt gedaan.

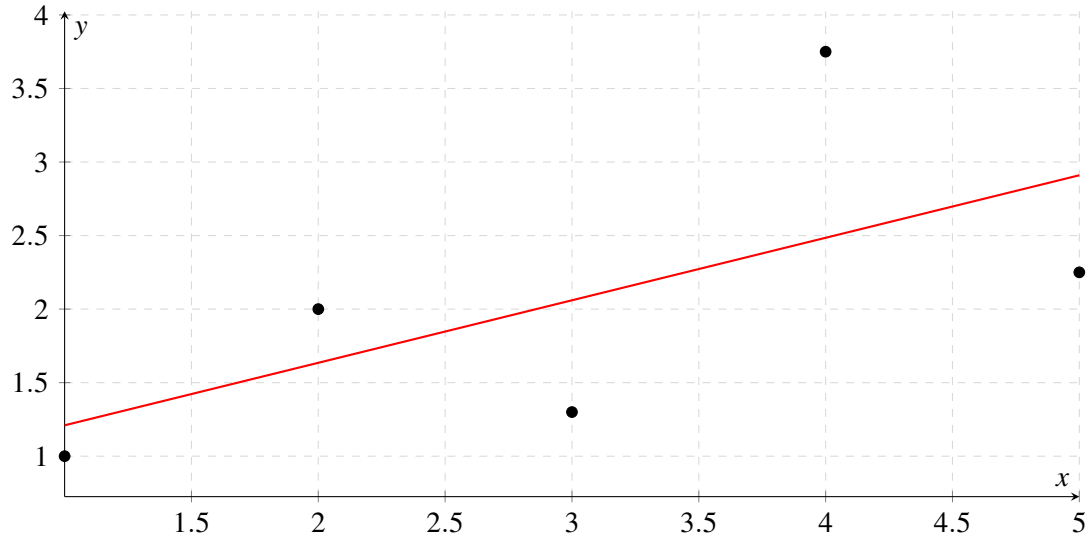
Stelling 1. *Een lineair verband wordt weergegeven als volgt:*

$$y = \beta_0 + \beta_1 x \tag{6.5}$$

met

- *y de afhankelijke*
- *x de onafhankelijke*

We willen hier de som van de kwadraten minimaliseren van de afwijkingen $e_i = y_i - (\beta_0 + \beta_1 x_i)$. Zo'n afwijking kan ook geschreven worden als (stel $X_i = x_i - \bar{x}$ en $Y_i = y_i - \bar{y}$):



Figuur 6.1: Een voorbeeld van een lineair verband

$$
\begin{aligned}
e_i &= y_i - \beta_1 x_i - \beta_0 && (6.6)\\
e_i &= (y_i - \overline{y}) - \beta_1 (x_i - \overline{x}) - (\beta_0 - \overline{y} + \beta_1 \overline{x}) && (6.7)\\
e_i &= Y_i - \beta_1 X_i - (\beta_0 - \overline{y} + \beta_1 \overline{x}) && (6.8)
\end{aligned}
$$

In stap 6.8 doen we eigenlijk $+\overline{x} - \overline{x} + \overline{y} - \overline{y}$, wat een nuloperatie is. Dit is een gedachtensprong die niet meteen voor de hand ligt, maar onthou dat dit een "shortcut" is naar de oplossing en dat het "echte" bewijs een stuk ingewikkelder is.

We willen de som van de kwadraten van e_i minimaliseren:

$$
\begin{aligned}
\sum_i^n e_i^2 &= \sum_i^n (y_i - (\beta_0 + \beta_1 x_i))^2 && (6.9)\\
&= \sum_i^n ((Y_i - \beta_1 X_i) - (\beta_0 - \overline{y} + \beta_1 \overline{x}))^2 && (6.10)\\
&= \sum_i^n (Y_i - \beta_1 X_i)^2 - 2 \sum_i^n (Y_i - \beta_1 X_i)(\beta_0 - \overline{y}) + \beta_1 \overline{x}) + (\beta_0 - \overline{y} + \beta_1 \overline{x}))^2 && (6.11)\\
&= \sum_i^n (Y_i - \beta_1 X_i)^2 + n(\beta_0 - \overline{y} + \beta_1 \overline{x})^2 && (6.12)
\end{aligned}
$$

We kunnen de stap maken van 6.11 naar 6.12 door volgende uit te werken:

$$\sum_i^n X_i = \sum_i^n (x_i - \overline{x}) = 0$$

en equivalent

$$\sum_i^n Y_i = \sum_i^n (y_i - \overline{y}) = 0$$

daardoor is

$$\sum_i^n (Y_i - \beta_1 X_i) = \sum_i^n Y_i - \beta_1 \sum_i^n X_i = 0$$

en bijgevolg dus ook

$$2\sum_i^n (Y_i - \beta_1 X_i)(\beta_0 - \overline{y})$$

Nu is e_i^2 geschreven als een som van twee positieve uitdrukkingen. Deze som is minimaal als beide uitdrukkingen minimaal zijn.

$$\begin{cases} \sum_i^n (Y_i - \beta_1 X_i)^2 \text{ is minimaal.} \\ n(\beta_0 - \overline{y} + \beta_1 \overline{x})^2 \text{ is minimaal} \end{cases} \tag{6.13}$$

Voor de eerste uitdrukking vinden we eigenlijk een kwadratische functie in β_1.

$$\begin{aligned} & \sum_i^n (Y_i - \beta_1 X_i)^2 \text{ is minimaal.} && (6.14) \\ \Leftrightarrow \quad & \sum_i^n (Y_i^2 - 2X_i Y_i \beta_1 + X_i^2 \beta_1^2 \text{ is minimaal.} && (6.15) \\ \Leftrightarrow \quad & \beta_1^2 \sum_i^n X_i^2 - 2\beta_1 \sum_i^n X_i Y_i + \sum_i^n Y_i^2 \text{ is minimaal.} && (6.16) \\ \Leftrightarrow \quad & \text{is minimaal als } \beta_1 = \tfrac{\sum_i^n X_i Y_i}{\sum_i^n X_i^2} && (6.17) \end{aligned}$$

Voor de tweede uitdrukking vinden we

$$\begin{aligned} & n(\beta_0 - \overline{y} + \beta_1 \overline{x})^2 \text{ is minimaal} \Leftrightarrow \quad n(\beta_0 - \overline{y} + \beta_1 \overline{x})^2 = 0 && (6.18) \\ \Leftrightarrow \quad & \beta_0 - \overline{y} + \beta_1 \overline{x} = 0 && (6.19) \\ \Leftrightarrow \quad & \beta_0 = \overline{y} - \beta_1 \overline{x} && (6.20) \end{aligned}$$

Eiwitgehalte%	Gewichtstoename (gram)
0	177
10	231
20	249
30	348
40	361
50	384
60	404

Tabel 6.7: De data die verzameld geweest is door de kerstman: per eiwitpercentage wordt de gewichtstoename beschouwd.

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})(y-\overline{y})$	$(x-\overline{x})^2$
0	177	-30	-130,71	3921,3	900
10	231	-20	-76,71	1534,2	400
20	249	-10	-58,71	587,1	100
30	348	0	40,29	0	0
40	361	10	53,29	532,9	100
50	384	20	76,29	1525,8	400
60	404	30	96,29	2888,7	900
				10990	2800

Tabel 6.8: Berekeningen die nodig zijn voor het toepassen van de kleinste kwadratenmethode.

met als oplossing

$$\begin{cases} \beta_1 = \frac{\sum_i^n X_i Y_i}{\sum_i^n X_i^2} \\ \beta_0 = \overline{y} - \beta_1 \overline{x} \end{cases} \tag{6.21}$$

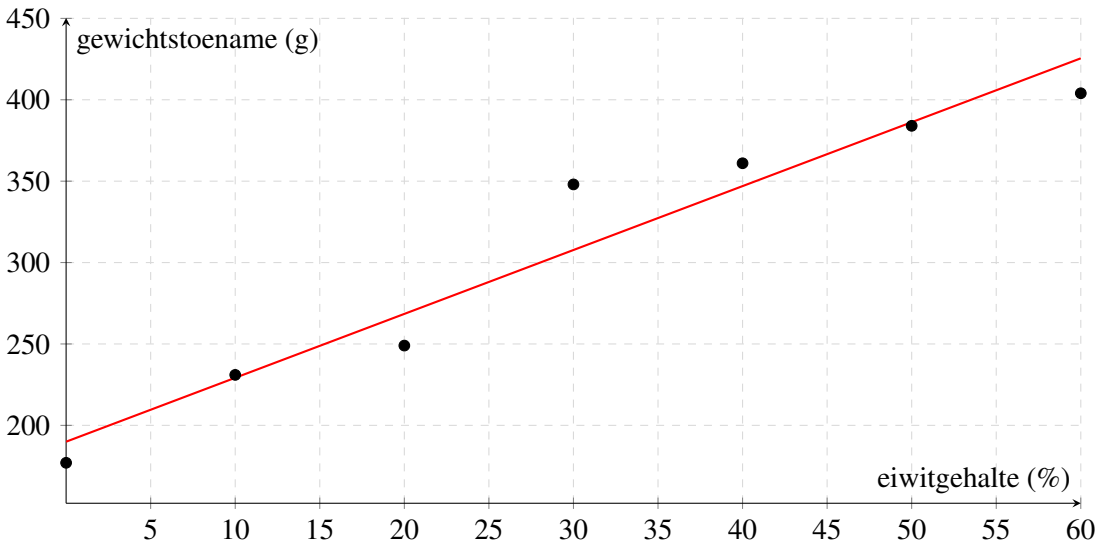
en dus

$$\beta_1 = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} \tag{6.22}$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x} \tag{6.23}$$

Voorbeeld 6.4. *We kijken naar het voorbeeld van de Kerstman en zijn rendieren. Hij wil zien of er een lineair verband bestaat tussen het eiwitgehalte van het voeder en de gewichtstoename van de rendieren. Hij voert een aantal proeven uit en bekomt de data in tabel 6.7. Door toepassing van de formules die hierboven staan bekomt men (zie tabel 6.8):*

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x - \overline{x})^2} = \frac{10990}{2800} = 3.925$$



Figuur 6.2: Lineair verband tussen eiwitgehalte en gewichtstoename

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 307.7143 - 3.925 \times 30 = 189.96$$

Men heeft dus een lineair verband gevonden die de kwadraten van de residuen minimaliseert. Let wel, er wordt niets gezegd over de sterkte of validiteit van dit verband. Dit verband wordt getekend in figuur 6.2. ∎

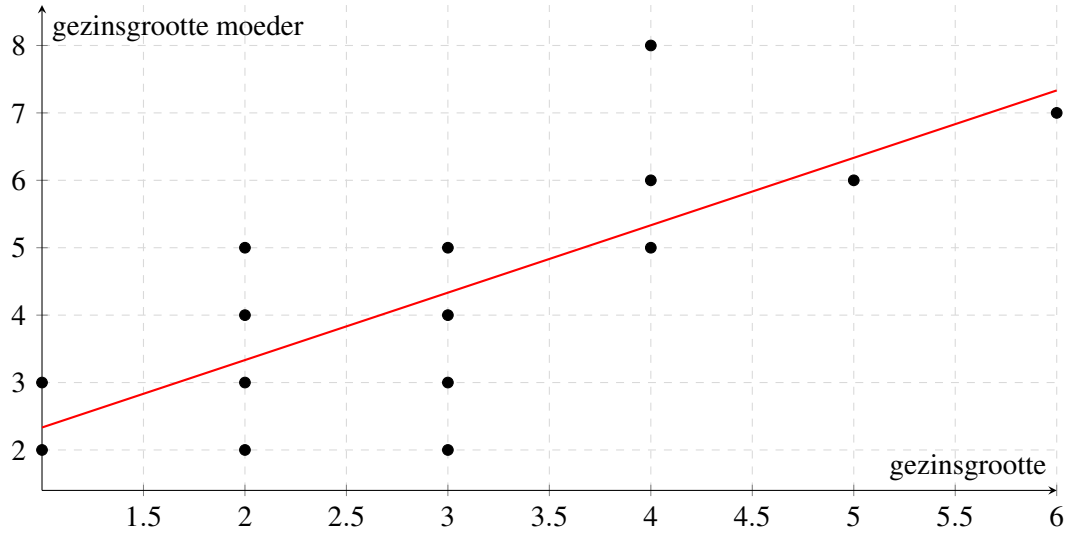
Voorbeeld 6.4 uitgewerkt in R (met plot van de regressierechte):

```
# Voorbeeld Lineaire Regressie
gewichtstoename <- read.csv('santa.txt', sep = "")
attach(gewichtstoename)

# x = eiwitgehalte (%)
# y = gewichtstoename (g)
plot(x, y,
     main = 'Gewichtstoename',
     xlab = 'eiwitgehalte (%)',
     ylab = 'gewichtstoename (g)')
# Bereken regressie ("Linear Model")
regr <- lm(y ~ x)
abline(regr, col = 'red') # Plot regressierechte
```

6.4 Correlatie

6.4.1 Pearsons product-momentcorrelatiecoëfficiënt

We kunnen twee statistieken bepalen die de sterkte van een lineair verband uitdrukken.



Figuur 6.3: Linear verband tussen grootte van een gezin en de grootte van de familie van de moeder

Definitie 6.4.1 (Pearsons product-momentcorrelatiecoëfficiënt)**.** *Pearsons product momentcorrelatiecoëfficiënt R (of kortweg correlatiecoëfficiënt) is een maat voor de sterkte van de lineaire samenhang tussen X en Y. De waarde kan variëren van -1 tot 1.*

- *Een waarde van +1 duidt een positief lineair verband aan.*
- *Een waarde van -1 duidt een negatief lineair verband aan.*
- *Een waarde van 0 wil zeggen dat er totaal geen lineaire samenhang is.*

Hoe dichter de correlatiecoëfficiënt bij 1 of -1, hoe beter de kwaliteit van het lineair model.

6.4.2 Determinatiecoëfficiënt

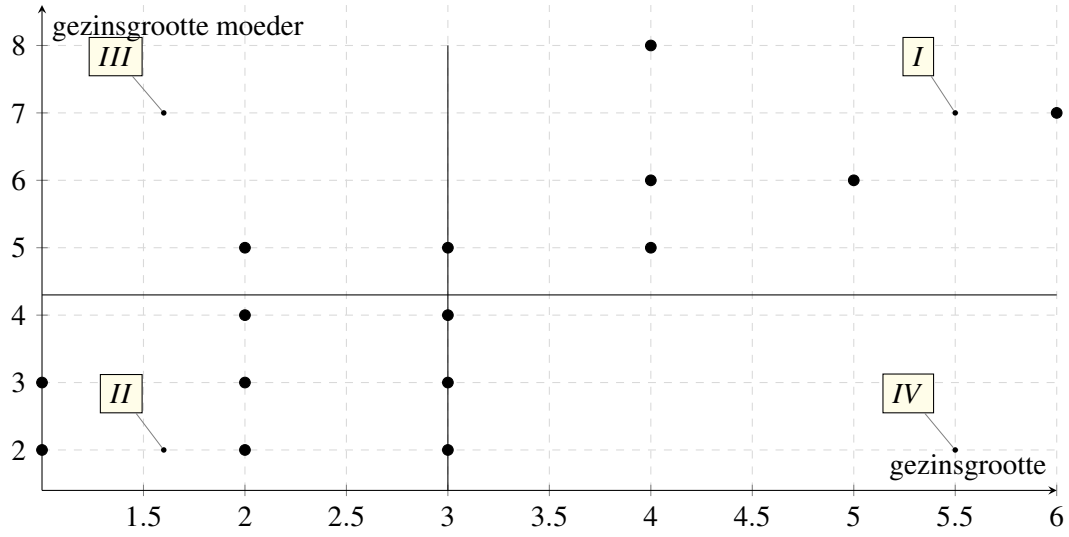
Definitie 6.4.2. *De determinatiecoëfficiënt (R^2) is het kwadraat van de correlatiecoëfficiënt en verklaart het percentage van de variantie van de waargenomen waarden t.o.v. de regressierechte.*

- *R^2 is de verklaarde variantie*
- *$1-R^2$ is de onverklaarde variantie*

Bepaling van R en R^2

Beschouw het voorbeeld in figuur 6.3: de grootte van een gezin vs. de grootte gezin moeder. We zien duidelijk dat er een linear verband is. Indien we de gemiddelde berekenen en de figuur in 4 kwadranten (kwadrant I, II, III, IV) volgens de gemiddelden verdelen krijgen we de figuur in 6.4. Dan kunnen we volgende situaties bekijken.

- Neem een element uit gebied I. Voor dit element is $x_i - \overline{x}$ positief en $y_i - \overline{y}$ ook. Dus is hun product. $(x_i - \overline{x})(y_i - \overline{y}) > 0$.
- Neem een element uit gebied II. Voor dit element is $x_i - \overline{x}$ negatief en $y_i - \overline{y}$ ook. Dus is hun product. $(x_i - \overline{x})(y_i - \overline{y}) > 0$.
- Neem een element uit gebied III. Voor dit element is $x_i - \overline{x}$ negatief en $y_i - \overline{y}$ positief. Dus is



Figuur 6.4: De figuur opgedeeld in 4 kwadranten

hun product. $(x_i - \bar{x})(y_i - \bar{y}) < 0$.

- Neem een element uit gebied IV. Voor dit element is $x_i - \bar{x}$ positief en $y_i - \bar{y}$ negatief. Dus is hun product. $(x_i - \bar{x})(y_i - \bar{y}) < 0$.

Aangezien dat er meer punten in gebieden I en II zijn dan in gebieden III en IV zal de som $\sum_i (x_i - \bar{x})(y_i - \bar{y})$ een positief getal zijn. Hoe meer punten in I en II, hoe groter het getal. We merken dus een sterk positief lineair verband.

Indien de punten ongeveer gelijk verdeeld zouden zijn over de vier gebieden vinden we dat deze soms dicht bij nul zal zijn. Omgekeerd, indien er een negatief lineair verband zou zijn vinden we een negatief getal.

We hebben dus een maat gevonden om het verband tussen twee variabelen te meten:

- Stijgende gecorreleerde verbanden is $\sum_i (x_i - \bar{x})(y_i - \bar{y})$ positief en groot.
- Dalende gecorreleerde verbanden is $\sum_i (x_i - \bar{x})(y_i - \bar{y})$ negatief en groot (in absolute waarde).
- Met niet gecorreleerde variabelen is $\sum_i (x_i - \bar{x})(y_i - \bar{y})$ klein in absolute waarde.

We kunnen deze maat onafhankelijk maken van de grootte van de steekproef door te delen door de steekproefgrootte n. Dit noemen we de co-variantie en wordt gedefinieerd als gemeenschappelijke spreiding:

$$Cov(X,Y) = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n} \tag{6.24}$$

Dit geeft ons de gemiddelde afwijking per meetpunt.

Om opnieuw te normaliseren (een variatie in X is niet per se van dezelfde grootteorde als een variatie in Y) gaan we de maatstaf voor het gezamelijk variëren onafhankelijk maken van het aantal waarnemingen en de orde van grootte van de getalswaarden. Zo kunnen we deze waarden universeel vergelijkbaar maken. Daarom delen we de co-variantie door het product van de standaardafwij-

kingen en noemen we de relatieve co-variantie of Pearson's correlatiecoëfficiënt ook bekend als product-moment-correlatiecoëfficiënt of kortweg als correlatiecoëfficiënt.

$$R = \frac{COV(X,Y)}{\sigma_x \sigma_y} \tag{6.25}$$

$$= \frac{COV(X,Y)}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}} \times \sqrt{\frac{\sum(y_i-\bar{y})^2}{n}}} \tag{6.26}$$

$$= \frac{\sum_i^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i^n (x_i-\bar{x})^2}\sqrt{\sum_i^n (y_i-\bar{y})^2}} \tag{6.27}$$

De correlatiecoëfficiënt is onafhankelijk van de meeteenheid terwijl de covariantie afhankelijk is van de meeteenheid.

R^2 interpretatie

Als we aannemen dat x niet bijdraagt aan de voorspelling van y dan is de beste voorspelling voor een waarde van y het steekproefgemiddelde $\bar{y}$, dat in figuur 6.6 als een horizontale lijn wordt weergegeven. De verticale lijnstukken zijn de afwijkingen van de waargenomen punten y van deze voorspelling (het steekproefgemiddelde). De som van de kwadraten van deze afwijkingen is:

$$SS_{yy} = \sum (y_i - \bar{y})^2$$

Indien we aannemen dat x wel een rol speelt bij de voorspelling van y, berekenen we de regressielijn bij dezelfde gegevensverzameling en de afwijkingen van de punten ten opzichte van de lijn zoals in figuur 6.5.

$$SSE_{yy} = \sum (y_i - \hat{y})^2$$

Als we nu de afwijkingen vergelijken met elkaar zien we het volgende:

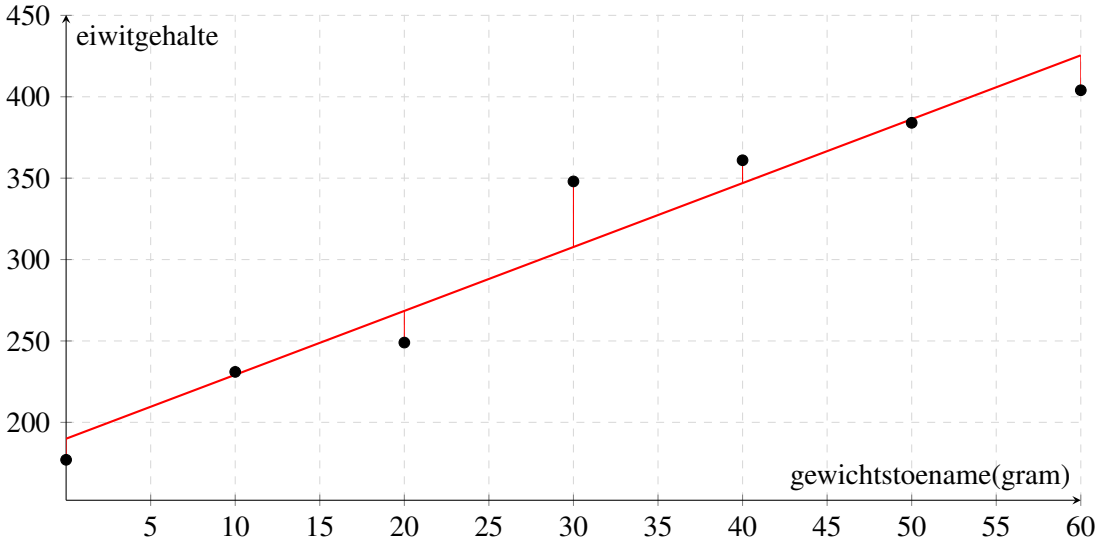
1. Als x weinig of niet bijdraagt in de voorspelling zullen de sommen van de kwadraten van de afwijkingen van de twee lijnen nagenoeg dezelfde zijn:

   $$SS_{yy} = \sum (y_i - \bar{y})^2$$

   en

   $$SSE_{yy} = \sum (y_i - \hat{y})^2$$

   waarbij $\hat{y}$ de voorspelde waarde is.



Figuur 6.5: Deviaties tot de regressierechte: aanname x geeft extra informatie voor het voorspellen van y.

2. Als x wel bijdraagt tot de voorspelling van y zal SSE kleiner zijn dan SS_{yy}. In feite zal

$$SSE_{yy} = \sum (y_i - \widehat{y})^2$$

gelijk zijn aan nul als alle punten perfect voorspeld worden (en dus op de regressierechte liggen).

De vermindering in de som van de kwadraten die toegeschreven kan worden aan het opnemen van x in het model is dan (uitgedrukt in fractie van SS_{yy})

$$\frac{SS_{yy} - SSE_{yy}}{SS_{yy}}$$

We noemen SS_{yy} de totale steekproefvariantie van de meetwaarden rond het steekproefgemiddelde $\bar{y}$ en SSE_{yy} de overblijvende niet-verklaarde steekproefvariantie, na het schatten van de lijn $\widehat{y} = \beta_0 + \beta_1 x$. Dus dan is $(SS_{yy} - SSE_{yy})$ de verklaarde variantie die toe te schrijven is aan de lineaire relatie met x.

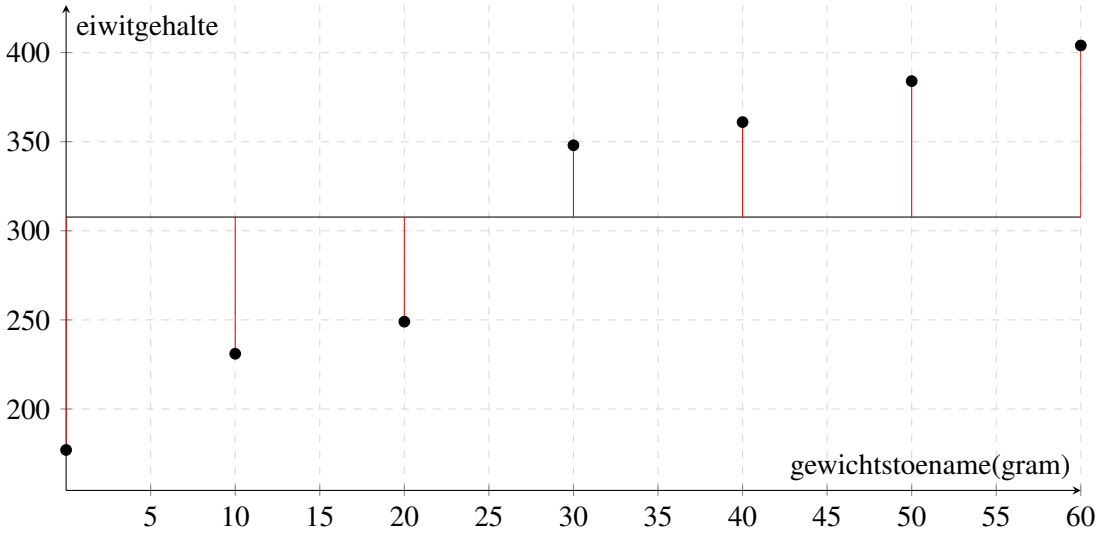
Er kan nu worden aangetoond dat bij enkelvoudige lineaire regressie deze fractie

$$\frac{SS_{yy} - SSE_{yy}}{SS_{yy}} = \frac{\text{verklaarde variantie}}{\text{totale steekproefvariantie}}$$

gelijk is aan het kwadraat van de pearsoncorrelatiecoëfficiënt. (= het deel van de totale variantie dat verklaard wordt door de lineaire rechte).

6.5 Conclusie

Er bestaan verschillende soorten verbanden tussen variabelen. Wij zijn geïnteresseerd in monotone en lineaire verbanden. We beschikken hier over een correlatiecoëfficiënt en lineaire regressie. Deze technieken mogen niet met nominale en ordinale variabelen gebruikt worden. Een kleine waarde



Figuur 6.6: Deviaties tot de gemiddelde van y: aanname x geeft geen informatie voor het voorspellen van y ($\bar{y} = 307.71$).

$(= 0)$ voor een maat voor verband betekent alleen dat het overeenkomend verband afwezig is: er kan een ander soort verband aanwezig zijn. Het gebruik van een spreidingsdiagram is dus altijd aan te raden.

Het feit dat twee variabelen gecorreleerd zijn betekent niet dat de ene de oorzaak is van de andere.

6.6 Samenvatting

In dit hoofdstuk zijn verschillende technieken voorgesteld om na te gaan of er een verband bestaat tussen twee variabelen. De ene variabele noemen we de *onafhankelijke*, de andere de *afhankelijke* variabele. Wat we willen uitzoeken is of de waarde van de onafhankelijke variabele een impact heeft op die van de afhankelijke.

De technieken die we kunnen gebruiken (hetzij rekenkundige, hetzij voor visualisatie), hangen af van het meetniveau van de onderzochte variabelen. Tabel 6.9 geeft een overzicht.

6.7 Oefeningen

De databestanden voor deze oefeningen zijn te vinden op Github (in de directory *oefeningen/data/hfst6_2variabelen*).

Oefening 6.1. *Marktonderzoek toont aan dat achtergrondmuziek in een supermarkt invloed kan hebben op het aankoopgedrag van de klanten. In een onderzoek werden drie methoden met elkaar vergeleken: geen muziek, Franse chansons en Italiaanse hits. Telkens werd het aantal verkochte flessen Franse, Italiaanse en andere wijnen geteld (***Ryan1998***).*

De onderzoeksdata bevindt zich in het csv-bestand MuziekWijn.

Meetniveau variabele			
Onafhankelijke	**Afhankelijke**	**Numeriek**	**Visualisatie**
Kwalitatief	Kwalitatief	χ^2 Cramér's V	mozaïekdiagram geclusterd staafdiagram rependiagram
Kwalitatief	Kwantitatief	t-toets voor 2 steekproeven	boxplot (evt. staafdiagram gemiddelde met standaardafwijking)
Kwantitatief	Kwantitatief	covariantie correlatiecoëfficiënt determinatiecoëfficiënt	spreidings-/XY-diagram regressierechte

Tabel 6.9: Overzicht technieken voor de analyse van twee variabelen.

Vragen:

1. *Stel de correcte kruistabel op. Gebruik hiervoor het R-commando table om de frequentietabel te bekomen.*
2. *Bepaal de marginalen.*
3. *Bepaal de verwachte resultaten.*
4. *Bereken manueel de χ^2 toetsingsgrootheid.*
5. *Bereken manueel de Cramér's V. Wat kan je hieruit besluiten?*

■

Oefening 6.2. *Gebruik dezelfde data.*

1. *Stel de percentages verkochte wijnen voor in een staafdiagram met de muziekconditie= Geen.*
2. *Stel de percentages verkochte wijnen voor in een geclusterd staafdiagram (clustered bar chart).*
3. *Stel de percentages verkochte wijnen voor in rependiagram (stacked bar chart).*

■

Oefening 6.3. *Lees het databestand "Aardbevingen.csv" in.*

1. *Maak een histogram en een boxplot van de variabele "Magnitudes".*
2. *Maak een lijngrafiek met het aantal aardbevingen per maand.*
3. *Onderzoek of er een verband bestaat tussen de variabelen "Type" en "Source". Bereken ook de Cramér's V-waarde. Wat is de conclusie?*

■

Oefening 6.4. *In onderstaande tabel vindt men voor elke rij (= persoon) het resultaat van een test en zijn examenscore. Gevraagd:*

- *Bepaal handmatig de regressierechte $\beta_0 + \beta_1 x$.*
- *Bepaal handmatig de correlatie- en determinatiecoëfficiënt (R, R^2)*
- *Geef uitleg bij de gevonden statistieken.*

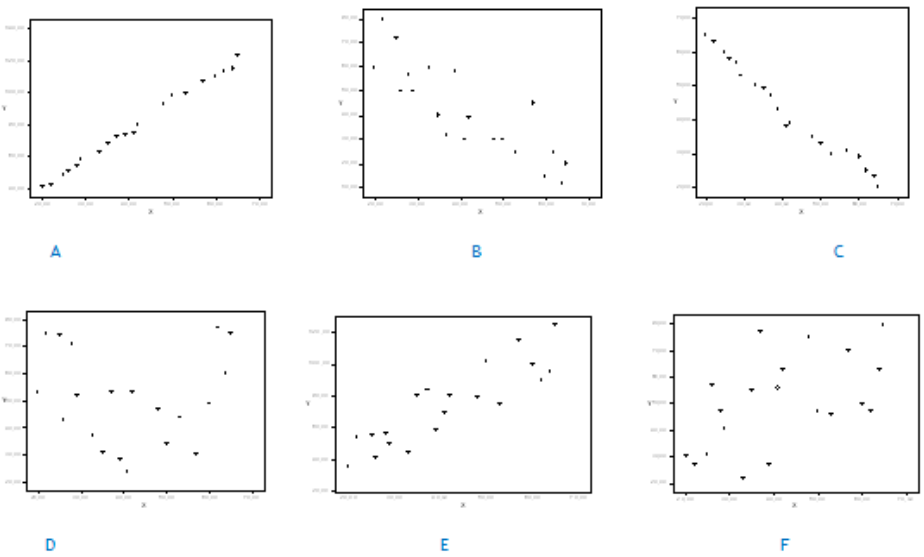
Resultaat Test (X)	*Examenresultaat (Y)*
10	*11*
12	*14*
8	*9*
13	*13*
9	*9*
10	*9*
7	*8*
14	*14*
11	*10*
6	*6*

Tabel 6.10: Scores test en examen voor aantal personen

■

Oefening 6.5. *Gegeven 6 scatterplots in volgende figuur en onderstaande correlatiecoëfficiënten. Match de coëfficiënten met de scatterplots. Er is dus één scatterplot waarvan geen correlatie gegeven staat hieronder.*

- $r_1 = 0.6$
- $r_2 = 0$
- $r_3 = -0.9$
- $r_4 = 0.9$
- $r_5 = 0.3$



Figuur 6.7: Correlaties

■

Oefening 6.6. *Lees het databestand "Cats.csv" in.*

1. *Voer een lineaire regressieanalyse uit op de variabelen Lichaamsgewicht (`Bwt`, afhankelijke variabele) en Gewicht hart (`Hwt`, onafhankelijke variabele).*
2. *Maak een spreidingsdiagram van beide variabelen.*
3. *Bereken en teken de regressielijn.*
4. *Bereken de correlatie- en de determinatiecoëfficiënt.*
5. *Geef een interpretatie van deze resultaten.*

■

Oefening 6.7. *Gebruik dezelfde data als in vorige oefening.*

1. *Voer een lineaire regressieanalyse uit op de variabelen Lichaamsgewicht (Bwt) en Gewicht hart (Hwt) per geslacht.*
2. *Maak een spreidingsdiagram van beide variabelen voor elk van de geslachten.*
3. *Bereken en teken telkens de regressielijn.*
4. *Bereken de correlatie- en de determinatiecoëfficiënt.*
5. *Geef een interpretatie aan deze resultaten.*

■

Oefening 6.8. *Lees het databestand "Pizza.csv" in.*

1. *Voer een volledige lineaire regressieanalyse uit op de variabelen Rating en CostPerSlice. Trek hieruit de juiste conclusies en ga deze ook grafisch na.*
2. *Onderzoek een mogelijk verband tussen Rating en Neighbourhood. Welke methode kan je hiervoor gebruiken? Kan je de gegevens van Rating hiervoor in dezelfde vorm gebruiken?*
3. *Geef een interpretatie aan deze resultaten.*
4. *Stel de kruistabel grafisch voor met een staafdiagram. Voorzie een legende.*

■

6.7.1 Antwoorden op geselecteerde oefeningen

Oefening 6.1

$\chi^2 \approx 18,2792$, Cramér's $V \approx 0,1939$

Oefening 6.4

- $\beta_0 \approx 0,6333$, $\beta_1 \approx 0.9667$
- $Cov \approx 6,444$, $R \approx 0,9352$, $R^2 \approx 0,8747$

Oefening 6.6 en 6.7

Selectie	β_0	β_1	Cov	R	R^2
Hele dataset	-0.3511	4.0318	0.9496	0.8041	0.6466
Female	2.9813	2.6364	0.1979	0.5320	0.2831
Male	-1.1768	4.3098	0.9419	0.7930	0.6289

7. De χ^2 toets

7.1 χ^2 toets voor verdelingen

Wanneer alle variabelen in het onderzoek nominaal zijn, is chi kwadraat de eenvoudigste (en populairste) techniek die men ter beschikking heeft voor het toetsten van hypothesen. De teststatistiek heet chi kwadraat, en is verdeeld volgens de chi kwadraat verdeling. De test kan gebruikt worden om na te gaan in welke mate de steekproef overeenstemt met een nulhypothese over de verdeling van de variabele. Men noemt dit een *goodness of fit* test.

In het voorbeeld op de slides willen we nagaan, of de verdeling van onze steekproef bij $n = 400$ superhelden overeenstemt met de verdeling die je verwacht in de volledige populatie (de verzameling van alle mogelijke superhelden).

Daartoe vergelijken we de aantallen in de steekproef met de aantallen die je zou verwachten als de steeproef exact representatief zou zijn naar de types van superhelden. Als deze verschillen relatief groot zijn dan komt de verdeling in de steekproef niet overeen met de verdeling in de populaties en zullen we moeten concluderen dat de steekproef niet representatief is. Om te oordelen of deze verschillen relatief groot zijn voeren we een χ^2toets uit.

7.1.1 Voorbeeld superhelden

We willen kijken of de steekproef voor onze superhelden representatief is. Als de steekproef exact representatief zou zijn zouden we verwachten dat in de steekproef 35% van de superhelden een mutant zou zijn. Het verwachte aantal of de verwachte frequentie voor deze categorie is dus gelijk aan $0,35 \times 400 = 140$. De verwachte frequenties worden genoteerd met de letter e (expected). Er geldt dus:

$$e = n \times \pi$$

met π de frequentie over de hele populatie. Als de verschillen $o - e$ (o staat voor observed) reletief klein zijn kunnen ze toegerekend worden aan toevallige steekproeffouten. We gaan nu een toetsingsgrootheid bepalen waarmee getoetst kan worden of de steekproefverdeling overeenkomt met de gegeven verdeling in de populatie.

Beschouw χ^2:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

We merken op:

- indien de verschillen klein zijn $\Rightarrow$ verdeling komt voldoende overeen
- indien de verschillen groot $\Rightarrow$ verdeling niet representatief

We bepalen nu een kritieke grenswaarde g die een χ^2 verdeling heeft. Hierbij speel het aantal vrijheidsgraden een rol (df). Er geldt:

$$df = k - 1$$

met k het aantal categorieën. In ons voorbeeld hebben we $df = 5 - 1 = 4$. Om de kritieke grenswaarde te bepalen, kan je gebruik maken van een tabel voor de χ^2-verdeling. Voor een gegeven significantieniveau α en vrijheidsgraad df kan je in zo'n tabel de grenswaarde aflezen.

In ons voorbeeld is $\chi^2 = 3,47$ met grenswaarde $g = 9,49$. Omdat de gevonden toetsingsgrootheid $\chi^2 = 3,47 < g = 9,49$, mogen we besluiten dat de steekproef representatief is.

7.1.2 Toetsingsprocedure

We volgen de stappen van een statistische toetsingsprocedure:

1. **Bepalen hypotheses** Als nulhypothese formuleren we dat de verdeling over de opleidingen in de steekproef gelijk is aan de verdeling in de populatie. Als alternatieve hypothese formuleren we dat de verdelingen verschillend zijn.
   - H_0: steekproef is representatief naar populatie
   - H_1: steekproef is niet representatief naar populatie
2. **Bepalen α en n :** $\alpha = 0,05$ en $n = 400$.
3. **Toetsingsgrootheid en waarde ervan in steekproef**:
   $$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$
4. **Bereken en teken kritiek gebied**: de toets is altijd rechtszijdig. Is de toetsingsgrootheid kleiner dan kritieke grenswaarde verwerp H_0 niet, anders verwerp H_0 en aanvaard H_1.

7.1.3 Voorbeeld 2

Beschouw alle gezinnen met 5 kinderen in een bepaalde gemeenschap. Met betrekking tot samenstelling zijn er 6 mogelijkheden.

1. 5 jongens
2. 4 jongens, 1 meisje
3. 3 jongens, 2 meisjes
4. 2 jongens, 3 meisjes
5. 1 jongen, 4 meisjes
6. 5 meisjes

Het onderzoek bevat 1022 gezinnen met 5 kinderen en resultaten staan beschreven in de slides (kolom = aantal jongens). Zijn de waargenomen aantallen in de 6 klassen representatief voor een populatie waar de kans om een jongen te krijgen = kans om een meisje te krijgen = 0,5?

Indien de veronderstelling waar is wordt de kans π_i om i jongens te krijgen bepaald door een binominaalvedeling met parameters $n = 5$ en $p = 0,5$.

Dit kan je eenvoudig nagaan aan de hand van voorbeeld. De kans om 2 jongens te krijgen met 5 kinderen is gelijk aan :

$$(0,5)^2 \times (1-0,5)^{5-2} \times \binom{5}{2}$$

Algemeen geldt dus:

$$\pi_i = \binom{5}{i} \times 0,5^i \times 0,5^{5-i} = \frac{5!}{i!(5-i)!} \times 0,5^i$$

Met deze π_i kunnen we dus de verwachte waarde bepalen en de stappen volgen zoals hierboven beschreven.

1. **Bepalen hypotheses**
   - H_0: steekproef is representatief naar populatie
   - H_1: steekproef is niet representatief naar populatie
2. **Bepalen α en n** : $\alpha = 0,01$ en $n = 1022$.
3. **Toetsingsgrootheid en waarde ervan in steekproef**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} = 29,5766$$

4. **Bereken en teken kritiek gebied**: kritieke grens is 15,0863. Onze toetsingsgrootheid ligt dus in het kritieke gebied dus verwerpen we H_0.

We vinden dus dat de steekproef niet representatief is naar een populatie waar geldt dat de kans op een jongen even groot is als de kans op een meisje. Het is interessant om te kijken naar de

	Longkanker	Niet	Wel	Totaal
Roker	**Wel**	21178	83	21261
	Niet	3092	1	3093
	Totaal	24270	84	24354

Tabel 7.1: Resultaten van het onderzoek van **Doll1954**

gestandaardiseerde residuen die aanduiden welke klassen de grootste bijdrage leveren aan de waarde van de grootheid.

$$r_i = \frac{O_i - n\pi_i}{\sqrt{n\pi_i(1-\pi_i)}}$$

Oefening 7.1. *Hoe komen we hier aan de noemer? Waar komt dit mee overeen? Hoe bepaal je de variantie van een binomiale verdeling? Antwoord:* $n \times \pi(1-\pi)$ ■

Er geldt algemeen dat waarden groter dan 2 of kleiner dan -2 extreem zijn. We kunnen dus besluiten dat het aantal gezinnen waarin alle kinderen hetzelfde geslacht hebben groter mag worden genoemd dan verwacht.

7.1.4 Voorwaarden

Om de toets te mogen toepassen dient aan de volgende voorwaarden te zijn voldaan (Regel van Cochran)

1. Voor alle categorieën moet gelden dat de verwachte waarde e groter is dan 1.
2. In ten hoogste 20 % van de categoriën mag de verwachte waarde e kleiner dan 5 zijn.

7.2 χ^2-kruistabeltoets

De Chi-kwadraattoets laat zich eenvoudig uitbreiden tot een onderzoeksontwerp met twee variabelen, met respectievelijk r en k niveaus. Hier gaan we onderzoeken of er een verband is tussen 2 variabelen. De procedure kan opnieuw geformuleerd worden.

We gaan de procedure na aan de hand van een studie door **Doll1954** over de relatie tussen roken en longkanker. Doll en Hill schreven in 1951 alle Britse huisartsen aan met het verzoek om gegevens over hun leeftijd en rookgedrag. Vervolgens hielden ze jarenlang de overlijdensberichten en de doodsoorzaak bij en herhaalden hun periodiek. De eerste uitkomsten, na circa vier jaar, zijn in tabel 7.1 samengevat. Uit de tabel kan makkelijk geconcludeerd worden dat er geen relatie is tussen roken en longkanker. In (ruim) vier jaar is slechts $(84/24354) * 100 = 0,35\%$ van de Britse artsen aan longkanker overleden en dat met slechts $(83/21261) * 100 = 0,39\%$ van de rokers onder hen. Dit is weinig, maar het is wel veel meer dan hetzelfde cijfer voor de niet-rokers $(1/3093) * 100 = 0,032\%$.

We zien in de tabel dat er wel een erg groot verschil is tussen de geobserveerde aantallen rokers die overlijden aan longkanker en de verwachte waarden in deze cel. Hetzelfde geldt voor het geringe aantal huisartsen dat niet rookt, maar wel aan longkanker overleden is. Deze observatie maakt ons wel wantrouwig of de eerdere tentatieve conclusie wel juist is. We kunnen om aan deze onzekerheid de toetsingsgrootheid χ^2 uitrekenen. Dat doen we op de vertrouwde manier:

1. **Bepalen hypotheses**
   - H_0: in de populatie is er geen samenhang tussen onafhankelijke en afhankelijke variabele
   - H_1: er bestaat wel een samenhang tussen de variabelen in de populatie
2. **Bepalen α en n :** $\alpha = 0,05$ en $n = 24354$.
3. **Toetsingsgrootheid en waarde ervan in steekproef**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{E_i} = 10,35$$

4. **Bereken en teken kritiek gebied**: kritieke grens is 3,8415 en aantal vrijheidsgraden $df = (r-1)(k-1)$ Onze toetsingsgrootheid ligt dus in het kritieke gebied dus verwerpen we H_0.

We moeten derhalve H_0, dat er geen relatie is tussen beide variabelen, verwerpen ten gunste van H_1 dat er wel een relatie is tussen beide variabelen: rokers sterven vaker aan longkanker dan niet-rokers.

Maar, is dit nu een bewijs dat zoals zo vaak verondersteld wordt dat roken longkanker veroorzaakt? Nee, dat is het absoluut niet. Een paar alternatieve verklaringen: niet alle rokers krijgen longkanker, de rokers zijn ouder dan de niet-rokers, de rokers wonen veelal in de grote steden met meer vervuilde lucht dan de niet-rokers die veelal op het platte land wonen, ook zo erg nog een speciale genetische dispositie kunnen zijn, die zowel van invloed is op de verslaving aan tabak, als op de kans om longkanker te krijgen. Voor een causale interpretatie van de gegevens (let wel, het betreft hier immers geen experiment), moeten we op zijn minst de beschikking hebben over een theorie die de relatie tussen roken en longkanker expliciteert.

7.3 Oefeningen

Oefening 7.2. *Voor deze oefening maken we gebruik van de dataset* `survey` *die is meegeleverd met R. De dataset is samengesteld uit een bevraging onder studenten. Om deze te laden, doe het volgende:*

```
library(MASS)
View(survey)   # Toont de "survey" dataset
?survey        # Help-pagina voor deze dataset met uitleg
    over de inhoud
```

Als je een foutboodschap krijgt bij het laden van de bibliotheek (eerste regel), betekent dit dat de package `MASS` *nog niet geïnstalleerd is. Dit kan je alsnog doen via Tools > Install Packages en het invullen van de package-naam in het tekstveld.*

We willen de relatie onderzoeken tussen enkele discrete (nominale of ordinale) variabelen in deze dataset. Voor elke hieronder opgesomde paren, volg deze stappen:

(a) *Denk eerst eens na welke uitkomt je precies verwacht voor de opgegeven combinatie van variabelen.*
(b) *Stel een frequentietabel op voor de twee variabelen. De (vermoedelijk) onafhankelijke variabele komt eerst.*
(c) *Plot een grafiek van de data, bv. geclusterde staafgrafiek, gestapelde staafgrafiek van relatieve frequenties, of een "mozaïekgrafiek" (eenvoudig met* `plot(table(data$col1, data$col2))`*).*
(d) *Als je de grafiek bekijkt, verwacht je dan een eerder hoge of eerder lage waarde voor de χ^2-statistiek? Waarom?*
(e) *Bereken de χ^2-statistiek en de kritieke grenswaarde g (voor significantieniveau $\alpha = 0.05$)*
(f) *Bereken de p-waarde*
(g) *Moeten we de nulhypothese aanvaarden of verwerpen? Wat betekent dat concreet voor de relatie tussen de twee variabelen?*

Hieronder zijn de te onderzoeken variabelen opgesomd. De vermoedelijke onafhankelijke variabele komt telkens eerst.

1. `Exer` *(sporten) en* `Smoke` *(rookgedrag)*
2. `W.Hnd` *(de hand waarmee je schrijft) en* `Fold` *(de hand die bovenaan komt als je de armen kruist)*
3. `Sex` *(gender) en* `Smoke`
4. `Sex` *en* `W.Hnd`

■

Oefening 7.3. *Laad de dataset* `Aids2` *uit package* `MASS` *(zie Oefening 7.2) die informatie bevat over 2843 patiënten die vóór 1991 in Australië met AIDS besmet werden. Deze dataset werd in detail besproken door* **Ripley2007**. *Onderzoek of er een relatie is tussen de variabele geslacht (*`Sex`*) en de manier van besmetting (*`T.categ`*).*

1. *Ga op de gebruikelijke manier te werk: visualiseren van de data, χ^2, g en p-waarde berekenen ($\alpha = 0{,}05$), en tenslotte een conclusie formuleren.*
2. *Bepaal de gestandaardiseerde residuën om te bepalen welke categorieën extreme waarden bevatten.*

■

Oefening 7.4. *Elk jaar voert Imec (voorheen iMinds) een studie uit over het gebruik van digitale technologieën in Vlaanderen, de Digimeter (***Vanhaelewyn2016***). In deze oefening zullen we nagaan of de steekproef van de Digimeter 2016 ($n = 2164$) representatief is voor de bevolking wat betreft de leeftijdscategorieën van de deelnemers.*

In Tabel 7.2a worden de relatieve frequencies van de deelnemers weergegeven. De absolute frequenties voor de verschillende leeftijdscategorieën van de Vlaamse bevolking worden samengevat in Tabel 7.2b. Deze gegevens zijn ook te vinden in bijgevoegd CSV-bestand `oefeningen/data/bestat-vl-ages.csv`*.*

1. *De tabel met leeftijdsgegevens van de Vlaamse bevolking als geheel heeft meer categorieën dan deze gebruikt in de Digimeter. Maak een samenvatting zodat je dezelfde categorieën*

overhoudt dan deze van de Digimeter. Tip: dit gaat misschien makkelijker in een rekenblad dan in R.

2. *Om de goodness-of-fit test te kunnen toepassen hebben we de absolute frequenties nodig van de geobserveerde waarden in de steekproef. Bereken deze.*
3. *Bereken ook de verwachte percentages (π_i) voor de populatie als geheel.*
4. *Voer de goodness-of-fit test uit over de verdeling van leeftijdscategorieën in de steekproef van de Digimeter. Is de steekproef in dit opzicht inderdaad representatief voor de Vlaamse bevolking?*

■

7.4 Antwoorden op geselecteerde oefeningen

Oefening 7.2

1. `Exer/Smoke`: $\chi^2 = 5.49$, $g = 12.5916$, $p = 0.35$
2. `W.Hnd/Fold`: $\chi^2 = 1.581399$, $g = 5.9915$, $p = 0.454$
3. `Sex/Smoke`: $\chi^2 = 3.554$, $g = 7.8147$, $p = 0.314$
4. `Sex/W.Hnd`: $\chi^2 = 0.236$, $g = 3.8415$, $p = 0.627$

Oefening 7.3

$\chi^2 = 1083.372914$, $g = 14.067140$, $p \approx 1.157 \times 10^{-229}$

Oefening 7.4

$\chi^2 = 6.6997$, $g = 12.5916$, $p = 0.35$

Tabel 7.2: Frequenties van de leeftijd van deelnemers aan de iMec Digimeter 2016 en de Vlaamse bevolking.

Leeftijdsgroep	**Percentage**
15-19	6,6%
20-29	14,2%
30-39	15,0%
40-49	16,3%
50-59	17,3%
60-64	7,3%
64+	23,2%

(a) Percentage van deelnemers aan de Digimeter 2016 van iMec ($n = 2164$), opgedeeld per leeftijdscategorie. (**Vanhaelewyn2016**)

Leeftijdsgroep	**Aantal**
–5	352017
5-9	330320
10-14	341303
15-19	366648
20-24	375469
25-29	387131
30-34	401285
35-39	409587
40-44	458485
45-49	493720
50-54	463668
55-59	413315
60-64	379301
65-69	299152
70-74	279789
75-79	249260
80-84	182352
85-89	104449
90-94	29888
95-99	7678
100+	923

(b) Absolute frequentie van de Vlaamse bevolking per leeftijdscategorie. Bron: BelStat (https://bestat.economie.fgov.be/bestat/, C01.1: Bevolking volgens verblijfplaats (provincie), geslacht, positie in het huishouden (C), burgerlijke staat en leeftijd (B)).

8. Tijdreeksen

In de voorgaande hoofdstukken hebben we telkens data geanalyseerd die op een bepaald moment in de tijd is verzameld en we hebben uitspraken gedaan over die data voor dat specifieke moment.

In de ict-beroepspraktijk zijn er echter ook vele toepassingen waar het nodig is om data op te volgen die voortdurend verandert. We denken dan bijvoorbeeld aan de belasting van een processor, evolutie van schijfgebruik op een opslagapparaat, de responstijd van een website, enz.

In dit hoofdstuk gaan we dit soort data onder de loep nemen en de belangrijkste analysemethoden bespreken.

8.1 Tijdreeksen & voorspellingen

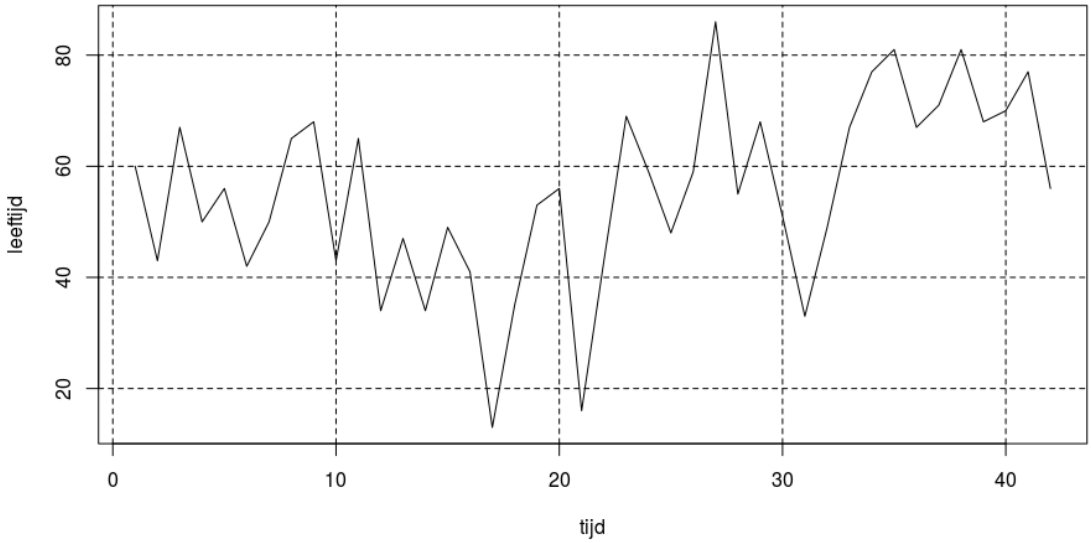
Definitie 8.1.1 (Tijdreeks)**.** *Een* tijdreeks *is een opeenvolging van observaties van een willekeurige variabele in functie van de tijd.*

Voorbeelden:

- maandelijkse vraag naar melk
- jaarlijkse instroom van studenten bij de Hogeschool
- dagelijks debiet van een rivier
- de buitentemperatuur over het verloop van een dag

Het voorspellen van tijdreeksen is een belangrijk onderdeel van onderzoek omdat ze vaak de basis vormen voor beslissingsmodellen. Voorbeelden hiervan zijn:

- algemene ontwikkeling van toekomstplannen (investeringen, capaciteit . . .)
- plannen van budgettering om tekortkomingen te vermijden (operationeel budget, marketing



Figuur 8.1: De tijdreeks die de leeftijden van de koningen voorstelt.

budget . . .)
- competitieve leveringstijden van een bedrijf
- ondersteuning van financiële objectieven
- onzekerheid vermijden
- de mogelijkheid om ontwikkelingen in de verkeersveiligheid kwantitatief te modelleren

Tijdreeksen modelleren is een statistisch probleem: we gaan ervan uit dat de observaties variëren volgens een bepaalde kansdichtheidsfunctie in functie van de tijd. Vaak gaan we ervan uit dat de observaties in een tijdreeks gecorreleerd zijn en dus niet uit een willekeurige steekproef komen.

Er zijn verschillende types modellen in gebruik voor het analyseren van tijdreeksen. Deze modellen hebben met elkaar gemeen dat ze in principe niet alleen de ontwikkeling in een geobserveerde tijdreeks kunnen beschrijven, maar dat we ze ook kunnen gebruiken om (i) verklaringen te vinden voor die ontwikkeling en (ii) om de toekomstige waarden van de tijdreeks te voorspellen. Hun geschiktheid voor het verwezenlijken van deze doelstellingen loopt echter sterk uiteen. In dit hoofdstuk beperken we ons tot het gebruik van tijdreeksen met een geschiedenis om tijdsafhankelijke modellen te bepalen. Een voorbeeld van een tijdreeks is bijvoorbeeld de leeftijd van de opeenvolgende koningen van Engeland startend van Willem De Veroveraar (**Hipel1994**).

```
kings <- scan(file = 'cursus/data/tijdreeksen/kings.data', skip
    = 3)
kingstimeseries <- ts(kings)
plot.ts(kingstimeseries, ylab='leeftijd', xlab="tijd")
grid(lty=2,lwd=1,col='black')
```

8.2 Tijdreeksmodellen

8.2.1 Wiskundig model

Ons doel is het opstellen van een model dat een verklaring vindt voor de geobserveerde data en dat toelaat om observaties in de toekomst zo goed mogelijk te voorspellen. Het simpelste model dat je kan bedenken is een model waarbij een constante b gebruikt wordt met variaties rond b bepaald door een willekeurige variabele ε_t zoals in vergelijking 8.1.

$$X_t = b + \varepsilon_t \tag{8.1}$$

X_t stelt een *variabele* voor dat de onbekende is op tijdstip t.
x_t stelt een *observatie* voor op tijdstip t (en is dus gekend).
ε_t noemt met de *storing* (Eng. *noise*) en wordt geacht een gemiddelde van 0 te hebben met variantie σ^2 en normaal verdeeld ($\varepsilon_t \sim Nor(0,\sigma)$).

We kunnen ook ervan uit gaan dat er een lineair verband is:

$$X_t = b_0 + b_1 \times t + \varepsilon_t \tag{8.2}$$

De vergelijking in 8.1 en 8.2 zijn speciale gevallen van het polynomiaal geval:

$$X_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_n t^n + \varepsilon_t \tag{8.3}$$

Oefening 8.1. *Wat zou volgende tijdreeks kunnen voorstellen?*

$$X_t = b_0 + b_1 \sin\left(\frac{2\pi t}{4}\right) + b_1 \cos\left(\frac{2\pi t}{4}\right) + \varepsilon_t \tag{8.4}$$

■

Antwoord: dit is een cyclische tijdreeks met periode = 4. Dit zou bijvoorbeeld kunnen gebruikt worden bij een tijdreeks voor seizoenen.

```
f <- function(a, b, t){
  return(a + b * sin((2 * pi*4)/4) + b * cos((2 * pi*4)/4) +
    rnorm(1))
}
t <- seq(from = 1, to = 100, by = 1)
X <- lapply(t,f,a=5,b=5)
plot(x = t, y=X, type = 'l')
```

4	16	12	25	13	12	4	8	9	14
3	14	14	20	7	9	6	11	3	11
8	7	2	8	8	10	7	16	9	4

Tabel 8.1: Voorbeeld van tijdreeksdata, gevisualiseerd in figuur 8.2

Algemeen

In elk model beschouwd is de tijdreeks een functie van tijd en parameters van het model. We kunnen algemeen stellen dat:

$$X_t = f(b_0, b_1, b_2, \ldots, b_t, t) + \varepsilon_t \tag{8.5}$$

We aanvaarden vervolgens nog volgende stellingen:

- Het model gaat uit van twee componenten van variabiliteit: het gemiddelde van de voorspellingen verandert met de tijd en de variaties tot dit gemiddelde variëren willekeurig.
- De residuen van het model $(X_t - x_t)$ zijn homoscedastisch : dat wil zeggen in de tijd een constante variantie hebben.

Eenmaal het model gekozen, rest enkel nog het probleem van het schatten van de parameters voor vergelijking 8.5. Dit is wat in de volgende stukken besproken zal worden.

8.3 Schatten van de parameters

Eenmaal een model geselecteerd wordt, is het aan de onderzoeker om de parameters te gaan schatten, i.e. parameters die ervoor zorgen dat het model de geobserveerde waarden zo goed mogelijk benaderen. Meestal gaan we ervan uit dat alle waarden gelijkwaardig zijn, maar dat is niet zo bij tijdreeksen. Aangezien onze onafhankelijke parameter de tijd is moeten we methoden bekomen die ervoor zorgen dat recentere data belangrijker zijn dat oude data of omgekeerd.

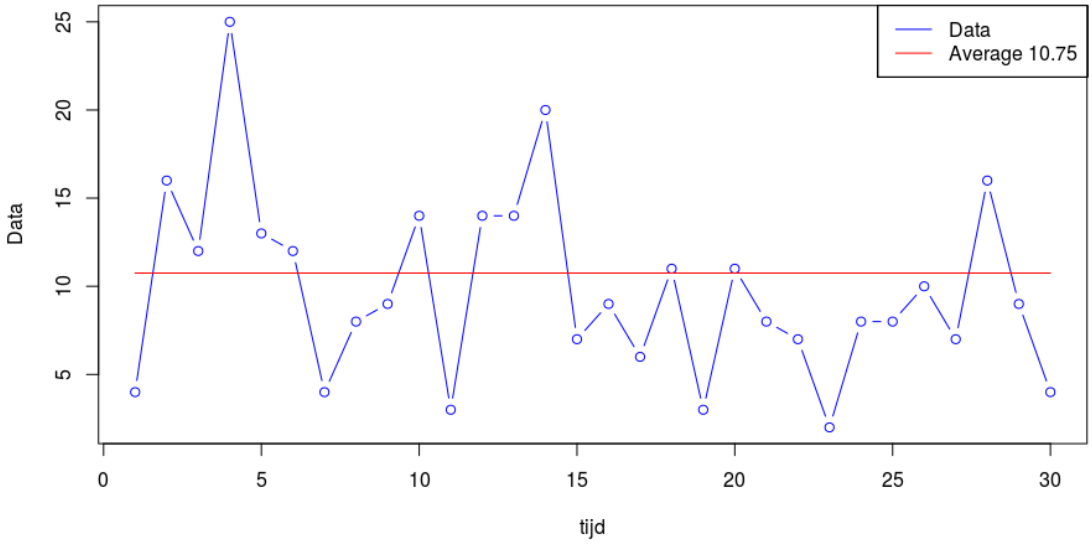
In wat volgt beschrijven we de tijdreeksen met geschatte waarden voor de parameters. We duiden schatters aan met een hoedje op de parameters:

$$\widehat{b}_1, \widehat{b}_2 \ldots \widehat{b}_n$$

8.3.1 Voortschrijdend gemiddelde

Stel dat de statisticus de data in tabel 8.1 tot het twintigste datapunt beschikbaar heeft (bekende data). De onderzoeker kent de datapunten vanaf het twintigste datapunt niet en moet deze gaan voorspellen. Een eerste model dat gebruikt zou kunnen worden is het constante model zoals in formule 8.1.

Volgens dit model worden de datapunten beschouwd als willekeurige waarden uit een populatie met gemiddelde b. De beste schatter voor b is het gemiddelde van deze twintig datapunten.



Figuur 8.2: Tijdreeks met constant gemiddelde 10.75

```
data <- c(4  , 16  , 12  , 25  , 13  , 12  , 4  , 8   , 9  , 14,
+           3  , 14  , 14  , 20  , 7   , 9   , 6  , 11  , 3  , 11,
+           8  , 7   , 2   , 8   , 8   , 10  , 7  , 16  , 9  , 4 )
mean(data[1:20])
```

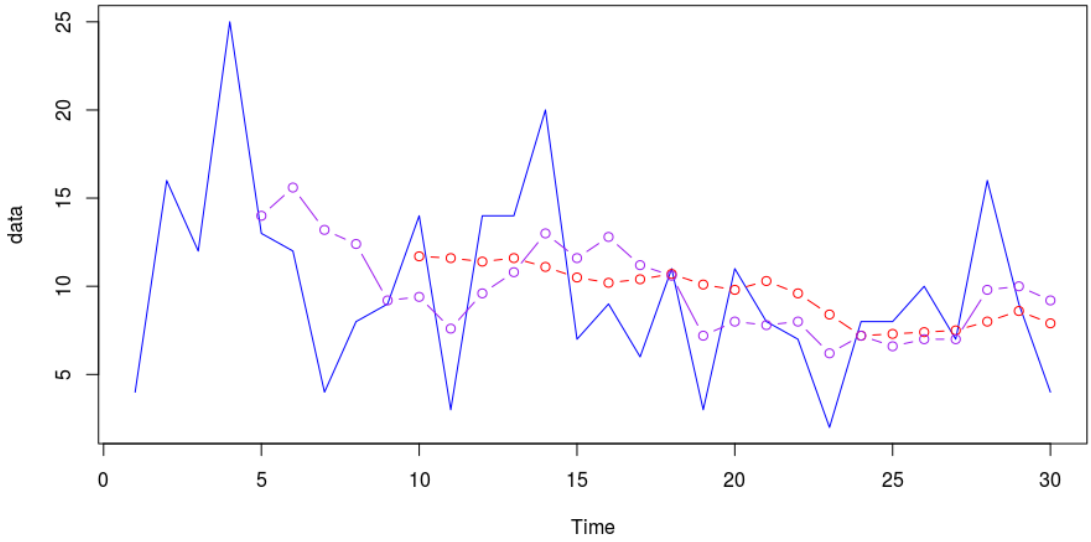
$$\widehat{b} = \frac{1}{20}\sum_{1}^{20} x_t = 10.75$$

Dit is de beste schatter vertrekkende van de 20 datapunten. We merken wel op dat $x_1 = 4$ evenveel *waarde* heeft als $x_{20} = 11$, of anders verwoord: de coëfficiënt van x_1 is dezelfde als die van x_{20}, namelijk $\frac{1}{20}$.

Indien we dit als schatter zouden gebruiken dan zien we dat dit in figuur 8.2 geen goed idee is.

```
AV20 <- matrix(10.75,30,1)
plot.ts(data, col="blue", type='b', xlab='tijd', ylab='Data')
lines(AV20,col='red', type='l')
legend(x= 'topright',legend = c("Data","Average 10.75"), lty = c
    (1,1), lwd = c(2.5,2.5), col=c('blue','red'))
```

Indien we veronderstellen dat de data verandert met de tijd is het beter om oude data minder te laten meetellen dan recentere. Een mogelijkheid is om enkel recente data te gebruiken, bijvoorbeeld de 10 of 5 laatste datapunten (zie figuur 8.3).



Figuur 8.3: Tijdreeks met voortschrijdend gemiddelde $m = 10$ en $m = 5$

.

$$\widehat{b} = \frac{1}{10}\sum_{10}^{20} x_t = 10.18$$

en

$$\widehat{b} = \frac{1}{5}\sum_{15}^{20} x_t = 7.83$$

```
sma10 <- SMA(x =data ,n=10)
sma5 <- SMA(x=data ,n=5)
plot.ts(x = data, col = 'blue',type = 'l')
lines(sma10, col='red', type = 'b')
lines(sma5, col='purple', type = 'b')
```

Dit worden *voortschrijdende gemiddelden* genoemd (Eng. *moving average*).

Welke schatter is nu de beste? We kunnen dit nu nog niet zeggen.

- De schatter die alle datapunten gebruikt is de beste indien de tijdreeks het model volledig volgt.
- De schatter met de recentere datapunten is de beste indien de tijdreeks verandert met de tijd.

Definitie 8.3.1 (voortschrijdend gemiddelde)**.** *Algemeen is het* voortschrijdend gemiddelde *(Eng.* moving average*) het gemiddelde van de m laatste observaties.*

$$\widehat{b} = \sum_{i=k}^{t} \frac{x_i}{m} \tag{8.6}$$

Tabel 8.2: Voorspellingsfout voor een moving average $m = 10$

met $k = t - m + 1$. m is de time range en is de parameter van de methode.

8.3.2 Meten van de nauwkeurigheid van voorspellingen

Een methode om de voorspelling te meten is het gemiddelde van de deviaties (Eng. *Mean Average Deviation*, afgekort *MAD*): gemiddelde absolute verschil tussen het voorspelde en de werkelijke waarden van de tijdreeks.

Definitie 8.3.2 (*MAD*)**.**

$$MAD = \frac{1}{n}\sum_{1}^{n}|e_i| \tag{8.7}$$

Je kan dit ook percenteren om zo tot de gemiddelde absolute procentuele afwijking (Eng. *Mean Absolute Percentage Error*, afgekort *MAPE*) te komen.

Definitie 8.3.3 (*MAPE*)**.**

$$MAPE = \frac{1}{n}\sum_{1}^{n}\left|\frac{e_i}{X_i}\right| \tag{8.8}$$

Je kan ook de variantie van de fouten bepalen:

Definitie 8.3.4 (*VAR*)**.**

$$s_e^2 = \frac{1}{m}\sum_{1}^{n}(e_i - \overline{e})^2 \tag{8.9}$$

Als laatste interessante parameter kan gekeken worden naar de wortel uit de gemiddelde kwadratische afwijking (Eng. *root-mean-squared error*, afgekort *RMSE*), als de wortel uit het gemiddelde gekwadrateerde verschil tussen de voorspelde en de werkelijke waarden van de tijdreeks.

Definitie 8.3.5 (*RMSE*)**.**

$$RMSE_e = \sqrt{\frac{1}{m}\sum_{1}^{n}(e_i)^2} \tag{8.10}$$

8.4 Exponentiële afvlakking

Bij een voortschrijdend gemiddelde krijgen alle voorgaande observaties een gelijk gewicht. Bij exponentiële afvlakking (Eng. *exponential smoothing*) worden kleinere gewichten toegekend aan oudere observaties. Met andere woorden, recentere observaties krijgen relatief meer gewicht dan oudere observaties.

In het geval van het eenvoudig voortschrijdend gemiddelde zijn de gewichten hetzelfde, namelijk $\frac{1}{m}$.

8.4.1 Enkelvoudige exponentiële afvlakking

Exponentiële afvlakking is een gewogen gemiddelde dat positieve gewichten toekent aan de huidige waarden en waarden uit het verleden van de tijdreeks. Een enkel gewicht, $0 \leq \alpha \leq 1$ of de afvlakkingsconstante (Eng. *smoothing constant*) wordt hiervoor gekozen. Voor een tijdseenheid t wordt de enkelvoudige exponentiële afvlakking gevonden door vergelijking 8.11.

Definitie 8.4.1 (Exponentiële afvlakking)**.**

$$X_t = \alpha x_t + (1-\alpha)X_{t-1}, 0 \leq \alpha \leq 1, t \geq 3 \tag{8.11}$$

Met andere woorden, X_t is een gewogen gemiddelde van de huidige waarneming x_t en de vorige exponentiële afvlakking X_{t-1}.

Intiële setting

Het bepalen van X_2 is een belangrijke parameter. Men kan kiezen om:

1. $X_2 = x_1$ te stellen
2. X_2 gelijk te stellen aan een bepaald objectief
3. Een gemiddelde te nemen van de eerste x observaties
4. ...

Waarom wordt dit een exponentiële methode genoemd? Als we zouden substitueren vinden we bv. voor X_{t-1}:

$$X_t = \alpha x_t + (1-\alpha)\left[\alpha x_{t-1} + (1-\alpha)X_{t-2}\right]$$

$$X_t = \alpha x_{t-1} + \alpha(1-\alpha)x_{t-1} + (1-\alpha)^2 X_{t-2}$$

of dus algemeen gesteld :

$$X_t = \alpha \sum_{i=0}^{t-2}(1-\alpha)^{i-1}x_{t-i} + (1-\alpha)^{t-2}X_2, t \geq 2$$

Zo merk je dat oudere componenten een exponentieel kleiner gewicht verkrijgen.

Waarde voor α

De snelheid waarmee de oude observaties "vergeten" worden hang af van α. Met een α dicht bij 1 vergeet je snel, terwijl een α dicht bij nul ervoor zorgt dat vergeten minder snel gaat (zoals aangetoond in tabel 8.3). Vaak wordt een waarde gebruikt tussen 0.10 en 0.30.

Bijvoorbeeld, het bestand `precip.data` bevat totale jaarlijkse neerslag in inches voor Londen, vanaf 1813-1912. Laten we dit eens analyseren met R.

α	$(1-\alpha)$	$(1-\alpha)^2$	$(1-\alpha)^3$	$(1-\alpha)^4$
0.9	0.1	0.01	0.001	0.0001
0.5	0.5	0.25	0.125	0.062
0.1	0.9	0.81	0.729	0.6561

Tabel 8.3: Waarden voor α en $(1-\alpha)^n$

```
rain <- scan("cursus/data/tijdreeksen/precip.data",skip=1)
rainseries <- ts(rain, start=c(1813))
plot.ts(rainseries)
plot(rainseriesforecasts)
```

Voorspelling met exponentiële effening

Stel dat het doel is om de volgende waarde X_{t+1} te voorspellen, dan wordt dit gelijk gesteld aan de afvlakkingswaarde op tijdstip t.

$$X_{t+1} = EMA_t = X_t \tag{8.12}$$

Met X_t de laatst voorspelde waarde.

We kunnen dit eenvoudig uitvoeren in R. Je krijgt hierbij een *prediction interval*, een interval waarin we verwachten dat de voorspelde waarde met een bepaalde waarschijnlijkheid zal liggen. Standaard krijg je een 80% en een 95% interval.

```
library('forecast')
rainseriesforecasts2 <- forecast.HoltWinters(rainseriesforecasts
    , h=8)
plot.forecast(rainseriesforecasts2)
```

We zouden correlaties mogen zien tussen de voorspellingsfouten voor opeenvolgende voorspellingen. Met andere woorden, als er sprake is van een correlatie tussen prognosefouten voor opeenvolgende voorspellingen, is het eerder waarschijnlijk dat de simpele exponentiële afvlakking kan worden verbeterd door een andere voorspellingstechniek te gebruiken.

Om te achterhalen of dit het geval is, kunnen we een correlogram verkrijgen van de in-sample voorspellingsfouten voor.

We weten nog uit hoofdstuk 6 dat de covariantie of correlatie de lineaire relatie beschrijft tussen twee variabelen. De autocovariantie en autocorrelatie meten de lineaire relatie tussen in de tijd verschoven waarden voor een tijdreeks.

Definitie 8.4.2 (Autocovariantie)**.** *We definiëren de autocovariantie bij vertraging k door* c_k.

$$c_k = \sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})$$

Definitie 8.4.3 (Autocorrelatie). *We definiëren de autocorrelatie bij vertraging k door r_k.*

$$r_k = \frac{c_k}{c_0}$$

Een correlogram is een grafiek van de autocorrelaties. In R kan je die plotten met de functie `acf()`. Deze berekent ook de voorspellingsfouten. Om de maximale vertraging te bepalen die we willen bekijken, gebruiken we de parameter `lag.max`.

Bijvoorbeeld, om een correlogram te berekenen van de prognosefouten voor de Londen-regenvalgegevens voor vertragingen 1 tot en met 20, typen we:

```
acf(rainseriesforecasts2$residuals, lag.max=20, na.action = na.
    pass)
```

Om te testen of er significant bewijs is voor significante correlaties bij vertraging 1-20, kunnen we een Ljung-Box test uitvoeren. Dit kan in R worden gedaan met de functie `Box.test()`. De maximale vertraging die we willen bekijken, wordt gespecificeerd met behulp van de parameter `Lag`.

De test volledig uitleggen valt buiten het bereik van deze cursus, maar de test gaat uit van de hieronder geformuleerde hypothesen H_0 en H_1. De teststatistieken kunnen dan gewoon geïntepreteerd worden zoals alle andere hypothesetesten die beschreven geweest zijn in vorige hoofdstukken.

- H_0 De gegevens zijn onafhankelijk verdeeld (d.w.z. de correlaties in de populatie waaruit de sample wordt genomen, zijn 0, zodat elke waargenomen correlatie in de data voortvloeien uit willekeurigheid).
- H_1 De gegevens zijn niet onafhankelijk verdeeld: ze tonen een linaire correlatie.

Bijvoorbeeld, om te testen of er geen nul autocorrelaties zijn op vertragingen 1-20, voor de in-sample voorspellingen fouten voor Londen regenval data, typen we:

```
Box.test(rainseriesforecasts2$residuals, lag=20, type="Ljung-Box
    ")
Box-Ljung test
data:  rainseriesforecasts2$residuals
X-squared = 17.4008, df = 20, p-value = 0.6268
```

Als laatste moeten we ook kijken naar de verdeling van de fouten van de voorspelling. Zoals boven vermeld gaan we ervan uit dat de fouten normaal verdeeld zijn met een gemiddelde $\mu = 0$ en een standaardafwijking die constant is. Om deze veronderstelling te controleren, kunnen we een histogram van de prognosefouten plotten, met een overlappende normale curve met gemiddelde nul en dezelfde standaardafwijking heeft als de verdeling van de voorspellingsfouten. Hiervoor kunnen we een R-functie `plotForecastErrors()` definiëren. Het is ook aangewezen de methoden zoals beschreven in sectie 4.4.2.

```
plotForecastErrors <- function(forecasterrors)
{
# make a histogram of the forecast errors:
mybinsize <- IQR(forecasterrors)/4
mysd   <- sd(forecasterrors)
```

Data	Enkelvoudige afvlakking
6.4	
5.6	6.4
7.8	6.2
8.8	6.7
11.0	7.3
11.6	8.4
16.7	9.4
15.3	11.6
21.6	12.7
22.4	15.4

Tabel 8.4: Enkelvoudige afvlakking met $\alpha = 0.3$

```
mymin <- min(forecasterrors) - mysd*5
mymax <- max(forecasterrors) + mysd*3
# generate normally distributed data with mean 0 and standard
    deviation mysd
mynorm <- rnorm(10000, mean=0, sd=mysd)
mymin2 <- min(mynorm)
mymax2 <- max(mynorm)
if (mymin2 < mymin) { mymin <- mymin2 }
if (mymax2 > mymax) { mymax <- mymax2 }
# make a red histogram of the forecast errors, with the normally
    distributed data overlaid:
mybins <- seq(mymin, mymax, mybinsize)
hist(forecasterrors, col="red", freq=FALSE, breaks=mybins)
# freq=FALSE ensures the area under the histogram = 1
# generate normally distributed data with mean 0 and standard
    deviation mysd
myhist <- hist(mynorm, plot=FALSE, breaks=mybins)
# plot the normal curve as a blue line on top of the histogram
    of forecast errors:
points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
}
```

8.4.2 Dubbele exponentiële afvlakking

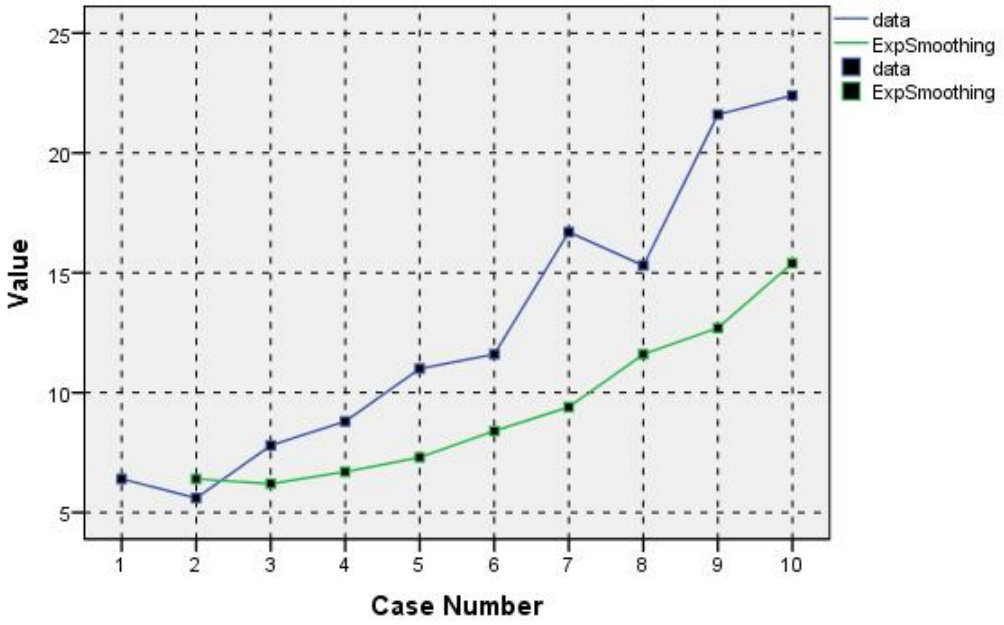
Enkelvoudige afvlakking wordt gebruikt wanneer er geen trend zichtbaar is. Wanneer er een trend (stijgend of dalend) is dan kan er iets fout gaan. Zie bijvoorbeeld de data in tabel 8.4 en figuur 8.4.

Daarom voegen we een extra constante toe om deze trap te overbruggen:

Definitie 8.4.4 (Holt-voorspelling of dubbele exponentiële afvlakking)**.**

$$X_t = \alpha x_t + (1-\alpha)(X_{t-1} + b_{t-1}) \quad 0 \leq \alpha \leq 1 \tag{8.13}$$

$$b_t = \beta(X_t - X_{t-1}) + (1-\beta)b_{t-1} \quad 0 \leq \beta \leq 1 \tag{8.14}$$



Figuur 8.4: Exponentiële afvlakking bij een trend

Initiële waarde

Net zoals in enkelvoudige afvlakking kan je verschillende methodes kiezen om initiële waardes voor X_t en b_t te kiezen:

- $X_1 = x_1$
- $b_1 = x_2 - x_1$
- $b_1 = \frac{1}{3}\left[(x_2 - x_1) + (x_1 - x_2) + (x_4 - x_3)\right]$
- $b_1 = \frac{x_n - x_1}{n-1}$

Voorspelling

Een voorspelling maken met dubbele exponentiële afvlakking gebeurt dan iets anders (noem F_{t+1} de voorspelling voor tijd $T+1$):

$$F_{t+1} = X_t + b_t$$

of

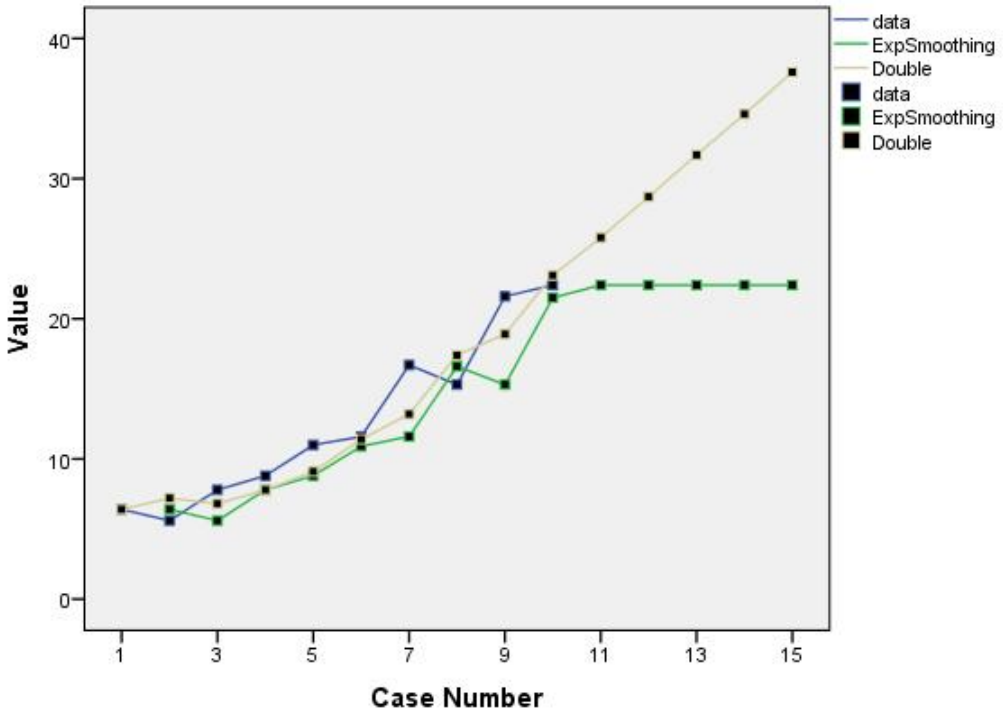
$$F_{t+m} = X_t + mb_t$$

Als we nu de tekening maken met enkelvoudige afvlakking ($\alpha = 0.977$) en dubbele afvlakking ($\alpha = 0.3623, \beta = 1.0, X_1 = x_1 = 6.4$ en $b_1 = \frac{1}{3}\left[(x_2 - x_1) + (x_1 - x_2) + (x_4 - x_3)\right] = 0.8$ vinden we volgende waarden in tabel 8.5 en figuur 8.5:

De manier om dit met R op te lossen is gelijkaardig als bij exponentiële afvlakking, alleen moet de parameter gamma β niet op NULL gezet worden. Het maken van de correlogram, de Ljung–Box test en het testen van de normaliteit van de errors gebeurt om dezelfde manier.

Data	Enkelvoudige afvlakking X_t	Dubbele afvlakking X_t	F_t
6.4		6.4	
5.6	6.4	6.6	7.2
7.8	5.6	7.2	6.8
8.8	6.7	8.1	7.8
11.0	8.8	9.8	9.1
11.6	10.9	11.5	11.4
16.7	11.6	14.5	13.2
15.3	16.6	16.7	17.4
21.6	15.3	19.9	18.9
22.4	21.5	22.8	23.1

Tabel 8.5: Tabel met enkelvoudige en dubbele afvlakking



Figuur 8.5: Enkelvoudige en dubbele afvlakking

8.4.3 Driedubbele exponentiële afvlakking

In vele tijdreeksen zie je bepaalde patronen terugkomen. Neem bijvoorbeeld is de dagelijkse omzet van een bakker: die zal elke dag anders zijn, maar als je die over een lange periode bekijkt, zal je wellicht elke week een gelijkaardig patroon terugzien (met bv. een topomzet op zondag).

Dit soort tijdreeksen kan benaderd worden met de Holt-Winters methode, of driedubbele exponentiële afvlakking.

$$X_t = \alpha \frac{x_t}{c_{t-L}} + (1-\alpha)(X_{t-1} + b_{t-1}) \qquad \text{Smoothing} \tag{8.15}$$

$$b_t = \beta(X_t - X_{t-1}) + (1-\beta)b_{t-1} \qquad \text{Trend smoothing} \tag{8.16}$$

$$c_t = \gamma \frac{x_t}{X_t} + (1-\gamma)c_{t-L} \qquad \text{Seasonal smoothing} \tag{8.17}$$

$$F_{t+m} = (X_t + mb_t)c_{t-L+m \mod L} \qquad \text{Voorspelling} \tag{8.18}$$

met

- x_t de observatie op tijdstip t
- X_t is de afgevlakte observatie op tijdstip t
- b_t is de trendfactor op tijdstip t
- c_t is de seizoensindex op tijdstip t
- F_t is de voorspelling op tijdstip t
- L is de periode (bv. van de seizoenen)

$\alpha, \beta en \gamma$ zijn constanten die geschat moeten worden. De manier om dit met R op te lossen is gelijkaardig als bij exponentiële afvlakking. Het maken van de correlogram, de Ljung–Box test en het testen van de normaliteit van de errors gebeurt op dezelfde manier.

8.5 Oefeningen

Oefening 8.2. *In bijgevoegd bestand* Budget.csv *vind je vanaf 1981 tot 2005 per kwartaal de omzet, het advertentiebudget en het BNP van een middelgroot bedrijf. Voeg zelf nog een kolom 'Kwartaalnummer' toe.*

1. *Bereken het voortschrijdend gemiddelde* (simple moving average) *over de periodes 4 en 12 voor deze data. Gebruik hiervoor de methode SMA. Maak een lijngrafiek van X, SMA*(4) *en SMA*(12).
2. *Welke techniek die we eerder gezien hebben (in het deel over beschrijvende statistiek) is ook geschikt om voorspellingen te maken over de waarden van X? Werk dit uit aan de hand van de daarvoor bestemde functie en plot het resultaat in de grafiek.*
3. *Gebruik de methode* forecast *om voorspellingen voor de 10 volgende periodes met elk van voorgaande methoden (dus moving average 4 en 10 en regressie) te maken. Teken deze eveneens op de grafiek.*
4. *Is het gebruik van één van deze technieken interessant om voor deze data voorspellingen te maken?*

5. *Maak van de data een tijdreeks via de methode* ts. *Gebruik de methode* decompose *om de tijdreeks op te delen en zo een idee te krijgen van de trend en de seizoenschommeling.*
6. *Bereken het exponentieel voortschrijdend gemiddelde* (exponential moving average, EMA) *door gebruik te maken van de methode* HoltWinters. *Maak opnieuw via de methode* forecast *een voorspelling voor 20 periodes. Gebruik als startwaarden* $s_1 = x_1$ *en* α *de door R gegenereerde waarde. Plot het resultaat op een nieuwe grafiek samen met X.*
7. *Doe nu hetzelfde met* $\alpha = 0.1$.
8. *Hoe zien de voorspellingen er nu uit?*
9. *Doe nu hetzelfde met* dubbele *exponentiële afvlakking. Gebruik als startwaarden* $s_1 = x_1$ *en* $b_1 = \frac{x_n - x_1}{n-1}$, $\alpha = 0.05$ *en* $\beta = 0.2$. *Plot het resultaat op de grafiek.*
10. *Gebruik dubbele exponentiële afvlakking om voorspellingen te berekenen voor 20 periodes. Plot de waarden op de grafiek. Is deze techniek beter of slechter dan de vorige voor deze dataset?*
11. *Speel met de waarden voor* α *en* β *en bekijk het resultaat, zowel voor enkele als dubbele exponentiële afvlakking.*
12. *Gebruik de* HoltWinters*-methode zonder trend. M.a.w. we stellen* $\beta = 0$. *Gebruik als startwaarden* $\alpha = 0.05$ *en* $\gamma = 0.9$. *Plot het resultaat op de grafiek.*
13. *Bereken opnieuw voorspellingen voor 20 periodes. Plot de waarden op de grafiek. Is deze techniek beter of slechter dan de vorige voor deze dataset?*
14. *Speel met de waarden voor* α, β *en* γ *en bekijk het resultaat.*
15. *Gebruik de* HoltWinters*-methode met de door R-gegeneerde waarden zonder trend. M.a.w. we stellen* $\beta = 0$. *Plot het resultaat op de grafiek.*
16. *Bereken opnieuw voorspellingen voor 20 periodes maar gebruik nu de methode* predict. *Plot de waarden op de grafiek. Is deze techniek beter of slechter dan de vorige voor deze dataset?*

■

Oefening 8.3. *In bestand* Passagiers2.csv *vind je vanaf januari 1949 tot december 1960 het aantal passagiers van een luchtvaartmaatschappij.*

1. *Bereken het voortschrijdend gemiddelde* (simple moving average) *over de periodes 4 en 12 voor deze data. Gebruik hiervoor de methode* ma. *Maak een lijngrafiek van X,* $MA(4)$ *en* $MA(12)$.
2. *Welke techniek die we eerder gezien hebben (in het deel over beschrijvende statistiek) is ook geschikt om voorspellingen te maken over de waarden van X? Werk dit uit aan de hand van de daarvoor bestemde functie en plot het resultaat in de grafiek.*
3. *Gebruik de methode* forecast *om voorspellingen voor de 10 volgende periodes met elk van voorgaande methoden (dus moving average 4 en 10 en regressie) te maken. Teken deze eveneens op de grafiek. Conclusie?*
4. *Is het gebruik van één van deze technieken interessant om voor deze data voorspellingen te maken?*
5. *Gebruik de methode* decompose *om de tijdreeks op te delen en zo een idee te krijgen van de trend en de seizoenschommeling.*
6. *Bereken het exponentieel voortschrijdend gemiddelde* (exponential moving average, EMA) *door gebruik te maken van de methode* ses *met* $\alpha = 0.2$. *Maak opnieuw via de methode* forecast *een voorspelling voor 20 periodes. Plot het resultaat op een nieuwe grafiek samen met X.*

7. *Doe nu hetzelfde met* $\alpha = 0.6$ *en* $\alpha = 0.89$.
8. *Hoe zien de voorspellingen er nu uit?*
9. *Doe nu hetzelfde met* dubbele *exponentiële afvlakking. Gebruik hiervoor de methode* holt $\alpha = 0.8$ *en* $\beta = 0.2$. *Plot het resultaat op de grafiek.*
10. *Gebruik dubbele exponentiële afvlakking om voorspellingen te berekenen voor 20 periodes. Plot de waarden op de grafiek. Is deze techniek beter of slechter dan de vorige voor deze dataset?*
11. *Gebruik in de methode de optie* $exponential = TRUE$. *Teken het resultaat. Wat is het verschil?*
12. *Gebruik de* hw*-methode met de door R gegeneerde waarden. Plot het resultaat op de grafiek.*
13. *Bereken opnieuw een aantal voorspellingen via de methode* predict. *Plot de waarden op de grafiek. Is deze techniek beter of slechter dan de vorige voor deze dataset?*
14. *Speel met de waarden voor* α, β *en* γ *en bekijk het resultaat.*

■

Appendices

A. Logistische regressie

A.1 Inleiding

In dit onderzoek gaan we een andere vorm van verband zoeken tussen variabelen waarbij de afhankelijke variabele twee waarden kan aannemen.

Voorbeeld A.1. *Stel dat je wil nagaan of het student al dan niet zal slagen voor het examen onderzoekstechnieken. We zijn dus geïnteresseerd in de voorspelling (door onafhankelijke variabelen) van de kans dat een student in de categorie 'examen slagen' of in de categorie 'niet slagen' valt.* ■

In bovenstaand voorbeeld zal een 'gewone' lineaire regressie analyse algemeen wel de juiste richting van de β-coëfficiënten opleveren. Maar de schatting is niet helemaal correct, omdat enkele belangrijke regressie assumpties geschonden worden, zoals de normaliteitsassumptie en de assumptie van homoscedasticiteit. Het grootste probleem is evenwel dat de door lineaire regressie voorspelde kansen groter kunnen zijn dan 1 en kleiner dan 0 en dat is niet te interpreteren.

Bij logistische regressie gaan we werken met kansverhoudingen. In voorbeeld A.1 hebben we een kansverdeling dat een student wel slaagt $(y = 1)$ met kans p gedeeld door de kans om niet te slagen $(y = 0)$ met kans $q = 1 - p$:

$$\text{verhouding} = \frac{p}{1-p}$$

We wensen dat de waarden van de verhouding gaan van $-\infty$ tot ∞. Daarom gaan we de natuurlijke logaritme nemen van de verhouding. Om de functie te tekenen van de logaritmische functie kan je onderstaande code gebruiken.

```
1 curve(log(x),0,20,n=50)
```

Als we de onafhankelijke variabelen X_1, X_2 $\ldots X_n$ noemen,dan ziet het logistische model er in formulevorm als volgt uit:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

We kunnen het kansmodel ook herschrijven (afzonderen van de p):

$$p = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}} \tag{A.1}$$

We kunnen het kansmodel dan ook herschrijven (afzonderen van de $(1-p)$):

$$1-p = \frac{1}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}$$

Aan deze formules is af te lezen dat de kansen p en $1-p$ bij elkaar opgeteld gelijk zijn aan één. Verder is te zien dat de kansen p en $1-p$ afhankelijk zijn van de variabelen $X_1, X_2 \cdots X_n$, maar dat deze afhankelijkheid niet lineair is. Een logistische regressielijn ziet er dus niet als een rechte lijn uit, maar als een S-vormige curve. (TODO: hier zou een tekening moeten komen van de sigmoïde functie).

Bij logistische regressie gaan we dus op zoek naar goede waarden voor $\beta_0 \cdots \beta_n$ die het model zo goed mogelijk beschrijven zodat we ook voorspellingen kunnen doen. Dit kan in R makkelijk door de methode `glm`.

Om de logistische functie te tekenen kunnen we gebruik maken van onderstaande code (twee parameters).

```
sigmoide <- function (alfa, beta, x){
  z <- alfa + beta * x;
  e <- exp(z);
  return(e/(1+e));
}
```

A.1.1 Intuïtie rond de oplossingsmethode

Om de waarden van $\beta_0, \beta_1 \cdots \beta_n$ te bepalen gaan we deze keer niet gebruik maken van de kleinste kwadratenmethode (zie sectie 6.3), maar wel van een meer algemene methode : maximum likelihood methode . Hierbij proberen we waarden voor de β_i te vinden die ervoor zorgen dat in de trainingsdataset (de dataset die we gebruiken om de parameters β_i te bepalen) de elementen die een label 1 krijgen zo goed mogelijk benaderd worden door 1 in vergelijking A.1 en de elementen die een label 0 krijgen zo goed mogelijk benaderd worden door 0. Dit doen we door volgende vergelijking te maximaliseren.

$$\Pi_{i:y_i=1} p(x_i) \Pi_{j=y_j=0} (1-p(x_j)) \tag{A.2}$$

De oplossingsmethode wordt geïmplementeerd in R en is buiten de scope van deze cursus. We refereren de geïnteresseerde lezer naar **Hastie2009** voor meer informatie rond deze methode.

A.1.2 Performantie van het model

Er zijn een aantal performantiematen die in rekening moeten gebracht worden wanneer aan logistische regressie gedaan wordt.

Akaike Information Criteria

Dit is een statistiek die wat overeenkomt met R^2 vanuit sectie 6.4.2. Het geeft aan hoe goed de opgenomen variabelen in ons model het resultaat weergeven en we wensen die AIC zo laag mogelijk te houden. Het geeft ons dus een inkijk in het gebruik van de variabelen en zorgt ervoor dat we niet te veel variabelen in ons model opnemen.

De waarde van de AIC is op zichzelf niet van belang, maar wordt vooral gebruikt wanneer de verschillende modellen wilt vergelijken: dan neem je best het model met de laagste AIC.

Null deviance

Dit is een indicatie hoe goed het model de data fit waarbij alleen gebruik gemaakt wordt van de intercept. Hoe lager deze waarde hoe beter.

Residual deviance

Dit is een indicatie hoe goed het model de data fit, waarbij de onafhankelijke variabelen toegevoegd zijn. Hier geldt ook, hoe lager deze waarde hoe beter.

Bij de output in R krijg je bovenstaande waarden. Waar je als onderzoeker vooral geïnteresseerd in bent is een lage AIC en een een significante daling van the Null Deviance naar de Residual deviance.

A.2 Logistische regressie in R

We gaan het voorbeeld nemen dan in Kaggle [1] gegeven wordt. Het bevat de informatie rond de mensen die de reis van de titanic ondernomen hebben en het overleefd hebben of niet. De analyse komt uit het blog artikel **michy**, maar is wat aangepast aangezien niet alle conclusies in dit artikel kloppen.

A.2.1 Data cleaning

Importeer de data, en zorg ervoor dat de juiste types voor de juiste variabelen gekozen zijn (Sex is bijvoorbeeld een `factor` variabele)

[1] https://www.kaggle.com/c/titanic/data

We gaan de data opruimen en kijken welke parameters er in het model kunnen zitten. We gaan dit na door te kijken welke parameters in de dataset niet voldoende aanwezig zijn.

```
sapply(train,function(x) sum(is.na(x)))
sapply(train, function(x) length(unique(x)))
missmap(train, main = "Missing values vs observed")
```

Hierbij zien we dat de variabelen `cabin` te weinig waarden bevat. Ook `tickets` laten we vallen aangezien dit weinig invloed zal hebben. We nemen bijgevolg een subset van de data en gaan hiermee aan de slag.

```
data <- subset(train,select=c(2,3,5,6,7,8,10,12))
```

We moeten ervoor zorgen dat de andere data elementen die er te kort zijn zinvol ingevuld worden. Je hebt hier verschillende methodieken voor. Je kan vervangen door:

- het gemiddelde
- de mediaan
- de modus
- een elementen uit een bepaalde distributie

We gaan voor de optie om de `NA` elementen te vervangen door hun gemiddelde.

```
data$Age[is.na(data$Age)] <- mean(data$Age,na.rm=T)
```

Voor de nominale en ordinale variabelen kunnen we kijken hoe ze gecodeerd worden door R.

```
contrasts(data$Sex)
```

A.2.2 Fitten van de data in R

We gaan de data opsplitsen in een trainingsset en een testset. We gaan hiervoor de library `caTools` gebruiken.

```
install.packages('caTools')
library(caTools)
```

Nu kunnen we het model laten opbouwen door R.

```
model <- glm(Survived ~.,family=binomial(link='logit'),data=
    train)
summary(model)
```

Je krijgt volgende output na het uitvoeren van dit commando:

Coefficient De schatting voor de coëfficiënt in het model
Std. error De standard errors op de coëffciënt.
z-statistic Dit komt overeen met de $\frac{\beta_i}{SE(\beta_i)}$.
P-value De p-waarde geassocieerd met de null-hypothese van de coëffciënt.

Deze laatste twee getallen hebben wat verduidelijking nodig. Voor elke β_i wordt een null-hypothese H_0^i opgesteld. Deze stelt dat

$$p(X_i) = \frac{e^{\beta_0+\cdots\beta_{i-1}+\beta_{i+1}+\cdots\beta_n}}{1+e^{\beta_0+\cdots\beta_{i-1}+\beta_{i+1}+\cdots\beta_n}}$$

wat eigenlijk neerkomt dat het model niet afhangt van X_i. Wanneer de $|z|$ groot genoeg is en bijgevolg de p-waarde klein is mag de H_0^i verworpen worden en kunnen we stellen dat X_i wel degelijk van belang is in het model.

Om een betrouwbaarheidsinterval te bouwen rond de geschatte parameter β_i kan je gewoonweg volgende formule gebruiken:

$$\beta_i + z_i \times SE(\beta_i)$$

Als output krijgen we:

```
Coefficients:
(Intercept)      Pclass2      Pclass3    Sexfemale          Age
      SibSp0       Parch1         Fare    EmbarkedC
   EmbarkedQ
1.36178      -0.96344      -2.19975       2.67728      -0.04503
   -0.49519      0.08984      0.00105       0.37631      0.68404

Degrees of Freedom: 666 Total (i.e. Null);  657 Residual
Null Deviance:      887.4
Residual Deviance: 582.5   AIC: 602.5
```

En met `summary` van het model bekomen we volgende output:

```
Call:
glm(formula = Survived ~ ., family = binomial(link = "logit"),
data = dresstrain)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.4971  -0.6377  -0.3730   0.6240   2.5854

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.361775   0.495174   2.750  0.00596 **
Pclass2     -0.963440   0.339019  -2.842  0.00449 **
Pclass3     -2.199753   0.335770  -6.551  5.7e-11 ***
Sexfemale    2.677281   0.224722  11.914  < 2e-16 ***
Age         -0.045028   0.009096  -4.951  7.4e-07 ***
SibSp0      -0.495189   0.252907  -1.958  0.05023 .
Parch1       0.089835   0.304233   0.295  0.76778
Fare         0.001050   0.002259   0.465  0.64190
EmbarkedC    0.376309   0.277878   1.354  0.17566
EmbarkedQ    0.684037   0.364547   1.876  0.06060 .
---
```

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 887.35  on 666  degrees of freedom
Residual deviance: 582.46  on 657  degrees of freedom
AIC: 602.46

Number of Fisher Scoring iterations: 5
```

Hieruit kunnen we volgende dingen zeggen:

- SibSp, Parch1, Fare, EmbarkedC en EmbarkedQ zijn niet statisch significant.
- We zien dat Sexfemale erg significant. De positieve coëfficiënt voor sexFemale toont aan dat vrouw zijn ervoor zorgt dat je meer kans hebt op overleven.

Bij de `anova` wordt getoond wat het effect is van een variabele een per een toe te voegen aan het model. **TODO: dit nog eens deftig interpreteren.**

De volledige code kan je hier nog eens bekijken:

```
# Creeer de modus function.
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}


# Geeft ons het aantal elementen die NA zijn per variabele
sapply(train,function(x) sum(is.na(x)))
# Geeft ons het aantal unique elementen
sapply(train, function(x) length(unique(x)))
library(Amelia)

#Geeft ons een visuele voorstelling van de missing waarden
missmap(train, main = "Missing values vs observed")

#We nemen een subset van de kolommen om mee verder te werken
data <- subset(train,select=c(2,3,5,6,7,8,10,12))

#Missende elementen opvullen met het gemiddelde
data$Age[is.na(data$Age)] <- mean(data$Age,na.rm=T)

sibSpMode <- getmode(data$SibSp)
parchMode <- getmode(data$Parch)


#Modus gebruiken voor resterende variabelen
data$SibSp[is.na(data$SibSp)] <- sibSpMode
data$Parch[is.na(data$Parch)] <- parchMode
```

```
data$Parch[is.na(data$Embarked)] <- embarkedMode

#Rijen verwijderen waarbij embarked niet juist is
data <- data[!is.na(data$Embarked),]
rownames(data) <- NULL

missmap(data, main = "Missing elements from dataset")

#We hebben nu een probere dataset

library(caTools)

set.seed(88)
split <- sample.split(data$Survived, SplitRatio = 0.90)

dresstrain <- subset(data, split == TRUE)
dresstest <- subset(data, split == FALSE)

#Train the data
model <- glm(Survived ~Pclass+Sex+Age+SibSp+Parch+Fare+Embarked,
    family=binomial(link='logit'),data=dresstrain)
summary(model)
anova(model)

fitted.results <- predict(model,newdata=dresstest,type='response
    ')
fitted.results <- ifelse(fitted.results > 0.5,1,0)

misClasificError <- mean(fitted.results != dresstest$Survived)
print(paste('Accuracy',1-misClasificError))
library(ROCR)
library(ROCR)
p <- predict(model, newdata=dresstest, type="response")
pr <- prediction(p, dresstest$Survived)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)

auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc

#Odds ratio berekenen in R
require(MASS)
exp(cbind(coef(model), confint(model)))
```

A.3 Oefeningen

Oefening A.1.

- *Beschouw de dataset* `Smarker` *van de package* `ISLR`*. Deze dataset bestaat uit het rendement voor de S & P 500 aandelenindex over 1250 dagen, van begin 2001 tot eind 2005. Voor elke datum hebben we het retourneer percentage opgenomen voor elk van de vijf vorige handelsdagen (Lag1 t.e.m. Lag5). We hebben ook het Volume opgenomen (het aantal verhandelde aandelen) en het percentage van vandaag. Daarnaast hebben we ook opgenomen of de markt daalde of steeg.*
- *Schrijf de algemene statistieken uit van de verschillende variabelen.*
- *Probeer eens een plot te maken die aanduid of het volume stijgt of daalt met de jaren.*
- *We gaan proberen een logistisch model op te stellen dat het stijgen of dalen in functie van lag1 t.e.m. lag5 en volume uitzet. Gebruik hiervoor het commando glm.*
- *Analyseer de coëfficiënten. Wat kan je erover zeggen?*
- *Kijk nu eens hoe goed het model de dataset zelf voorspelt. Dit kan je doen door aan het predict commando geen dataset mee te geven.*
- *Zet de voorspelde probabiliteit om in juiste labels (≥ 0.5 up)*
- *Creeër een matrix die de vals positieven en ware positieven e.a. uitzet t.o.v. elkaar. Gebruik hiervoor de methode table.*
- *Wat kom je hier nu voor uit?*

■

Oefening A.2.

- *Beschouw dezelfde dataset als hierboven, maar train nu de dataset met de elementen van voor 2005 en gebruik als testset de elementen boven 2005. Wat kom je nu uit?*
- *Probeer nu het model aan te passen door de juiste variabelen te kiezen om mee te nemen in het model.*
- *Wanneer je tevreden bent met het model, probeer dan een voorspelling te doen van een willekeurige dataset.*

■

B. Notatie

Notatie	**Betekenis**
$X = \{x_1, x_2, \ldots, x_n\}$	Een stochastische variabele X met n waarnemingen x_i (voor $i : 1 \ldots n$)
N	De populatieomvang
n	De steekproefgrootte
μ (mu)	Het gemiddelde (ook: verwachtingswaarde) over heel de *populatie*.
$\bar{x}$	Het gemiddelde over de *steekproef*
σ (sigma)	De standaardafwijking over heel de populatie
σ^2 (sigma)	De variantie over heel de populatie
s	De standaardafwijking van de steekproef
s^2	De variantie van de steekproef
$X \sim Nor(\mu, \sigma)$	De variabele X is *normaal verdeeld* met gemiddelde μ en standaardafwijking σ
$Z \sim Nor(0, 1)$	Z is een variabele met een kansverdeling die de *standaardnormaalverdeling* volgt, dus met gemiddelde 0 en standaardafwijking 1
$M \sim Nor(\mu_{\bar{x}}, \sigma_{\bar{x}})$	De kansverdeling van het steekproefgemiddelde (cfr. de centrale limietstelling, Sectie 4.5)
$\mu_{\bar{x}}$	De verwachtingswaarde bij de kansverdeling van het steekproefgemiddelde
$\sigma_{\bar{x}}$	De standaardafwijking bij de kansverdeling van het steekproefgemiddelde
α	Een significantieniveau (voor een statistische toets)
$1 - \alpha$	Een betrouwbaarheidsniveau (voor een betrouwbaarheidsinterval)