FACULTEIT PSYCHOLOGIE EN
PEDAGOGISCHE WETENSCHAPPEN

STATISTIEK I

Prof. dr. Jan De Neve

Academiejaar 2020–2021

UNIVERSITEIT
GENT

Voorwoord

De cursus 'Statistiek I' is een inleiding tot de statistiek. De statistische technieken die behandeld worden, vormen de basis voor verschillende vervolgcursussen (zoals o.a. Statistiek II, Psychometrie, Onderzoeksmethoden I en II, Methoden in de psychologie en Toegepaste data-analyse). Kennis in statistiek kan pas verworven worden nadat het duidelijk is *waarom* statistiek noodzakelijk is. Daarom zijn delen van de cursus opgebouwd rond studies uit de gedragswetenschappen die gebruik maken van statistische analyses om een antwoord te geven op een onderzoeksvraag. Dit is ook hoe statistiek in de praktijk aan bod komt: ze vormt een ondersteunende, maar essentiële schakel in kwantitatief onderzoek.

De hoofdstukken in deze syllabus bouwen verder op elkaar. Voordat je een hoofdstuk kan aanvatten, moet je dus eerst alle voorgaande hoofdstukken bestuderen. Deze syllabus bevat enkele wijzigingen ten opzichte van de syllabus van vorig academiejaar. Nieuw dit jaar is de wijze waarop de cursus zal aangeboden worden: in plaats van klassieke hoorcolleges en werkcolleges zal de cursus worden aangeboden via blended learning. Meer informatie hierover kan je vinden op Ufora.

Het studeren van enkel deze syllabus is zeker niet voldoende om de inhoud van de cursus te beheersen: de *oefeningen* vormen ook een essentieel onderdeel. Statistische kennis verwerf je door afwisselend theorie te studeren en oefeningen te maken: eerst studeer je een hoofdstuk, vervolgens maak je enkele oefeningen om dan terug hetzelfde hoofdstuk te bestuderen. Deze cyclus herhaal je tot je zowel de theorie als de oefeningen *begrijpt*. Het begrijpen van de cursus is het ultieme doel, meer dan het memoriseren of reproduceren.

De syllabus bevat vele figuren en tabellen en die staan vaak in de buurt van de tekst waar ze besproken worden. Soms kan het echter gebeuren dat ze enkele pagina's verder staan. Begrippen die belangrijk zijn om het vervolg van de syllabus te kunnen begrijpen, worden met een uitroepingsteken aangeduid:

! Dit is een belangrijk begrip om het vervolg van de syllabus te begrijpen.

Sommige figuren maken gebruik van kleuren en zijn zo opgesteld dat ze ook duidelijk zouden moeten zijn in zwart-wit. Via Ufora is het mogelijk om gratis de digitale versie van de syllabus te raadplegen indien sommige figuren toch onduidelijk zouden zijn.

De syllabus werd zorgvuldig nagelezen, maar bevat zonder twijfel nog een aantal fouten. Deze mogen steeds worden gemeld door te mailen naar `Jan.DeNeve@UGent.be`.

Inhoudsopgave

Hoofdstuk 1

Inleiding

Statistiek is de wetenschap van het verzamelen en interpreteren van data[a], meestal met een duidelijke onderzoeksvraag voor ogen. Het is echter ook een wetenschap waar veel misvattingen rond bestaan. We bespreken enkele van die misvattingen in paragraaf 1.1. In paragraaf 1.2 gaan we dieper in op de betekenis van statistiek en we illustreren dit aan de hand van een voorbeeld. In paragraaf 1.3 bespreken we enkele eigenschappen van *variabelen* en introduceren we *schaalfamilies*. In paragraaf 1.4 illustreren we de software R en RStudio en geven we duiding bij de drie delen waaruit de cursus is opgebouwd.

1.1 Enkele misvattingen

1.1.1 "Met statistiek kan je alles bewijzen"

De uitspraak "Met statistiek kan je alles bewijzen" is volkomen incorrect. Echter, door statistische analyses verkeerdelijk toe te passen, kan je wel de impressie wekken dat je kan aantonen wat je wil en dit kan nefaste gevolgen hebben.

We illustreren dit kort aan de hand van een rechtszaak waar statistisch bewijs een belangrijk onderdeel vormde van het proces.

Op 4 september 2001 wordt aangifte gedaan van een mogelijks onnatuurlijke dood van een baby in het Juliana Kinderziekenhuis (Den Haag) waar verpleegster Lucia de B.

[a]Met data bedoelen we een verzameling van gegevens, vaak afkomstig van personen vb. de leeftijd, het IQ, het gewicht, etc.

werkzaam is. Haar aanwezigheid bij dit overlijden wordt als verdacht beschouwd en men onderzoekt ook voorgaande sterfgevallen waar Lucia de B. bij aanwezig was. In totaal zijn er 9 onverwachte en medisch onverklaarbare overlijdens. Naar aanleiding hiervan arresteert de politie haar op 13 december 2001. Op 24 maart 2003 wordt ze veroordeeld tot een levenslange gevangenisstraf voor de moord op 4 patiënten en poging tot moord op 3 patiënten. Op 18 juni 2004 wordt ze in hoger beroep zelfs schuldig verklaard aan 7 moorden en 3 pogingen tot moord en wordt ze levenslang veroordeeld samen met een terbeschikkingstelling[b].

Tijdens het proces was er een gebrek aan harde bewijzen: Lucia de B. werd nooit op heterdaad betrapt en ze heeft altijd ontkend schuldig te zijn. Omwille van dit gebrek aan harde bewijzen werd er onder andere gebruik gemaakt van statistisch bewijs. Een professor strafrecht liet hierover optekenen in een uitzending:

In de Lucia de B.-zaak is het statistisch bewijs ontzettend belangrijk geweest. Ik zie niet hoe men zonder dat bewijs tot een veroordeling zou zijn gekomen.

Dit statistisch bewijs bestond uit de berekening en interpretatie van de volgende kans:

De kans dat een verpleegkundige, werkzaam op de drie ziekenhuisafdelingen, bij toeval bij zoveel van de onverklaarbare overlijdensgevallen en reanimaties op elk van de drie afdelingen aanwezig was, is één op 342 miljoen.

Men interpreteerde deze uitspraak als volgt: het is zeer onwaarschijnlijk dat Lucia de B. per toeval aanwezig was bij deze verdachte overlijdens en bijgevolg moet ze wel schuldig zijn. Deze kans werd berekend op basis van meerdere gegevens, waaronder die in Tabel 1.1. Verschillende wetenschappers waren het echter niet eens met de berekening en interpretatie van deze kans omdat o.a. foute gegevens en incorrecte statistische analyses waren gebruikt. Volgens statisticus Richard Gill ligt de geschatte kans ergens tussen één op 48 (2%) en één op 5 (20%), wat dus veel groter is dan één op 342 miljoen.

Als gevolg van deze verkeerde statistische analyses en andere onregelmatigheden besliste de Hoge Raad der Nederlanden op 7 oktober 2008 dat de zaak moest heropend worden wegens twijfel aan het bewijs. Terecht, zo bleek: op 14 april 2010 wordt Lucia de B. formeel vrijgesproken. Bij de uiteindelijke uitspraak bleek zelfs dat de onverwachte en medisch onverklaarbare overlijdens *niet* het gevolg waren van een misdrijf. In totaal verbleef Lucia de B. meer dan zes jaar onschuldig in de gevangenis als gevolg van,

[b]Omdat de combinatie levenslang en terbeschikkingstelling niet kan, besliste de Hoge Raad in Cassatie op 14 maart 2006 om deze terbeschikkingstelling te laten vallen.

	Aantal diensten met incident	Aantal diensten zonder incident	Totaal
Lucia dienst	9	133	142
Lucia geen dienst	0	887	887
Totaal	9	1020	1029

Tabel 1.1: In deze tabel staat het overzicht van diensten en incidenten in het Juliana Kinderziekenhuis tussen 1 oktober 2000 en 9 september 2001. Met incidenten wordt het overlijden van een patiënt, een reanimatie of een verdachte gebeurtenis bedoeld. Er zijn in totaal drie tabellen voor verschillende periodes en verschillende werkplekken van Lucia de B. gebruikt. Bron: Smeets, I. (2007, 20 juni). Toch statistiek in de zaak Lucia de B. Geraadpleegd op 2 september, 2015, van `http://www.kennislink.nl/publicaties/toch-statistiek-in-de-zaak-lucia-de-b`

onder andere, foutieve statistische analyses. Voor meer informatie verwijzen we naar Buchanan (2007) en Meester et al. (2006) (deze laatste referentie bevat uitgebreide informatie over de statistische analyses).

Dit voorbeeld geeft aan dat je met foutieve statistische analyses volledig verkeerde besluiten kan trekken. Gelukkig hebben niet alle foutieve analyses dergelijke drastische gevolgen.

1.1.2 "Statistiek is nutteloos voor de gedragswetenschappen"

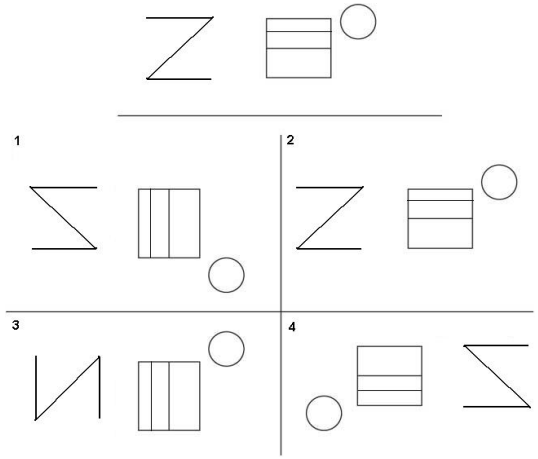
De gedragswetenschappen zijn gebaseerd op empirisch onderzoek. Met empirisch wordt bedoeld dat kennis wordt verworven door middel van observaties en metingen. Deze observaties en metingen geven aanleiding tot data (gegevens). De data worden[c] op hun beurt vaak statistisch geanalyseerd om zo meer inzicht te verwerven in de processen die bestudeerd worden. In dit opzicht vormt statistiek een belangrijke schakel in de totstandkoming en het begrijpen van vele inzichten binnen de gedragswetenschappen. We geven enkele voorbeelden om dit te illustreren.

Het visueel geheugen onderzoeken

Om leermoeilijkheden bij kinderen te identificeren, kan het visueel geheugen onderzocht worden via de Benton Visual Retention Test. Elk kind krijgt een afbeelding gelijkaardig aan de bovenste in Figuur 1.1 gedurende tien seconden te zien. Daarna moet hij/zij deze afbeelding trachten te selecteren uit de vier mogelijkheden onderaan. Dit wordt een aantal keer herhaald met verschillende afbeeldingen. Vervolgens zal de onderzoeker het

[c]Het woord 'data' is een meervoudsvorm.

aantal correcte antwoorden scoren per kind. Deze scores vormen dan de basis voor de statistische analyses die dan kunnen onthullen welke kinderen leermoeilijkheden hebben.



Figuur 1.1: Een voorbeeldafbeelding (bovenaan) en de meerkeuze mogelijkheden die gebruikt worden in de Benton Visual Retention Test.

Impliciete voorkeuren bepalen

De Impliciete Associatie Test (IAT) is een reactietaak die gebruikt wordt om impliciete voorkeuren te meten. Op een computer moeten personen verschillende taken uitvoeren en bij elke taak wordt de reactietijd geregistreerd. Deze tijden bevatten informatie over de impliciete voorkeur en via statistische analyses kunnen we deze informatie beter begrijpen. In Hoofdstukken 2 en 3 bespreken we dit in detail en illustreren we hoe we impliciete raciale voorkeuren kunnen onderzoeken en hoe statistiek ons hierbij kan helpen.

Intelligentie en hersengrootte bestuderen

Zijn mensen met grotere hersenen slimmer? Om deze vraag te beantwoorden, zullen onderzoekers de samenhang bestuderen tussen de hersengrootte en het IQ.

Hiervoor zal men via intelligentietesten het IQ meten van 40 personen. Vervolgens zal men van elke persoon de hersengrootte bepalen via een MRI-hersenscan (Magnetic Resonance Imaging). Tabel 1.2 geeft de IQ-scores en hersengroottes weer voor 5 van de 40 personen. Via statistische methodes kunnen we meer inzicht krijgen in de samenhang tussen het IQ en de hersengrootte. In Hoofdstuk 4 wordt dit experiment en de statistische analyse in detail besproken.

IQ-score	Hersengrootte
132	816.93
150	1001.12
123	1038.44
129	965.35
132	951.54

Tabel 1.2: Illustratie van de IQ-scores en de hersengrootte (uitgedrukt in 1000 pixel) voor 5 van de 40 personen.

Voorgaande voorbeelden geven aan dat statistiek kan bijdragen tot het beter begrijpen van informatie bekomen door het verzamelen van gegevens (d.m.v. een Benton Visual Retention Test, een IAT, het meten van de hersenen, etc.). Het is daarom niet verwonderlijk dat vele van de inzichten binnen de psychologie en de pedagogische wetenschappen tot stand zijn gekomen door gebruik te maken van statistiek. Een basiskennis statistiek is dan ook noodzakelijk om deze inzichten te begrijpen en te kaderen.

1.1.3 "Statistiek is enkel wiskunde"

Het is correct dat wiskunde een belangrijke rol speelt binnen de statistiek, maar naast wiskundige kennis zijn ook andere vaardigheden vereist. Statistiek is dus zeker niet enkel wiskunde. In een cursus statistiek komen typisch volgende aspecten aan bod:

- Wiskunde: De methodes die we zullen bestuderen zijn geschreven in de taal van de wiskunde. Vooral (basis) algebra en kansrekening spelen een prominente rol. Bij deze cursus ligt de nadruk echter op het correct analyseren van data en zullen we de wiskunde tot een minimum beperken.
- Software: Om al die wiskundige methodes op data te kunnen toepassen, hebben we statistische software nodig. Voorbeelden hiervan zijn R, S-Plus, SPSS en SAS. Wij zullen gebruik maken van het gratis softwarepakket R en de omgeving RStudio.
- Interpretatie en besluitvorming: Statistiek wordt vaak aangewend om op basis van gegevens een onderzoeksvraag te beantwoorden. Eenmaal alle gegevens verwerkt zijn aan de hand van statistische software moeten de resultaten geïnterpreteerd worden om zo een besluit te kunnen formuleren rond de onderzoeksvraag.

1.2 De betekenis van statistiek

Dan rest ons de vraag: *wat is statistiek?* Er bestaat hier geen eenduidig antwoord op, maar volgende zin vat de essentie samen (Davidian and Louis, 2012):

Statistiek is de wetenschap van het leren uit data en van het meten, controleren en communiceren van onzekerheid.

We illustreren dit aan de hand van een voorbeeld rond intelligentie en we voeren enkele belangrijke begrippen in. Erna wordt er een overzicht gegeven van deze begrippen.

1.2.1 Een voorbeeld rond intelligentie

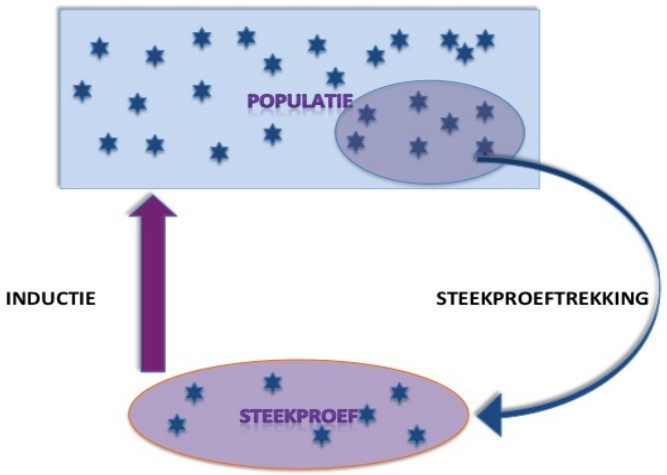
Binnen het kader van een masterproef is een student geïnteresseerd in het gemiddeld IQ van studenten eerste bachelor Psychologie en eerste bachelor Pedagogische Wetenschappen. Het IQ wordt hier de *variabele* genoemd: het IQ *varieert* (verschilt, wijzigt) immers van persoon tot persoon.

In totaal zijn er 1000 studenten. Deze verzameling van studenten wordt de *populatie* genoemd. Figuur 1.2 geeft een schematische weergave. De studenten die hier de populatie uitmaken worden de *elementen* van de populatie genoemd.

Om de IQ-scores te bekomen, zal de masterstudent de Wechsler Adult Intelligence Scale gebruiken. Echter, de masterstudent kan onmogelijk alle studenten deze IQ-test laten afnemen omwille van verschillende redenen: de groep is te groot, sommige studenten zijn moeilijk te bereiken omdat ze niet naar de les komen, etc. Het is dus zo goed als onmogelijk om de IQ-scores te bekomen van alle elementen in de populatie. Een oplossing bestaat erin om slechts een deel van de populatie te ondervragen. Dit wordt de *steekproef* genoemd.

Op basis van de inschrijvingslijst kunnen bijvoorbeeld 50 namen op willekeurige wijze gekozen worden[d]. Via email kunnen deze studenten gecontacteerd worden met de vraag deel te nemen aan het onderzoek. Voor de eenvoud gaan we ervan uit dat alle 50 studenten bereid zijn om mee te werken. Deze 50 studenten vormen dan de steekproef. De steekproef is dus een deelverzameling van de populatie. Eenmaal alle 50 studenten uit de steekproef de IQ-test hebben afgelegd, kan de masterstudent het gemiddelde IQ

[d]We kunnen 50 willekeurige namen bekomen door bijvoorbeeld alle namen te nummeren en vervolgens 50 willekeurige getallen te genereren met een computer.



Figuur 1.2: Schematische weergave van de populatie, steekproef *en de* elementen *(de elementen zijn weergegeven door sterren). De pijl rechts stelt de selectie voor van de steekproef uit de populatie. De pijl links stelt het* inductieprincipe *voor: op basis van de gegevens uit de steekproef zullen we besluiten formuleren over de populatie.*

berekenen. Dit gemiddelde heeft echter betrekking op deze 50 studenten en niet de 1000 studenten uit de populatie. Het is echter net deze populatie waarin de masterstudent geïnteresseerd is en uitspraken over wil doen.

Nu komt de echte meerwaarde van statistiek naar voor: via statistische technieken is het mogelijk om de conclusies op basis van de steekproef te veralgemenen naar de populatie. Op basis van de IQ-scores van slechts 50 studenten kan de masterstudent toch een uitspraak doen over het gemiddelde IQ van de 1000 studenten. Dit wordt *inductie* genoemd. De wijze waarop de steekproef wordt bekomen is zeer belangrijk: een goede keuze is noodzakelijk om de conclusies te kunnen veralgemenen. In Hoofdstuk 2 komen we hierop terug.

Omdat we niet over de IQ-scores beschikken van alle 1000 studenten uit de populatie, zullen we echter nooit met volledige zekerheid een uitspraak kunnen doen over de populatie[e]. We zullen bijgevolg *onzekerheid* moeten toelaten in onze uitspraken. In Deel III komen we hier uitgebreid op terug.

[e]We kunnen enkel met 100% zekerheid een uitspraak doen over het gemiddelde IQ van de populatie, indien we alle 1000 studenten een IQ-test laten afleggen.

1.2.2 Enkele definities

Op een meer formele wijze kunnen we volgende begrippen omschrijven:

- **Populatie**: de volledige verzameling van objecten of personen waarover informatie wordt gewenst[f].
- **Elementen**: de individuele leden van de populatie (de objecten of personen).
- **Steekproef**: een deelverzameling van de populatie die feitelijk zal onderzocht worden om informatie te bekomen. Vaak is het onmogelijk om de volledige populatie te bestuderen (omdat het te duur is, omdat de populatie te groot is, etc.), vandaar dat men dan een steekproef uit de populatie zal nemen voor verder onderzoek[g].
- **Variabele**: een eigenschap die bij de elementen van de populatie of steekproef varieert. Vaak worden er bij een steekproef verschillende variabelen gemeten[h].
- **Data**: de verzameling van gegevens die wordt bekomen door de variabelen te meten.
- De **verdeling** van een variabele geeft aan welke waarden worden aangenomen en hoe vaak.
- **Inductie**: uitgaande van het bijzondere het algemene besluiten[i]. Bij inductie proberen we op basis van een aantal waarnemingen tot een algemeen besluit te komen.

In het voorbeeld rond intelligentie werd slechts 1 variabele gemeten, namelijk het IQ. Het zou evenwel mogelijk zijn om naast IQ nog meerdere variabelen te meten: vb. geslacht, leeftijd, etc.

---

[f] Statistiek kent haar oorsprong in het verzamelen en analyseren van staatsgegevens (vandaar de *Stat* in Statistiek). Deze staatsgegevens hadden betrekking op een bevolking, vandaar de benaming *populatie*.

[g] Er zijn verschillende etymologische verklaringen voor het woord *steekproef*. Eén verklaring stelt dat een steekproef afkomstig is van de kaasmarkt waar een keurmeester de kwaliteit van de kazen moest nagaan door ze te steken met een kaasboor en vervolgens te proeven.

[h] Met variabelen *meten* bedoelen we het vaststellen van de waarden van de variabelen. Dus meten hoeft hier niet noodzakelijk een numerieke betekenis te hebben, het kan bijvoorbeeld het 'meten' zijn van het geslacht met de waarden 'man' of 'vrouw' of 'andere'.

[i] Het tegenovergestelde van inductie is *deductie*: uitgaande van het algemene het bijzondere besluiten. Een voorbeeld van deductie: alle mensen zijn sterfelijk (het algemene), Socrates is een mens (het bijzondere), dus Socrates is sterfelijk (conclusie).

1.3 Eigenschappen van variabelen

Op basis van hoe ze gemeten zijn, kan je verschillende soorten variabelen onderscheiden. In de volgende subparagrafen gaan we hier dieper op in.

1.3.1 Schaalfamilies

Getallen kunnen een verschillende betekenis hebben in verschillende situaties. Het getal 12 bijvoorbeeld: het kan een leeftijd voorstellen van een kind dat 12 jaar oud is, het rugnummer van een voetballer, de twaalfde plaats in de marathon van Londen, de temperatuur in graden Celsius, etc. Afhankelijk van de context verschilt de betekenis van het getal 12. De informatie die het getal 12 bevat, hangt af van de *meetschaal* die werd gebruikt om het getal te bekomen. In deze cursus onderscheiden we vier meetschalen: nominaal, ordinaal, interval en ratio[j]. Om de informatie in data ten volle te kunnen benutten, moeten we het verschil tussen deze meetschalen goed begrijpen. De wiskundige bewerkingen die we kunnen toepassen op data, zullen ook afhangen van de meetschaal.

Nominale schaal

Bij de nominale schaal worden de waarden van de variabele gebruikt voor identificatie zonder dat ze een hoeveelheid aanduiden.

Enkele voorbeelden:

- *Variabele*: Geslacht. *Waarden*: man, vrouw of andere.
- *Variabele*: Land van herkomst. *Waarden*: België, Frankrijk, Spanje, etc.
- *Variabele*: Politieke voorkeur in de VS. *Waarden*: Democraat, Republikein of Andere.

Bij deze voorbeelden is het duidelijk dat de waarde van de variabele geen numerieke betekenis heeft, het zijn zelfs geen getallen (man, vrouw, België, etc.). Er bestaan

[j]Niettegenstaande deze vier meetschalen algemeen worden aanvaard, staan ze soms toch ter discussie. Er zijn ook onderzoekers die de meetschalen onderverdelen in bijvoorbeeld 5 of 10 schalen. In 'Statistiek II' wordt hier dieper op ingegaan.

echter ook nominale variabelen die numerieke waarden aannemen zonder dat het getal zelf een specifieke betekenis heeft (het wordt louter gebruikt ter identificatie).

Enkele voorbeelden:

- *Variabele*: Rekeningnummer. *Waarden*: 37 0000 0000 2828 (Oxfam), 73 0000 0000 6060 (Artsen Zonder Grenzen), etc.
- *Variabele*: Rugnummer in het voetbal. *Waarden*: 1, 2, 10, 31, etc.
- *Variabele*: Het nummer van de tram in Gent. *Waarden*: 1, 2 en 4.

Het kan ook zijn dat een onderzoeker de waarden van een nominale variabele, zoals geslacht, numeriek codeert. Een 1 kan dan bijvoorbeeld overeenkomen met de mannen, een 2 met de vrouwen en een 3 met personen van het derde geslacht. Deze getallen hebben op zich geen betekenis, maar ze kunnen de verwerking van de gegevens eenvoudiger maken. Het coderen van variabelen wijzigt de schaal niet: als geslacht de waarden 'man', 'vrouw' of 'andere' aanneemt of '1', '2' of '3', het blijft een variabele gemeten op nominale schaal. In plaats van '1', '2' of '3' kunnen ook andere getallen gekozen worden. Het maakt niet uit welke numerieke waarden er worden gebruikt. Bij een nominale schaal worden dus alleen nummers toegekend aan de waarden van een variabele om ze van elkaar te kunnen onderscheiden.

Variabelen gemeten op nominale schaal worden ook nominale variabelen genoemd. Nominaal drukt uit dat de waarden van de variabelen slechts 'namen' zijn.

Ordinale schaal

De ordinale schaal erft alle eigenschappen van de nominale schaal samen met een extra eigenschap: de waarden duiden een volgorde aan. Behalve om een volgorde weer te geven, is de waarde van de variabele niet van belang.

Enkele voorbeelden:

- *Variabele*: Uitslag wedstrijd. *Waarden*: goud, zilver of brons.
- *Variabele*: Officiersgraad. *Waarden*: Onderluitenant, Luitenant, Kapitein, Kapitein-commandant, etc.
- *Variabele*: Mate van instemming met een bepaalde uitspraak. *Waarden*: volledig mee oneens, mee oneens, neutraal, mee eens, volledig mee eens.

Net zoals bij de nominale schaal kunnen we de waarden van de variabelen vervangen door getallen. Bij de uitslag van een wedstrijd bijvoorbeeld: 1 = goud, 2 = zilver en 3 = brons. Een kleinere waarde wil zeggen dat de atleet beter is. Het verschil tussen de getallen speelt echter geen rol: het verschil in prestatie tussen de eerste en tweede is niet noodzakelijk gelijk aan het verschil in prestatie tussen de tweede en derde.

Een andere keuze van numerieke waarden is: 1 = goud, 10 = zilver en 100 = brons. Nog steeds drukt een kleiner getal een beter resultaat uit. Opnieuw kunnen dus verschillende coderingen gebruikt worden, maar met de restrictie dat de codering de volgorde moet behouden. Een voorbeeld van een incorrecte codering zou zijn: 1 = goud, 3 = zilver en 2 = brons.

Variabelen gemeten op ordinale schaal worden ook ordinale variabelen genoemd. Ordinaal drukt uit dat de waarden van de variabele geordend kunnen worden. Bij een nominale variabele is dit niet het geval: we kunnen niet zeggen dat voetballers met een kleiner rugnummer slechter zijn dan die met een groter nummer (of omgekeerd).

Intervalschaal

De intervalschaal erft alle eigenschappen van de ordinale schaal met de extra eigenschap dat verschillen tussen waarden een betekenis hebben. Er is echter geen absoluut nulpunt[k].

Enkele voorbeelden:

- *Variabele*: Temperatuur in graden Celsius. *Waarden*: 0, 10, -30, etc.
- *Variabele*: IQ[l]. *Waarden*: 96, 100, 130, etc.

Niettegenstaande de temperatuur de waarde 0 kan aannemen, is het geen absoluut nulpunt omdat 0 °C niet willen zeggen dat er geen temperatuur aanwezig is. Het is de aanduiding van een bepaalde temperatuur[m]. Een andere manier om te zien dat het nulpunt bij temperatuur niet absoluut is, is door een andere meeteenheid te gebruiken. Nul graden Celsius komt overeen met 32 graden Fahrenheit. Dit geeft aan dat het nulpunt hier relatief is: ze hangt af van de gebruikte meeteenheid (Celsius of Fahrenheit).

[k] Met absoluut nulpunt bedoelen we een getal dat aangeeft dat niets van de variabele aanwezig is.

[l] Het IQ is een variabele waarover nog steeds discussie bestaat tot welke klasse ze behoort. Sommige wetenschappers argumenteren dat ze ordinaal is. We gaan hier echter niet dieper op in en zullen, zoals de meeste handboeken, veronderstellen dat het IQ gemeten is op intervalschaal.

[m] Nul graden Celsius duidt de temperatuur aan waarbij water bevriest bij een luchtdruk van 1 bar.

Variabelen gemeten op intervalschaal worden ook intervalvariabelen genoemd. De benaming interval drukt uit dat gelijke verschillen op de meetschaal (intervallen) duiden op gelijke verschillen in de variabele. Een stijging van 10 °C naar 20 °C is evenveel als een stijging van 20 °C naar 30 °C in termen van het warmteverschil. Dit blijft behouden bij omzetting naar Fahrenheit (10 °C = 50 °F, 20 °C = 68 °F en 30 °C = 86 °F): het verschil tussen 68 °F en 50 °F is gelijk aan het verschil tussen 86 °F en 68 °F. Bij een ordinale variabele is dit niet het geval: bv. het verschil in uitslag tussen de eerste (gouden medaille) en de tweede (zilveren medaille) hoeft niet gelijk te zijn aan het verschil in uitslag tussen de tweede en de derde (bronzen medaille).

Ratioschaal

De ratioschaal erft alle eigenschappen van de intervalschaal én heeft daarbij ook een absoluut nulpunt.

Enkele voorbeelden:

- *Variabele*: Lengte in cm. *Waarden*: 0, 1, 354, etc.
- *Variabele*: Geldbedrag in euro. *Waarden*: 0, 5, 2400, etc.
- *Variabele*: Reactietijd tussen stimulus en gedrag in seconden. *Waarden*: 0, 5, 2, 58, etc.

Hier is het nulpunt absoluut: als je 0 euro hebt, wil dit zeggen dat je geen geld hebt. Een lengte van 0 centimeter blijft 0, ook als je de schaal omzet naar meter. Een reactietijd van 0 seconden wil zeggen dat er geen tijd zit tussen stimulus en gedrag.

Variabelen gemeten op ratioschaal worden ook ratiovariabelen genoemd. De benaming ratio drukt uit dat verhoudingen (ratio's) een betekenis hebben: 10 euro is bijvoorbeeld dubbel zoveel als 5 euro. Bij intervalvariabelen zijn verhoudingen niet zinnig: een temperatuur van 10 °C is niet dubbel zo warm 5 °C omdat deze uitspraak bij een omzetting naar Fahrenheit (5 °C = 41 °F en 10 °C = 50 °F) niet meer opgaat: 50 is niet dubbel zoveel als 41.

Voor dit vak (en in de praktijk) is het vooral van belang om het onderscheid te kennen tussen de nominale, ordinale en interval/ratioschaal. Het onderscheid tussen de interval- en ratioschaal is minder van belang.

Als we terugkeren naar de verschillende situaties van het nummer 12, kunnen we ze als volgt indelen in de verschillende schaalfamilies:

- Nominaal: rugnummer van een voetballer.
- Ordinaal: de twaalfde plaats in de marathon van Londen.
- Interval: de temperatuur in graden Celsius.
- Ratio: de leeftijd van een kind dat 12 jaar oud is.

Niettegenstaande het getal 12 steeds dezelfde blijft, verschilt de informatie die het getal weergeeft van situatie tot situatie.

1.3.2 Discrete en continue variabelen

Naast de vier schaalfamilies kunnen we ook een onderscheid maken tussen continue en discrete variabelen.

Continue variabelen kunnen tussenwaarden aannemen. Met tussenwaarde bedoelen we dat er tussen elke twee willekeurige waarden een derde waarde ligt. Dit impliceert dat er tussen twee waarden (theoretisch gezien) oneindig veel andere waarden kunnen liggen. We zeggen ook dat variabelen gemeten op continue schaal continu variëren.

Enkele voorbeelden:

- Lengte in cm: tussen 2 en 3 cm liggen nog vele andere waarden: vb. 2.5 cm, 2.34 cm, 2.323548 cm, etc[n].
- Temperatuur in °C: tussen twee willekeurige temperaturen ligt altijd een derde. Tussen 25.8 °C en 25.9 °C ligt bijvoorbeeld 25.85 °C, 25.82147 °C, etc.
- De tijd in seconden: tussen twee willekeurige tijdstippen ligt altijd een derde tijdstip.

Bij *discrete variabelen* bestaan steeds twee waarden waar geen derde waarde kan tussen liggen. Dit impliceert dat de variabele maar een eindig aantal waarden kan aannemen[o].

Enkele voorbeelden:

[n] Voor kommagetallen wordt een punt gebruikt, dus 2.5 lees je als 2 komma 5.

[o] Strikt genomen kan het aantal ook aftelbaar oneindig zijn, maar dit behoort niet tot de inhoud van deze cursus.

- Het aantal kinderen. Tussen 0 en 1 kind ligt geen derde waarde: het is niet mogelijk om 1.5 kinderen te hebben.
- Het aantal keer dat 'munt' wordt geworpen bij 4 worpen met een geldstuk. Het is niet mogelijk om 1.5 keer munt te werpen, ze kan slechts de waarden 0, 1, 2, 3, of 4 aannemen.
- Het aantal volgers op Twitter. Barack Obama kan bijvoorbeeld 63 379 938 of 63 379 939 volgers hebben, maar geen 63 379 938.5.

Het aantal volgers op Twitter is theoretisch gezien een discrete variabele, maar omdat deze variabele zéér veel verschillende waarden kan aannemen, wordt er ook gezegd dat ze *bijna-continu* is. In de praktijk zullen we variabelen die bijna-continu zijn, als continu beschouwen. In Deel II zal dit duidelijk worden.

1.4 Informatie over de syllabus

1.4.1 Het softwarepakket R

Zoals aangeven in paragraaf 1.1.3 speelt statistische software een belangrijke rol bij het analyseren van data. In deze cursus wordt gebruikt gemaakt van het softwarepakket R (`https://www.r-project.org`). Via commando's kunnen we in R data analyseren en om dit op een overzichtelijk manier te doen, zullen we gebruik maken van de omgeving RStudio. RStudio is gratis beschikbaar via `https://www.rstudio.com` of via Athena. Figuur 1.3 toont het RStudio icoon op Athena[p]. Voor meer informatie rond het gebruik van Athena, zie `https://helpdesk.ugent.be/athena/`. Athena zelf kan je vinden via `https://athena.ugent.be`.

RStudio heeft een beperkte grafische gebruikersomgeving en wordt voornamelijk gebruikt door R-code in te typen en uit te voeren. Figuur 1.4 toont via verschillende screenshots hoe je dit kan doen.

De tekst in deze cursus is afgewisseld met R-code. Deze R-code is gekopieerd van de console in RStudio[q] (zie Figuur 1.4 rechtsonder) en wordt als volgt weergegeven:

[p]Op Athena zullen er vaak meerdere versies van RStudio beschikbaar zijn en je kan dan de meest recente versie gebruiken. Doorheen het academiejaar zal RStudio op regelmatige basis worden geüpdatet, dus het kan zijn dat de versienummers van de screenshots in de cursus verschillen van diegene op Athena. Dit heeft geen invloed op het functioneren van de R-code.

[q]De console in RStudio is het venster linksonder: ze toont de code en de output.

Figuur 1.3: Icoon van RStudio op Athena (het versienummer kan verschillen)

```
> 1 + 1
```

```
[1] 2
```

Zowel het prompt-symbool '`>`' als de tweede lijn '`[1] 2`' maken deel uit van de output, terwijl '`1 + 1`' het R-commando is. Indien je deze code zelf wil uitvoeren, moet je enkel '`1 + 1`' invoeren (zie Figuur 1.4).

Alle stukken code zijn ook beschikbaar via Minerva dus in principe hoef je geen code uit de cursus over te typen. Het overtypen wordt wel aangeraden om zo vertrouwd te geraken met de code.

De output is

```
[1] 2
```

Hier kunnen we '`[1]`' voorlopig negeren en '`2`' geeft de oplossing van de som: $1+1=2$.

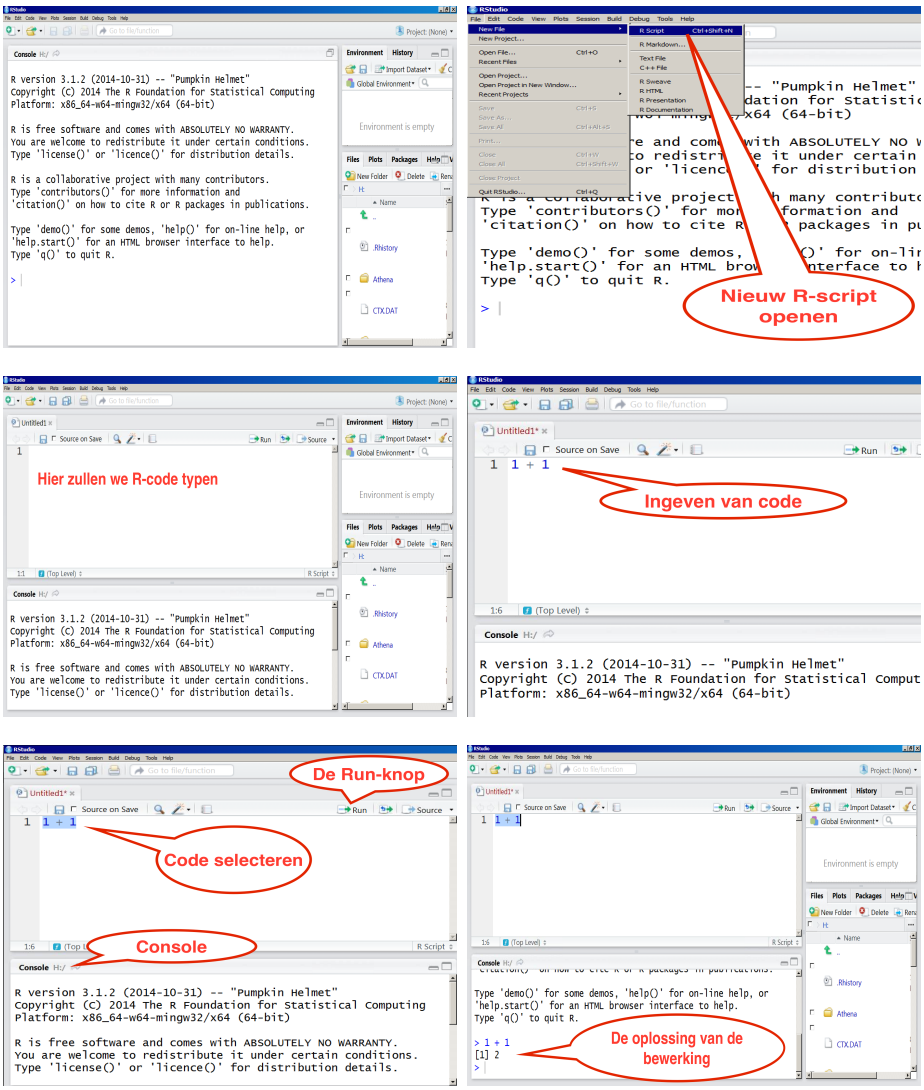
Opgelet: je moet de code *exact* overtypen en R is hoofdlettergevoelig. Foutmeldingen worden vaak veroorzaakt door het incorrect overtypen van code.

In wat volgt geven we een beknopte uitleg over hoe de code werkt. Het is de bedoeling dat je zelf vertrouwd geraakt met R en RStudio door tijdens het studeren van de cursus RStudio te openen en de R-code uit te voeren. In het begin zal dit moeilijk zijn, maar naargelang je meer code zelf uitvoert, zal je er beter mee vertrouwd geraken.

Zoals in het voorbeeld, kunnen we in R eenvoudige bewerkingen uitvoeren:

```
> 1 + 1
```

```
[1] 2
```



Figuur 1.4: Verschillende stappen die je na het openen van RStudio kan uitvoeren om een R-script te openen, code in te geven en code uit te voeren. Via de pdf-versie van de syllabus die beschikbaar is op Ufora kan je inzoomen op de verschillende figuren. Stap 1 (linksboven): screenshot van RStudio na het openen (sommige details, zoals de folders in het venster rechtsonder kunnen verschillen). Stap 2 (rechtsboven): Via File - New File - R Script kan je een nieuw R-script openen. Het is in dit bestand dat we code zullen typen. Stap 3 (linksmidden): screenshot van RStudio na het openen van een R-script. Stap 4 (rechtsmidden): ingeven van code. Stap 5 (linksonder): door de code te selecteren en Ctrl en Enter te drukken (of via de 'Run'-knop rechtsboven in het R-script) kan je code uitvoeren: RStudio zal automatisch de code copy-pasten in het onderste venster (de console) en uitvoeren. Stap 6 (rechtsonder): onderaan zie je in de console de oplossing van de bewerking: 1 + 1 = 2.

We kunnen de oplossing van deze som ook opslaan en een naam geven, we gebruiken hiervoor de toekenningsoperator '`<-`':

```
> som <- 1+1
```

Nu hebben we de oplossing van de som 1+1 bewaard en de naam `som` gegeven. Door dit nu uit te voeren, zien we dat `som` inderdaad het getal 2 bevat:

```
> som
```

```
[1] 2
```

Doorheen de cursus zullen we in R geregeld gebruik maken van *vectoren*. Deze vectoren zijn sequenties van elementen: deze elementen kunnen bijvoorbeeld getallen of woorden zijn. Later zal het duidelijk worden waarom we dit nodig hebben.

In R kan je een vector ingeven via `c()` met tussen de haakjes de getallen of woorden van de vector, gescheiden door komma's. Voor een vector van woorden moeten aanhalingstekens gebruikt worden. Bijvoorbeeld, een vector met de namen Billie, Edith en Dani kan je als volgt ingeven:

```
> c("Billie", "Edith", "Dani")
```

```
[1] "Billie" "Edith"  "Dani"
```

Indien de elementen van de vector getallen zijn, zijn er geen aanhalingstekens nodig. De rij 1, 2, 3, 4, bijvoorbeeld, wordt bekomen door:

```
> c(1, 2, 3, 4)
```

```
[1] 1 2 3 4
```

We kunnen deze rij opslaan en een naam geven, bijvoorbeeld `Rij`:

```
> Rij <- c(1, 2, 3, 4)
> Rij
```

```
[1] 1 2 3 4
```

Op deze rij kan je nu bewerkingen uitvoeren. De functie `max()` geeft bijvoorbeeld de grootste waarde weer:

```
> max(Rij)
```

```
[1] 4
```

Indien je meer informatie wil bekomen over hoe je de functie `max()` kan gebruiken, kan je de hulppagina openen door een vraagteken voor de functie te plaatsen en deze R-code uit te voeren (dus de code selecteren en Ctrl + Enter drukken):

```
> ?max
```

Zie Figuur 1.5: binnen RStudio is de hulppagina automatisch geopend (onderaan rechts). Deze bevat informatie over hoe je de functie kan gebruiken en geeft ook enkele voorbeelden weer.

Vergeet niet om je R-script op te slaan. Dit kan via File - Save (of Save As...). RStudio zal automatisch de extensie `.R` toevoegen aan het bestand. Het is aangeraden om al je R-files te bewaren in een folder op je persoonlijke H-schijf. Als je werkt via Athena zal je automatisch je files kunnen opslaan op je H-schijf. Voor meer informatie over de persoonlijke H-schijf verwijzen we naar de UGent helpdesk: `https://helpdesk.ugent.be/netdisk/schijfruimte.php`.

Tot slot geven we enkele basisbewerkingen mee in R.

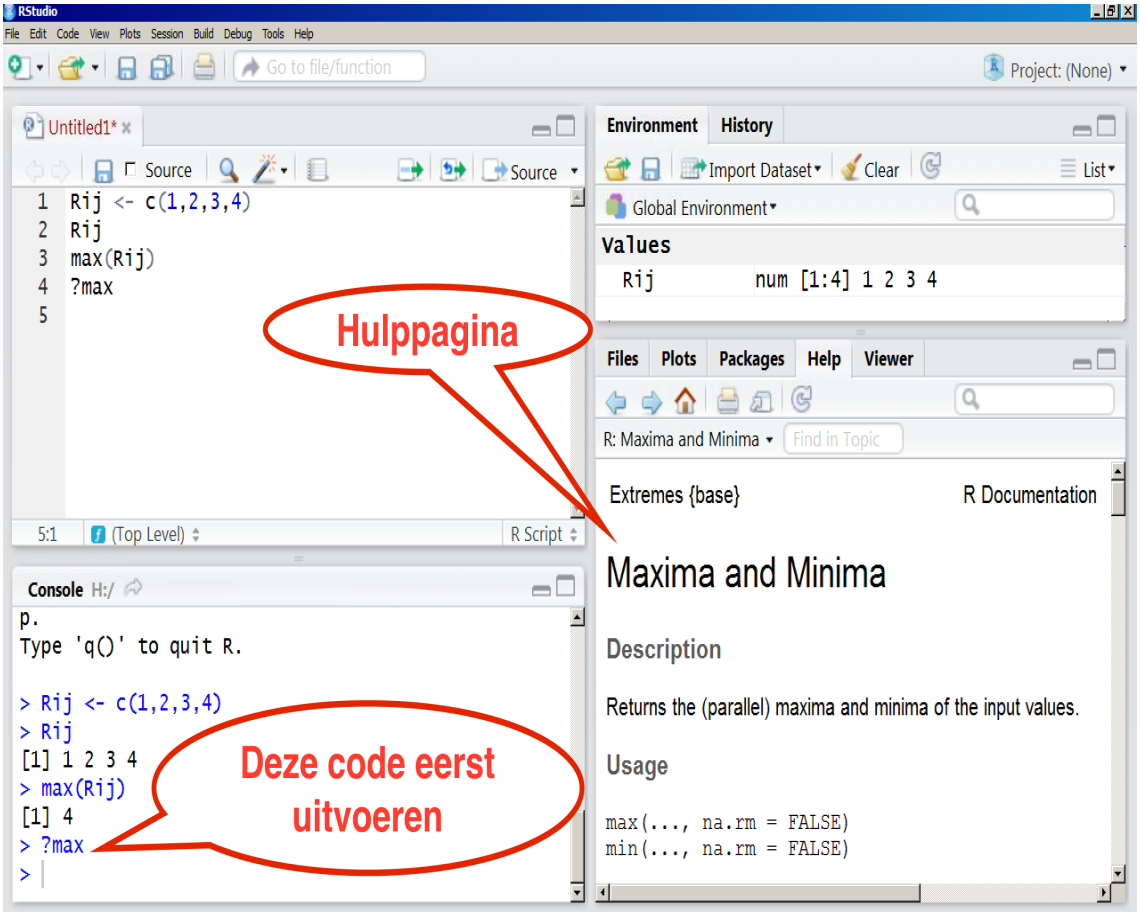
Optellen `+`:

```
> 2+3
```

```
[1] 5
```

Aftrekken `-`:

```
> 3-2
```



Figuur 1.5: De hulppagina kan je openen door een vraagteken voor de functie ***max*** *te plaatsen en deze code uit te voeren (uitvoeren wil zeggen: de code selecteren en dan op Ctrl + Enter drukken).*

```
[1] 1
```

Vermenigvuldigen `*`:

```
> 2*3
```

```
[1] 6
```

Delen `/`:

```
> 2/3
```

```
[1] 0.6666667
```

Haakjes '`( )`' kunnen ook gebruikt worden:

```
> 3*(5+4)
```

```
[1] 27
```

```
> (3*5)+4
```

```
[1] 19
```

Zonder het gebruik van haakjes worden de klassieke voorrangsregels gebruikt (eerst vermenigvuldigen en delen, dan pas optellen en aftrekken):

```
> 3*5+4
```

```
[1] 19
```

Voorlopig gaan we niet verder in op het gebruik van R en RStudio. Bij verschillende paragrafen in de cursus wordt R-code gegeven en op die manier zullen we stap voor stap leren hoe R/RStudio werkt en hoe het gebruikt kan worden bij statistische analyses.

Opgelet: het is *niet* de bedoeling dat je R-code tracht te *memoriseren*. Je moet echter *wel* in staat zijn om de *output* van de R-code te *begrijpen*.

1.4.2 Indeling van de syllabus

De syllabus is opgedeeld in 3 stukken:

- Deel I: Beschrijvende statistiek. In dit deel worden statistische methodes bestudeerd die gebruikt zullen worden om informatie uit de steekproef te visualiseren en samen te vatten.
- Deel II: Kansrekening. Het veralgemenen van besluiten uit de steekproef naar de populatie (het inductieprincipe) resulteert in onzekerheid: de kans dat de uitspraak over de populatie correct is, is kleiner dan 100%. Om deze kansen te kunnen berekenen, zullen we beroep doen op kansrekening.
- Deel III: Inductieve statistiek. Eenmaal we kansrekening begrijpen, kunnen we ze aanwenden om besluiten uit de steekproef te veralgemenen naar de populatie. Hier passen we kansrekening toe binnen de statistiek. In dit deel zien we slechts enkele voorbeelden van inductieve statistiek. In de cursus Statistiek II komt dit meer uitgebreid aan bod.

Deel I

Beschrijvende statistiek

Hoofdstuk 2

Visualiseren van data

In dit hoofdstuk staat het *visualiseren* van data centraal. Dit kan door middel van tabellen en figuren. Om dit te illustreren maken we gebruik van data uit een studie naar raciale voorkeur. In paragraaf 2.1 geven we details over dit onderzoek (de onderzoeksvraag, de experimenten, etc.). Dit dient voornamelijk om de context van het onderzoek te schetsen en heeft weinig met statistische methoden te maken. Om de meerwaarde van statistiek te kunnen inzien, is het echter noodzakelijk om deze context te geven.

In paragrafen 2.2-2.5 komen verschillende statistische technieken aan bod en deze vormen de essentie van dit hoofdstuk.

Tot slot illustreren we in paragraaf 2.6 hoe we deze technieken kunnen gebruiken om de resultaten van de studie naar raciale voorkeur beter te begrijpen.

2.1 Onderzoek naar raciale voorkeur

2.1.1 De onderzoeksvraag

Bij een sociaalpsychologisch onderzoek in de Verenigde Staten is men geïnteresseerd in raciale voorkeur. Meer specifiek wensen onderzoekers volgende onderzoeksvraag te beantwoorden:

Verkiezen mensen hun eigen ras?

Het ras wordt beperkt tot twee mogelijkheden: zwarten (Afro-Amerikanen) of blanken (Euro-Amerikanen). Gegeven deze beperkingen kan de onderzoeksvraag onderverdeeld worden in twee concrete onderzoeksvragen:

Verkiezen blanke Amerikanen blanken boven zwarten?
Verkiezen zwarte Amerikanen zwarten boven blanken?

Door middel van het opzetten van computerexperimenten, het verzamelen van data en het statistisch analyseren van deze data, zullen we trachten een antwoord te geven op deze vragen.

2.1.2 De populatie en de steekproef

De populatie omvat alle zwarte en blanke Amerikanen. Volgens recente schattingen zijn er ongeveer 220 miljoen Euro-Amerikanen en 38 miljoen Afro-Amerikanen. Het is duidelijk dat we niet van al deze personen hun raciale voorkeur kunnen onderzoeken. Daarom zullen we een steekproef nemen uit de populatie. De onderzoekers hebben de mogelijkheid om in totaal 90 personen op te nemen in hun steekproef.

De wijze waarop men de steekproef zal nemen, is belangrijk. De steekproef moet immers een afspiegeling zijn van de totale populatie. Stel dat er in de steekproef enkel mannen voorkomen, dan is het duidelijk dat deze steekproef geen goede afspiegeling is: ongeveer de helft van de Amerikanen in de populatie zijn vrouwen. Een steekproef die een goede afspiegeling is van de populatie, wordt ook *representatief* genoemd.

Het nemen van een steekproef uit een grote populatie als deze is niet eenvoudig. Stel dat de onderzoekers 90 van hun collega's aan de universiteit vragen om deel te nemen aan het experiment, zou dit dan een representatieve steekproef zijn? Neen, de collega's aan de universiteit zijn immers allemaal hoogopgeleid terwijl de populatie ook laagopgeleiden bevat. Ook zullen de meeste collega's van de onderzoekers in dezelfde staat wonen, terwijl er in de populatie 50 staten zijn.

Het is duidelijk dat de keuze van een steekproef vaak complex is en dat een goede keuze belangrijk is om een uitspraak te kunnen doen over de populatie. Er bestaan verschillende technieken om (theoretisch) een representatieve steekproef te bekomen. Een voorbeeld is de *aselecte steekproef*: men selecteert op willekeurige wijze 90 Amerikanen uit de populatie. Dit is echter niet haalbaar in de praktijk: hoe kan je 90 personen willekeurig selecteren uit een populatie van miljoenen?, hoe zal je er voor zorgen dat die 90 Amerikanen deelnemen aan het experiment?, etc. Het bekomen van een steekproef

uit een populatie zal bijgevolg een compromis zijn tussen de theoretisch beste keuze en een praktisch haalbare keuze.

Voor het onderzoek naar raciale voorkeur wordt een computerexperiment ontworpen (zie paragraaf 2.1.3 voor meer details) en men zal als volgt een steekproef bekomen: het computerexperiment wordt online beschikbaar gesteld en via advertenties rekruteert men 90 deelnemers. Omdat de experimenten thuis kunnen worden afgelegd, is het mogelijk om deelnemers te bekomen verspreid over de Verenigde Staten. De steekproef die op deze manier wordt bekomen, is niet noodzakelijk representatief: het is nog steeds mogelijk dat bijvoorbeeld enkel mannen deelnemen aan het experiment. Ook nemen er enkel mensen deel die internettoegang hebben en dit kan resulteren in een steekproef van vooral jongeren. Daarom zullen de onderzoekers bij het computerexperiment extra gegevens opvragen van de deelnemer, zoals geslacht en leeftijd. Op die manier kunnen ze dan zelf evalueren hoe representatief de steekproef is[a].

2.1.3 Het IAT-experiment

De Impliciete Associatie Test (IAT) is een populaire methode om impliciete voorkeur te meten. Alvorens we hier in detail op ingaan, schetsen we aan de hand van twee andere experimenten de meerwaarde van onderzoek naar impliciete voorkeur.

Om raciale voorkeur te onderzoeken, kan men een *gevoelsthermometer* gebruiken: de deelnemers moeten aangeven hoe warm of koud ze zich voelen tegenover zowel zwarten als blanken. Ze kunnen antwoorden op een 11-punten schaal waar 0 staat voor koude gevoelens, 5 voor neutrale gevoelens en 10 voor warme gevoelens. Op basis hiervan kunnen de onderzoekers nagaan of de scores systematisch hoger zijn voor een ras. De gegevens bekomen op deze wijze kan je echter niet altijd vertrouwen, het kan bijvoorbeeld zijn dat bepaalde personen sociaal-wenselijk zullen antwoorden: niettegenstaande iemand een afkeer heeft van bijvoorbeeld zwarten, kan het zijn dat hij/zij een score van 5 geeft (dus een neutraal gevoel) omdat dit sociaal meer aanvaard is.

Om dit sociaal-wenselijk gedrag te vermijden kan men een tweede experiment opzetten waar elke deelnemer moet kiezen of hij/zij naast een zwarte persoon of een blanke persoon gaat zitten. De deelnemer weet zelf niet dat de raciale voorkeur wordt onderzocht. Men kan dan vervolgens het aantal keer tellen dat een deelnemer ervoor gekozen heeft om naast een zwarte persoon te zitten. Deze strategie geeft aanleiding tot een *impliciete maat* van de raciale voorkeur. Voor meer informatie verwijzen we naar De Houwer

[a]Naast geslacht en leeftijd zouden de onderzoekers nog meer variabelen (vb. diploma, beroep, etc.) moeten meten om een uitspraak te kunnen doen over hoe representatief de steekproef is. We gaan hier echter niet dieper op in.

(2006).

De IAT (zoals reeds kort besproken in paragraaf 1.1.2) is een populair instrument om indirect impliciete voorkeuren te meten. De experimenten kunnen uitgevoerd worden op een computer en dat maakt ze praktisch zeer interessant: het opzetten van de experimenten vereist bijvoorbeeld weinig logistieke voorbereiding en, zoals reeds eerder aangegeven, men kan via het internet experimenten afnemen van personen die zich geografisch op een andere plaats bevinden.

De IAT om de raciale voorkeur te onderzoeken, wordt de Zwart-Blank IAT genoemd. **De eenvoudigste manier om het computerexperiment te begrijpen, is door ze *zelf* eens uit te voeren.**

Je kan een IAT zelf uitvoeren door volgende stappen volgen (dit duurt 5 tot 10 minuten):

- Ga naar de website `https://implicit.harvard.edu/implicit/belgium/`
- Klik op *Ga naar de Demonstratietest.*
- Klik op *Dit wetende, wens ik verder te gaan.*
- Selecteer *Ras (Zwart-Blank IAT)* .
- Volg de instructies op het scherm.

De volgende alinea's vatten de essentie van een IAT samen (als je de test zelf hebt uitgevoerd, kan je dit stuk diagonaal lezen). Niettegenstaande de informatie in deze paragraaf belangrijk is om de context van het onderzoek naar raciale voorkeur te begrijpen, maakt ze **geen** deel uit van de leerstof die op het examen kan gevraagd worden.

Opzet IAT experiment

Tijdens het experiment worden er foto's getoond van 6 blanken (3 vrouwen en 3 mannen) en 6 zwarten (3 vrouwen en 3 mannen). Deze foto's zijn afkomstig van spelers en coaches van de 1998-1999 NBA (National Basketball Association) en WNBA (Women's National Basketball Association) basketbalcompetitie. De deelnemers moeten de foto's onderbrengen in twee categorieën: *Blanke mensen* en *Zwarte mensen.* Naast deze foto's worden ook woorden getoond die kunnen ondergebracht worden in twee categorieën:

- *Goed*: Vreugde, Liefde, Vrede, Mooi, Genot, Glorieus, Lachen of Gelukkig.

- *Slecht*: Pijn, Erg, Horror, Naar, Slecht, Pijnlijk, Mislukking of Gekwetst.

Via het toetsenbord moeten de deelnemers de foto's en de woorden in de juiste categorieën onderbrengen. De deelnemers weten op voorhand welke woorden bij welke categorie horen en het is ook duidelijk welke foto's blanke personen voorstellen en welke zwarte personen. Er wordt verwacht dat ze weinig tot geen fouten zullen maken. Indien er toch fouten worden gemaakt, kunnen ze gecorrigeerd worden. De deelnemers moeten zo snel mogelijk de foto's of woorden onderbrengen in de juiste categorie en bij elke toetsindruk wordt de reactietijd gemeten (in milliseconden).

De deelnemer krijgt duidelijke instructies over hoe het experiment werkt en krijgt verschillende voorbeeldopdrachten om te oefenen. Vervolgens worden opdrachten gegeven gelijkaardig aan die in Figuur 2.1. Hier horen zwarte mensen en slechte woorden samen enerzijds en blanke mensen en goede woorden anderzijds. Dit worden de *congruente opdrachten* genoemd. Het label 'congruent' (congruent = overeenstemmend) wordt gebruikt in overeenstemming met bestaande racistische stereotypes bij een bepaalde groep. Hoe sneller de reactietijden van de congruente opdrachten, hoe meer men het impliciet eens is met die stereotypes.

Tijdens de congruente opdrachten worden woorden afgewisseld met foto's. Als er een woord van de categorie Goed of een foto van een blanke persoon tevoorschijn komt, moet men de `I`-toets indrukken. Indien een woord van de categorie Slecht of een foto van een zwarte persoon tevoorschijn komt, moet men de `E`-toets indrukken. Voor de 4 opdrachten bij Figuur 2.1 moet men dus de volgende toetsen indrukken: `I` (afbeelding linksboven), `E` (rechtsboven), `I` (linksonder) en `E` (rechtsonder).

Na 40 van deze opdrachten te hebben doorlopen (waar foto's van blanken worden afgewisseld met foto's van zwarten en goede woorden worden afgewisseld met slechte), worden de categorieën links- en rechtsboven het scherm gewisseld: zwarte mensen en goede woorden horen nu samen enerzijds en blanke mensen en slechte woorden horen samen anderzijds, zie Figuur 2.2. Dit worden de *incongruente opdrachten* genoemd. Opnieuw moet men de juiste foto's en woorden bij de juiste categorie plaatsen, maar nu horen de blanke mensen bij de `E`-toets en de zwarten bij de `I`-toets (dit is dus verschillend van de eerste reeks opdrachten).

Voor zowel de congruente als de incongruente opdrachten wordt de gemiddelde reactietijd geregistreerd[b].

De experimenten laten toe om de sterkte van de samenhang tussen concepten (blanken

[b]Elke deelnemer heeft in totaal 40 congruente opdrachten afgelegd waardoor we 40 reactietijden hebben. Hiervan wordt het gemiddelde berekend: alle 40 reactietijden worden opgeteld en gedeeld door 40. Analoog voor de 40 incongruente opdrachten.

Figuur 2.1: Enkele voorbeeldopdrachten van de IAT voor de congruente opdrachten

Figuur 2.2: Enkele voorbeeldopdrachten van de IAT voor de incongruente opdrachten

of zwarten) en evaluaties (goed of slecht) te meten. Het idee is dat een antwoord geven eenvoudiger (en dus sneller) is wanneer aansluitende foto's en woorden bij dezelfde toets horen. Indien je dus een impliciete voorkeur zou hebben voor blanken verwacht men dat de reactietijden bij de congruente opdrachten sneller zullen zijn dan bij de incongruente opdrachten. Door de reactietijden van de congruente en incongruente opdrachten te vergelijken, kunnen we de impliciete raciale voorkeur onderzoeken.

Er wordt aan de deelnemers ook gevraagd om via de gevoelsthermometer hun gevoelens aan te geven tegenover zwarten en blanken. Dus via een IAT wordt zowel impliciete als expliciete informatie rond raciale voorkeur bekomen. Van elke deelnemer wordt ook het ras (zwart of blank), het geslacht (man of vrouw) en de leeftijd (in jaren) geregistreerd.

2.1.4 De data

Tabel 2.1 geeft de data weer van 6 van de 90 deelnemers uit de steekproef. In totaal hebben we 7 variabelen gemeten: Geslacht, Leeftijd (in jaar), Ras, Gevoelens t.o.v. zwarten (op basis van de gevoelsthermometer), Gevoelens t.o.v. blanken (op basis van de gevoelsthermometer), de reactietijd van de congruente[c] opdrachten (in milliseconden) en de reactietijd van de incongruente opdrachten[d] (in milliseconden). De variabelen zijn gemeten op volgende schalen:

- Nominaal: Geslacht en Ras.
- Ordinaal: Gevoelens t.o.v. zwarten, Gevoelens t.o.v. blanken.
- Ratio: Leeftijd, Reactietijd congruente opdrachten, Reactietijd incongruente opdrachten.

De data moeten we horizontaal rij per rij lezen: de eerste persoon uit de steekproef is een zwarte vrouw van 40 jaar die aangeeft warme gevoelens te hebben t.o.v. zwarten en blanken met iets warmere gevoelens voor zwarten (een score van 10 voor zwarten en een score van 9 voor blanken). Haar reactietijd voor de congruente opdrachten is 953.050 milliseconden terwijl ze voor de incongruente opdrachten iets sneller is: 849.775 milliseconden. Een ander voorbeeld: de derde persoon uit de steekproef is een blanke man van 22 jaar die weergeeft eerder neutrale gevoelens te hebben t.o.v. blanken en

[c]Bij de congruente opdrachten horen het concept 'Blank' en de evaluatie 'Goed' samen. Bij de incongruente opdrachten horen het concept 'Zwart' en de evaluatie 'Goed' samen. Zie pagina 33 voor meer informatie.

[d]Herinner je dat de reactietijd eigenlijk een gemiddelde is van 40 reactietijden.

Geslacht	Leeftijd	Ras	Gevoelens t.o.v. zwarten	Gevoelens t.o.v. blanken	Reactietijd congruente opdrachten	Reactietijd incongruente opdrachten
vrouw	40	zwart	10	9	953.050	849.775
man	21	zwart	10	7	850.275	672.575
man	22	blank	5	5	447.725	441.500
man	22	blank	8	8	1531.650	1056.150
vrouw	18	blank	8	8	623.500	864.125
vrouw	29	zwart	10	5	1019.650	602.800

Tabel 2.1: De data van de eerste zes personen in de steekproef

zwarten en waarvoor de reactietijd voor de congruente opdrachten 447.725 milliseconden bedraagt en dit voor de incongruente opdrachten bijna gelijk is: 441.500 milliseconden. Analoog kan je de informatie aflezen voor de andere personen.

We illustreren nu hoe we R kunnen gebruiken om data in te lezen en op te vragen.

Illustratie in R

De data zijn online beschikbaar via de url `https://users.ugent.be/~jrdeneve/DataStatistiek1/DataIAT.txt`

We lezen de data in via `read.table()` en we geven ze de naam `DataIAT`:

```
> # Eerst geven we de lange url een korte naam (hier url.IAT)
> url.IAT <- "https://users.ugent.be/~jrdeneve/DataStatistiek1/DataIAT.txt"
> # Nu kunnen we netjes de data inlezen
> DataIAT <- read.table(file = url.IAT)
```

Merk op dat we de R-code van commentaar kunnen voorzien door gebruik te maken van een hashtag # (R zal de code/tekst na een hashtag enkel kopiëren zonder uit te voeren). Deze commentaar kan de code beter leesbaar maken.

De tabel `DataIAT` heeft 90 rijen en 7 kolommen[e]. Dit kunnen we nagaan door de dimensie op te vragen via `dim()`: het eerste getal geeft het aantal rijen weer, het tweede het aantal kolommen:

[e]Rijen worden horizontaal gelezen en kolommen verticaal.

```
> dim(DataIAT)
```

```
[1] 90  7
```

Het cijfer tussen vierkante haakjes `[1]` mag je negeren. De 90 rijen stellen de 90 personen voor uit de steekproef en de 7 kolommen de 7 variabelen die de onderzoekers gemeten hebben.

Via `head()` kunnen we de eerste 6 rijen en alle kolommen opvragen:

```
> head(DataIAT)
```

```
  Geslacht Leeftijd   Ras GevoelensZwart GevoelensBlank TijdCongruent TijdIncongruent
1    vrouw       40 zwart             10              9       953.050         849.775
2      man       21 zwart             10              7       850.275         672.575
3      man       22 blank              5              5       447.725         441.500
4      man       22 blank              8              8      1531.650        1056.150
5    vrouw       18 blank              8              8       623.500         864.125
6    vrouw       29 zwart             10              5      1019.650         602.800
```

De namen van de variabelen kunnen we opvragen via `names()`

```
> names(DataIAT)
```

```
[1] "Geslacht"        "Leeftijd"        "Ras"
[4] "GevoelensZwart"  "GevoelensBlank"  "TijdCongruent"
[7] "TijdIncongruent"
```

Merk op dat deze namen wat verschillen van deze in Tabel 2.1. Binnen statistische softwarepakketten zal men vaak geen spaties gebruiken bij de benaming van variabelen en zal men de namen ook trachten kort te houden.

In de volgende paragrafen illustreren we hoe tabellen en figuren kunnen gebruikt worden om data overzichtelijker te maken. In de laatste paragraaf zullen we deze figuren dan gebruiken om de onderzoeksvraag rond de impliciete raciale voorkeur te beantwoorden.

2.2 Cirkeldiagram

Het cirkeldiagram is een grafische voorstelling die voornamelijk gebruikt wordt voor variabelen van nominaal meetniveau. We illustreren dit aan de hand de variabele 'Geslacht'. Tabel 2.2 geeft het geslacht weer van de 90 personen uit de steekproef. Het zou echter overzichtelijker zijn om het *aantal* mannen en vrouwen weer te geven. Dit worden de *absolute frequenties* genoemd. Tabel 2.3 geeft deze absolute frequenties: er zijn 54 vrouwen en 36 mannen in de steekproef.

Het is duidelijk dat de absolute frequentie mannen (hier 36) opgeteld met de absolute frequentie vrouwen (hier 54) gelijk moet zijn aan de steekproefgrootte (hier 90). Tabel 2.3 wordt ook de *absolute frequentieverdeling* genoemd.

vrouw	man	man	man	vrouw	vrouw	vrouw	man	vrouw	man
man	man	man	vrouw	vrouw	vrouw	man	vrouw	vrouw	man
man	vrouw	man	man	man	vrouw	man	man	vrouw	vrouw
man	man	vrouw	vrouw	vrouw	vrouw	man	vrouw	vrouw	vrouw
vrouw	vrouw	man	vrouw	vrouw	vrouw	vrouw	vrouw	man	man
vrouw	vrouw	vrouw	vrouw	vrouw	man	vrouw	vrouw	man	man
vrouw	man	man	man	vrouw	vrouw	vrouw	vrouw	vrouw	man
vrouw	vrouw	vrouw	vrouw	vrouw	man	man	man	vrouw	man
vrouw	vrouw	vrouw	vrouw	man	vrouw	vrouw	man	vrouw	man

Tabel 2.2: Het geslacht van de personen in de steekproef.

Geslacht	vrouw	man
Absolute frequentie	54	36

Tabel 2.3: Absolute frequentieverdeling van de variabele Geslacht in de steekproef.

! Om deze begrippen formeler te kunnen omschrijven, hebben we wat **notatie** nodig. Een variabele zullen we symbolisch weergeven door een hoofdletter, vaak (maar niet altijd) zal dit de letter X zijn. De waarden die de variabele aanneemt, worden weergeven door kleine letters met cijfers als subscript. Bijvoorbeeld x_1, x_2, x_3, tot en met de laatste waarde x_n waarbij het symbool n staat voor het aantal elementen in de steekproef. Bij ons voorbeeld stelt X de variabele *Geslacht* voor, $n = 90$ (we hebben 90 personen in de steekproef) en op basis van Tabel 2.2 lezen we af (als we horizontaal rij per rij lezen) $x_1 =$ vrouw, $x_2 =$ man, $x_3 =$ man, ..., $x_{90} =$ man[f]. Met een kleine letter zonder cijfer in subscript, hier x, duiden we één

[f]Als er bij een rij volgende 3 puntjes staan '...' wil dit zeggen dat de rij doorloopt zonder dat we

van de mogelijke waarden van de variabele X aan. Bij ons voorbeeld kan x dus de waarde *man* of *vrouw* aannemen.

Met deze notatie zijn we nu in staat om de absolute frequentie en de absolute frequentieverdeling formeel te omschrijven.

! De **absolute frequentie** van x is het aantal keer dat de waarde x in de steekproef voorkomt.

! De **absolute frequentieverdeling** van X is een tabel met twee rijen waar de eerste rij de mogelijke waarden van X weergeeft en de tweede rij de overeenkomstige absolute frequenties. In plaats van een tabel met twee rijen, kan het ook een tabel zijn met twee kolommen.

Aan het voorbeeld van het geslacht zie je dat de absolute frequentie en de absolute frequentieverdeling vrij eenvoudige eigenschappen zijn, in tegenstelling tot de abstracte definities. Dit komt vaak voor bij definities en het is daarom belangrijk om ze te illustreren aan de hand van voorbeelden: ze helpen je de definities beter te begrijpen.

Op basis van de absolute frequentieverdeling (Tabel 2.3) kunnen we de *relatieve frequenties* berekenen: dit zijn de absolute frequenties gedeeld door de *steekproefgrootte*.

! De **steekproefgrootte** (symbool n) is gelijk aan het aantal elementen in de steekproef.

Tabel 2.4 geeft de relatieve frequenties weer: $54/90 = 0.60$ en $36/90 = 0.40$. De som van de relatieve frequenties moet 1 zijn. Het is ook mogelijk om de relatieve frequenties uit te drukken in percentages, zoals weergegeven in Tabel 2.5.

Geslacht	vrouw	man
Relatieve frequentie	0.60	0.40

Tabel 2.4: Relatieve frequentieverdeling van de variabele Geslacht.

We kunnen de relatieve frequentie ook formeel definiëren:

ze expliciet uitschrijven. Dit doen we om een rij bondig te kunnen weergeven. Bijvoorbeeld: x_1, x_2, x_3, x_4, x_5 kan bondig worden geschreven als x_1, x_2, …, x_5 of zelfs als x_1, …, x_5. Je moet minstens het eerste en het laatste element geven, maar voor de rest ben je vrij hoe bondig je het schrijft.

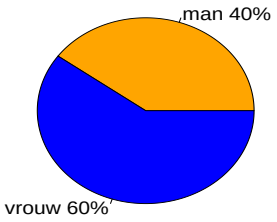
Geslacht	vrouw	man
Relatieve frequentie	60%	40%

Tabel 2.5: Relatieve frequentieverdeling (in procent) van de variabele Geslacht.

! De **relatieve frequentie** van x is de absolute frequentie gedeeld door de steekproefgrootte n.

Vaak spreken we kortweg over frequenties als het uit de context duidelijk is of het absolute of relatieve frequenties zijn. Soms zullen we ook spreken over *de verdeling van een variabele*, waarbij we het geheel van mogelijke waarden bedoelen, samen met de absolute of relatieve frequenties. Dit komt overeen met de omschrijving van de **verdeling** in paragraaf 1.2.2 op pagina 14.

Op basis van de relatieve frequenties kunnen we data visualiseren door middel van een cirkeldiagram, zie Figuur 2.3. De relatieve oppervlaktes van de stukken zijn gelijk aan de relatieve frequenties. Niettegenstaande cirkeldiagrammen populair zijn, voornamelijk omdat ze visueel aantrekkelijk zijn, wordt het toch afgeraden ze te gebruiken. Dit komt omdat het menselijk oog niet goed in staat is om de oppervlaktes van een cirkeldiagram te beoordelen. Een staafdiagram heeft dit nadeel niet en wordt daarom verkozen boven een cirkeldiagram. Voordat we een staafdiagram bespreken, illustreren we eerst hoe we in R de frequenties kunnen berekenen en hoe we een cirkeldiagram kunnen bekomen.



Figuur 2.3: Cirkeldiagram van de variabele Geslacht

Illustratie in R

Via het `$`-teken kunnen we de waarden van een variabele uit de tabel `DataIAT` opvragen. Na het `$`-teken geef je de naam van de variabele op, zoals die weergeven is in de tabel `DataIAT`. Hier zijn we geïnteresseerd in de variabele `Geslacht` uit de tabel `DataIAT`.

```
> DataIAT$Geslacht
```

```
 [1] vrouw man   man   man   vrouw vrouw vrouw man   vrouw man   man
[12] man   man   vrouw vrouw vrouw man   vrouw vrouw man   man   vrouw
[23] man   man   man   vrouw man   man   vrouw vrouw man   man   vrouw
[34] vrouw vrouw vrouw man   vrouw vrouw vrouw vrouw vrouw man   vrouw
[45] vrouw vrouw vrouw vrouw man   man   vrouw vrouw vrouw vrouw vrouw
[56] man   vrouw vrouw man   man   vrouw man   man   man   vrouw vrouw
[67] vrouw vrouw vrouw man   vrouw vrouw vrouw vrouw vrouw man   man
[78] man   vrouw man   vrouw vrouw vrouw vrouw man   vrouw vrouw man
[89] vrouw man
Levels: man vrouw
```

Via `table()` kunnen we de absolute frequenties berekenen:

```
> table(DataIAT$Geslacht)
```

```
  man vrouw
   36    54
```

De relatieve frequenties worden bekomen door de absolute frequenties te delen door 90:

```
> table(DataIAT$Geslacht)/90
```

```
  man vrouw
  0.4   0.6
```

Indien je de relatieve frequenties in percentages wil uitdrukken, moet je met 100 vermenigvuldigen.

```
> (table(DataIAT$Geslacht)/90)*100


  man vrouw
   40    60
```

Omdat we de absolute en relatieve frequenties nog een aantal keer nodig hebben, zullen we ze opslaan en een naam geven. We geven de namen `abs.freq.geslacht`, `rel.freq.geslacht` en `rel.freq.perc.geslacht`[g].

```
> abs.freq.geslacht <- table(DataIAT$Geslacht)
> rel.freq.geslacht <- table(DataIAT$Geslacht)/90
> rel.freq.perc.geslacht <- (table(DataIAT$Geslacht)/90)*100
```

Je kan de tabellen nu opvragen via hun naam:

```
> abs.freq.geslacht


  man vrouw
   36    54


> rel.freq.geslacht


  man vrouw
  0.4   0.6


> rel.freq.perc.geslacht


  man vrouw
   40    60
```

Een cirkeldiagram kan bekomen worden via `pie()`[h].

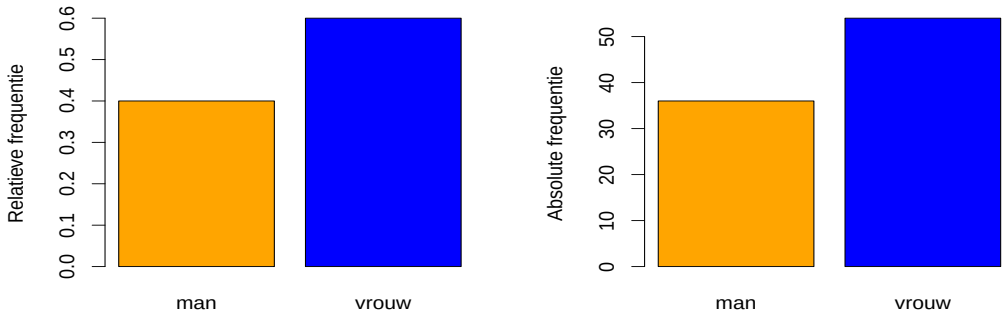
[g] De keuze van namen is vrij zolang er maar geen speciale tekens of spaties gebruikt worden.
[h] Indien je de output van de figuren wil zien, zal je de R-code zelf moeten uitvoeren.

```
> pie(rel.freq.perc.geslacht)
```

Indien je een figuur wenst gelijkaardig aan Figuur 2.3 moet je nog enkele argumenten aanvullen in de functie `pie()`[i]. Dit komt echter niet aan bod in de theoriecursus.

2.3 Staafdiagram

Het staafdiagram van de relatieve frequenties wordt weergegeven in Figuur 2.4 links. De verschillende waarden van de variabele worden horizontaal weergegeven en bij elke waarde wordt een rechthoek getekend waarbij de hoogte gelijk is aan de relatieve frequentie. De breedte van de rechthoek kan vrij gekozen worden zolang alle rechthoeken maar even breed zijn. De afstand tussen de verschillende rechthoeken moet ook dezelfde zijn. Staafdiagrammen worden voornamelijk gebruikt voor variabelen van nominaal of ordinaal meetniveau. Men kan ook een staafdiagram maken op basis van de absolute frequenties, zie Figuur 2.4 rechts. In plaats van de relatieve frequenties, staan nu de absolute frequenties op de verticale as.



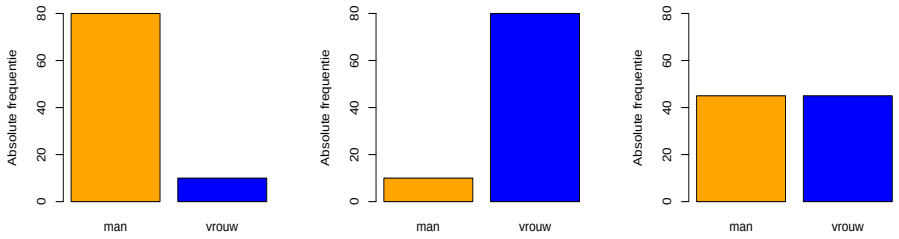
Figuur 2.4: Links: staafdiagram van de relatieve frequentie van Geslacht. Rechts: staafdiagram van de absolute frequentie van Geslacht.

Het visualiseren van data laat toe om in een oogopslag een idee te krijgen over de verdeling van de variabele. Veronderstel dat we het IAT-experiment opnieuw doen: de onderzoekers zetten het computerexperiment terug online en via advertenties rekruteren ze 90 deelnemers. Het is duidelijk dat deze steekproef (deels of volledig) uit andere

[i]Als je het commando `pie(rel.freq.perc.geslacht)` uitvoert, zal je zien dat de figuur niet volledig gelijk is aan die uit de cursus.

personen zal bestaan dan de eerste steekproef. De Figuur 2.5 links toont het staafdiagram van het geslacht in deze nieuwe steekproef. Visueel is het meteen duidelijk dat er veel meer mannen dan vrouwen zijn in deze steekproef.

Het middelste staafdiagram in Figuur 2.5 geeft de verdeling van het geslacht indien de onderzoekers het experiment nogmaals herhalen (en dus opnieuw een andere steekproef bekomen). Hier is het duidelijk dat er veel minder mannen dan vrouwen zijn. Tenslotte geeft de figuur rechts het staafdiagram weer indien men nogmaals een nieuwe steekproef bekomt: hier zijn er evenveel mannen als vrouwen aanwezig.



Figuur 2.5: Staafdiagram voor de variabele Geslacht voor 3 verschillende steekproeven.

Illustratie in R

Een staafdiagram kan bekomen worden in R via `barplot()`. Dit wordt hier toegepast op zowel de relatieve als de absolute frequenties.

```
> barplot(rel.freq.geslacht)
> barplot(abs.freq.geslacht)
```

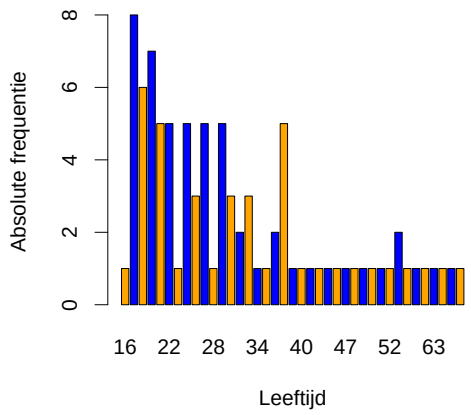
2.4 Histogram

Tabel 2.6 geeft de leeftijden weer van de 90 personen in de steekproef. Net als bij het geslacht is deze weergave van de data niet overzichtelijk. Figuur 2.6 toont een staafdiagram van de leeftijd. Deze figuur bevat echter te veel categorieën om in één oogopslag een idee te krijgen van de verdeling.

Het histogram is een betere figuur om de leeftijden te visualiseren. Om deze figuur te kunnen maken, moeten we eerst de *data groeperen*. We kunnen bijvoorbeeld de leeftijd

40	21	22	22	18	29	20	38	21	24
25	24	37	20	20	19	20	24	31	47
20	18	44	49	48	22	29	32	20	37
32	19	26	23	32	18	36	35	24	28
16	53	29	22	53	18	20	55	41	31
37	61	26	21	19	74	18	21	68	36
34	19	19	25	29	52	18	46	24	29
64	18	30	25	51	18	21	22	26	37
26	30	26	30	50	37	54	63	19	43

Tabel 2.6: De leeftijden van de 90 personen in de steekproef



Figuur 2.6: Staafdiagram van de absolute frequentie van Leeftijd. Omdat Leeftijd veel waarden kan aannemen, is een staafdiagram niet de beste keuze.

onderverdelen in 8 groepen (ook wel *klassen* of *intervallen* genoemd) zoals weergegeven in Tabel 2.7.

De **notatie** $]a,\ b]$ staat voor alle leeftijden groter dan a (a niet meegerekend), maar kleiner dan of gelijk aan b (b wel meegerekend), waarbij a en b getallen voorstellen[j]. De frequentie geeft het aantal personen weer in de overeenkomstige klasse. Er zijn bijvoorbeeld 19 personen die ouder zijn dan 20, maar jonger of gelijk aan 25. Je kan dit zelf nagaan door deze leeftijden te tellen in Tabel 2.6. Analoog voor de andere groepen. Tabel 2.7 wordt de gegroepeerde frequentieverdeling genoemd.

Om een histogram te kunnen tekenen, moeten we eerst de breedte van een klasse vastleggen.

! **Klassenbreedte**. De klassenbreedte van een interval $]a, b]$ wordt gegeven door $b - a$.

We stellen verder dat de klassenbreedte van elke type interval dezelfde is:

! De klassenbreedtes van de intervallen $]a, b]$, $[a, b]$, $[a, b[$ en $]a, b[$ zijn gelijk.

De gegroepeerde frequentieverdeling hangt af van de keuze van de klassen en het is mogelijk dat twee personen verschillende keuzes maken (andere klassen, meer of minder groepen). Dit illustreert dat het verdelen in klassen subjectief is.

Klasse	Frequentie
]15,20]	22
]20,25]	19
]25,30]	14
]30,35]	7
]35,40]	9
]40,50]	8
]50,60]	6
]60,90]	5

Tabel 2.7: De gegroepeerde frequentieverdeling van de variabele Leeftijd.

[j]De notatie $[a, b]$ staat voor alle getallen tussen a en b met a en b meegerekend. $]a, b[$ staat voor alle getallen tussen a en b zonder a en b meegerekend en $[a, b[$ staat voor alle getallen tussen a en b, a meegerend en b niet meegerekend.

! De **gegroepeerde frequentieverdeling** van een variabele X is een tabel met twee kolommen (of twee rijen) waar de eerste kolom de klassen van X weergeeft en de tweede de overeenkomstige frequenties.

Op basis van de gegroepeerde frequentieverdeling kunnen we de relatieve frequenties per klasse bepalen door de (absolute) frequenties te delen door de steekproefgrootte (hier 90). Tabel 2.8 geeft deze relatieve frequenties weer en we zullen die gebruiken om het histogram op stellen.

Bij een histogram liggen de waarden van de variabele op de horizontale as en boven elke klasse tekent men een rechthoek waarbij de breedte van de rechthoek gelijk is aan de breedte van de klasse. De hoogte van de rechthoek is gelijk aan de relatieve frequentie gedeeld door de breedte van de klasse, zodat de oppervlakte van de rechthoek gelijk is aan de relatieve frequentie. We illustreren dit aan de hand van het histogram in Figuur 2.7.

De breedtes van de rechthoeken zijn gelijk aan de breedtes van de klasse. Klasse]15,20] heeft breedte 5, terwijl klasse]50,60] breedte 10 heeft. De hoogte van de rechthoek is gelijk aan de relatieve frequentie gedeeld door de breedte. Voor de klasse]15,20] is dit $0.24/5 = 0.048$. Voor de klasse]50,60] is dit $0.07/10 = 0.007$. Analoog voor de andere klassen. Uit het histogram leren we dat er meer jongeren dan ouderen zijn in de steekproef: de meeste personen zijn 40 of jonger met vooral veel personen tussen 15 en 30 jaar. Slechts een minderheid is ouder dan 40 jaar[k].

Deze informatie kan je niet op het zicht zien door te kijken naar Tabel 2.6. Het histogram geeft ons dus meer inzicht in de data.

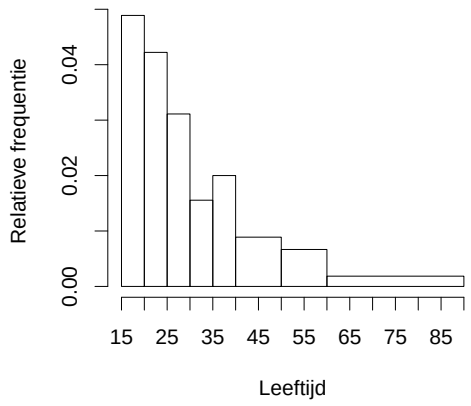
Klasse	Relatieve frequentie
]15,20]	0.24
]20,25]	0.21
]25,30]	0.16
]30,35]	0.08
]35,40]	0.10
]40,50]	0.09
]50,60]	0.07
]60,90]	0.06

Tabel 2.8: De gegroepeerde relatieve frequenties van de variabele Leeftijd.

Indien alle klassen dezelfde breedte hebben, is het ook mogelijk om een histogram op te stellen waar de hoogte (i.p.v. de oppervlakte) van de rechthoek gelijk is aan de

[k]Uit het histogram kunnen we vele besluiten trekken, maar typisch beperken we ons tot enkele.

absolute frequentie. Er bestaan dus verschillende varianten van het histogram (op basis van klassenbreedte en type van frequentie).



Figuur 2.7: Histogram van Leeftijd.

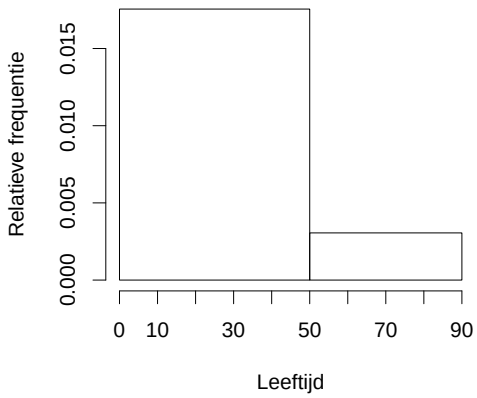
Een staafdiagram en een histogram lijken op elkaar, maar zijn zeker niet gelijk. We sommen enkele verschillen op:

- Bij een histogram raken de rechthoeken elkaar en kunnen de breedtes van de rechthoeken verschillen.
- Een staafdiagram wordt vooral gebruikt voor ordinale en nominale variabelen (omdat ze vaak een beperkt aantal waarden hebben).
- Een histogram wordt vaak gebruikt voor interval- en ratioschaal variabelen (omdat ze vaak een groot aantal waarden hebben).

Een histogram wordt gemaakt op basis van gegroepeerde data die bekomen worden door variabelen op te delen in klassen. Deze klassen worden bepaald door de gebruiker en bijgevolg is het histogram gebruikersafhankelijk. Figuur 2.8 toont een histogram waar te weinig klassen zijn aangemaakt. Hierdoor verliezen we te veel informatie: we zien dat de meeste personen jonger zijn dan 50, maar verder kunnen we niet veel afleiden, terwijl we uit Figuur 2.7 hadden besloten dat er veel personen tussen 15 en 30 jaar oud zijn. Er bestaan geen exacte regels over hoe je data in klassen moet verdelen. Vaak is het wenselijk om klassen te hebben van dezelfde breedte. Echter, als de uiterste klassen een lage frequentie hebben, voeg je ze best samen (dit is het geval bij Figuur 2.7). Er bestaan verschillende vuistregels om het *aantal* klassen te bepalen. Eén vuistregel stelt

dat je data moet indelen in ongeveer $\sqrt{n}$ klassen (n is de steekproefgrootte). Voor onze steekproef is dit $\sqrt{n} = \sqrt{90} = 9.487$, dus ongeveer 9 of 10 klassen. Voor het histogram in Figuur 2.7 hebben we 8 klassen gebruikt, wat in de buurt ligt van het aantal volgens deze vuistregel[1]. Niettegenstaande er geen exacte regels zijn voor het groeperen van data, bestaan er zeker slechte keuzes zoals geïllustreerd in Figuur 2.8. Deze gebruikersafhankelijkheid wordt meestal als een nadeel ervaren: door een slechte keuze te maken van de klassen kan de figuur een vertekend beeld geven. Verder in de cursus zullen we ook figuren zien die niet gebruikersafhankelijk zijn.

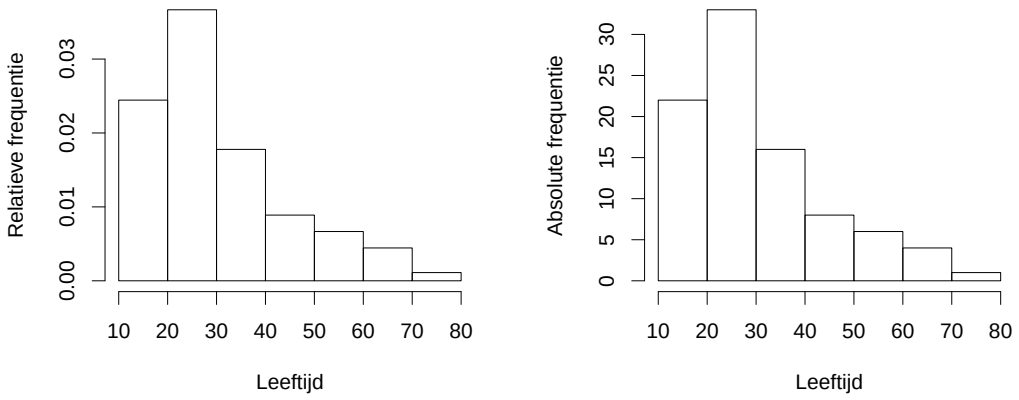
Zoals eerder aangegeven is het mogelijk om een histogram op te stellen op basis van gegroepeerde data met gelijke klassenbreedtes. De linkerfiguur in Figuur 2.9 toont dergelijk histogram. De oppervlaktes van de rechthoeken komen nog steeds overeen met de relatieve frequenties. Het is ook mogelijk om de absolute frequenties op de verticale as te plaatsen (zie Figuur 2.9 rechts). Nu komt de hoogte (en niet de oppervlakte) van de rechthoek overeen met de absolute frequentie.



Figuur 2.8: Een voorbeeld van een histogram met te weinig klassen.

Net als het staafdiagram laat het histogram toe om in één oogopslag een idee te krijgen over de verdeling van een variabele in een steekproef. Figuur 2.10 toont de histogrammen voor 3 hypothetische steekproeven en op basis van deze figuren zien we meteen of er veel jongeren of ouderen in de steekproef zitten. De steekproef horende bij de linkerfiguur bestaat voornamelijk uit jonge mensen, omdat de hoogste rechthoeken bij de leeftijden van 10 tot 30 jaar voorkomen. Men zegt ook dat de data (hier de leeftijd)

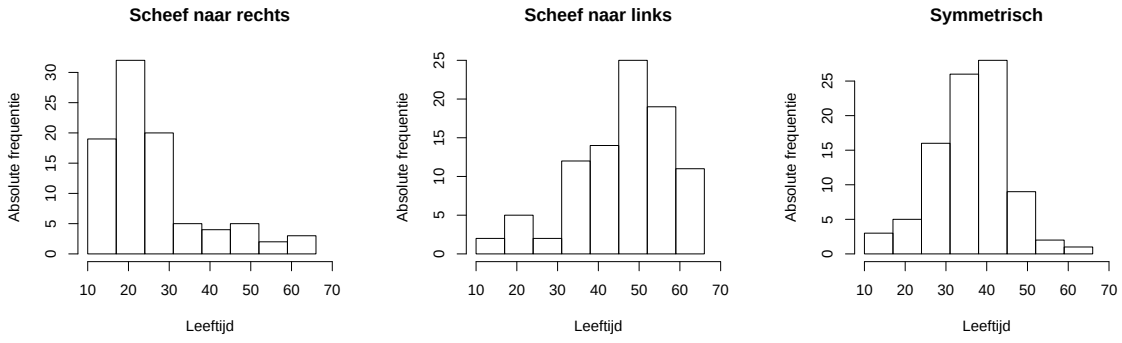
[1]Vuistregels zijn handig als hulpmiddel, maar het is belangrijk te beseffen dat het slechts praktische regels zijn. Indien iemand een histogram maakt met 8 klassen en iemand anders een histogram met 9 klassen, is er niemand juist of fout - het zijn gewoon twee verschillende keuzes.



Figuur 2.9: Een voorbeeld van een histogram waarbij de klassen even breed zijn. Links: de oppervlaktes komen overeen met de relatieve frequenties. Rechts: de hoogte komt overeen met de absolute frequentie.

scheef verdeeld zijn. Indien de meeste massa van het histogram links ligt en het uiteinde rechts uitloopt, dan zijn de data *scheef naar rechts* verdeeld. De uiteinden van een verdeling worden ook *de staarten* genoemd.

Voor de figuur in het midden, is dit net omgekeerd: ze bestaat voornamelijk uit oudere mensen met een uitlopende linkerstaart. Dit wordt *scheef naar links* genoemd. De rechterfiguur tenslotte, toont het histogram van een steekproef waar de meeste personen rond de 40 zijn en de linker- en rechterstaarten ongeveer gelijk zijn. Dit wordt een *symmetrische verdeling* genoemd.



Figuur 2.10: De histogrammen van de variabele Leeftijd bij drie (hypothetische) steekproeven. Deze figuren geven een idee hoe de leeftijden verdeeld zijn voor de verschillende steekproeven.

Illustratie in R

De klassen kunnen bekomen worden via `cut()`. Het eerste argument is de variabele waarin we geïnteresseerd zijn (hier Leeftijd) en het tweede de grenzen van de verschillende klassen:

```
> klassen <- cut(DataIAT$Leeftijd,breaks=c(15, 20, 25, 30, 35, 40, 50, 60, 90))
> klassen
```

```
 [1] (35,40] (20,25] (20,25] (20,25] (15,20] (25,30] (15,20] (35,40]
 [9] (20,25] (20,25] (20,25] (20,25] (35,40] (15,20] (15,20] (15,20]
[17] (15,20] (20,25] (30,35] (40,50] (15,20] (15,20] (40,50] (40,50]
[25] (40,50] (20,25] (25,30] (30,35] (15,20] (35,40] (30,35] (15,20]
[33] (25,30] (20,25] (30,35] (15,20] (35,40] (30,35] (20,25] (25,30]
[41] (15,20] (50,60] (25,30] (20,25] (50,60] (15,20] (15,20] (50,60]
[49] (40,50] (30,35] (35,40] (60,90] (25,30] (20,25] (15,20] (60,90]
[57] (15,20] (20,25] (60,90] (35,40] (30,35] (15,20] (15,20] (20,25]
[65] (25,30] (50,60] (15,20] (40,50] (20,25] (25,30] (60,90] (15,20]
[73] (25,30] (20,25] (50,60] (15,20] (20,25] (20,25] (25,30] (35,40]
[81] (25,30] (25,30] (25,30] (25,30] (40,50] (35,40] (50,60] (60,90]
[89] (15,20] (40,50]
8 Levels: (15,20] (20,25] (25,30] (30,35] (35,40] ... (60,90]
```

Voor elke persoon in de steekproef wordt weergegeven tot welke categorie hij/zij behoort. Het ronde haakje in R is gelijk aan het vierkantje haakje naar buiten uit de cursus. Dus `(35,40]` is hetzelfde als]35,40]. Via `table()` kunnen we de absolute en relatieve gegroepeerde frequenties bekomen zoals weergegeven in Tabellen 2.7 en 2.8:

```
> table(klassen)
```

```
klassen
(15,20] (20,25] (25,30] (30,35] (35,40] (40,50] (50,60] (60,90]
     22      19      14       7       9       8       6       5
```

```
> table(klassen)/90
```

```
klassen
   (15,20]    (20,25]    (25,30]    (30,35]    (35,40]    (40,50]
0.24444444 0.21111111 0.15555556 0.07777778 0.10000000 0.08888889
   (50,60]    (60,90]
0.06666667 0.05555556
```

Een histogram wordt bekomen via `hist()`, met als argumenten de variabele waarin we geïnteresseerd zijn (hier Leeftijd) en de grenzen van de klassen (bij de optie `breaks`).

```
> hist(DataIAT$Leeftijd, breaks = c(15, 20, 25, 30, 35, 40, 50, 60, 90))
```

Indien je de functie `hist()` gebruikt zonder de optie `breaks` zal R automatisch de data groeperen in klassen van gelijke breedte. Het kiezen van de klassen wordt gedaan op basis van algoritmes en vaak geven deze klassen een goede weergave van de data. Ook kiest R er standaard voor om een histogram op te stellen met de absolute frequenties op de verticale as.

```
> hist(DataIAT$Leeftijd)
```

2.5 Cumulatieve frequentiecurve

Om de cumulatieve frequentiecurve te kunnen opstellen, moeten we eerst de betekenis van cumulatieve frequenties begrijpen. Deze frequenties kunnen voor zowel ongegroepeerde data (zoals de originele variabele Leeftijd) als gegroepeerde data (zoals de verschillende leeftijdsklassen in Tabel 2.7) berekend worden.

2.5.1 Ongegroepeerde data

De cumulatieve absolute frequentie bekomen we door absolute frequenties op te tellen. Op een gelijkaardige wijze kunnen we de cumulatieve relatieve frequentie bekomen. Voor de eenvoud spreken we soms over de cumulatieve frequentie en uit de context moet het dan duidelijk zijn of we absolute of relatieve frequenties bedoelen.

We illustreren dit aan de hand van de variabele Leeftijd zoals weergegeven in Tabel 2.6. Op basis van deze gegevens kunnen we de (absolute) frequenties voor elke leeftijd berekenen, zie Tabel 2.9. De frequenties geven het aantal personen weer die een bepaalde

leeftijd hebben, terwijl de cumulatieve frequenties het aantal personen weergeven die een bepaalde leeftijd hebben *of jonger zijn*. Voor de leeftijd 19 is de cumulatieve frequentie 15 omdat er 6 personen 19 jaar zijn, 8 personen 18 jaar en 1 persoon 16 jaar (en niemand is jonger dan 16). Dus samengeteld zijn er 15 personen 19 jaar of jonger. De cumulatieve frequenties bekom je dus door de frequenties op te tellen. Uiteraard is de cumulatieve frequentie van de oudste persoon (hier 74 jaar) gelijk aan de steekproefgrootte: alle personen zijn immers 74 jaar of jonger. Op een gelijkaardige wijze kan je de cumulatieve relatieve frequenties bekomen door de relatieve frequenties op te tellen. Cumulatieve frequenties spelen een belangrijke rol binnen de statistiek, dit zal voornamelijk in de hoofdstukken rond kansrekening duidelijk worden.

Formeel kunnen we de begrippen als volgt omschrijven.

! De **cumulatieve absolute frequentie** van een waarde x is gelijk aan het aantal elementen in de steekproef die kleiner dan of gelijk aan x zijn. We duiden dit aan door het symbool $F(x)$.

! De **cumulatieve absolute frequentieverdeling** van X is een tabel met twee kolommen (of twee rijen), waar in de eerste kolom de waarden van de variabele X worden weergegeven en in de tweede kolom de overeenkomstige cumulatieve absolute frequenties.

Omdat de cumulatieve frequentie van 19 jaar gelijk is aan 15 kunnen we symbolisch schrijven $F(19) = 15$. Analoog lezen we af uit Tabel 2.9 dat $F(16) = 1$, $F(18) = 9$, etc.

Op een gelijkaardige wijze kan je dezelfde begrippen omschrijven voor de relatieve frequenties.

We kunnen deze cumulatieve frequenties visualiseren door middel van een cumulatieve frequentiecurve. Figuur 2.11 toont in verschillende stappen hoe je deze curve kan opstellen. Op de horizontale as staan de leeftijden en op de verticale as de cumulatieve (absolute) frequenties. De leeftijden op de horizontale as kunnen waarden bevatten die niet voorkomen in de steekproef. Bij Figuur 2.11 gaat de leeftijd van 0 jaar tot 80 jaar, terwijl dit in de steekproef van 16 jaar tot 74 jaar gaat. In een eerste stap (figuur linksboven) duiden we alle waarden van de cumulatieve frequentieverdeling aan: we zetten een punt bij een leeftijd van 16 en de cumulatieve frequentie van 1, bij een leeftijd van 18 en een cumulatieve frequentie van 9, etc (zie Tabel 2.9). In een tweede stap (figuur rechtsboven) verbinden we al deze punten trapsgewijs: voor een bepaalde leeftijd trekken we een horizontale lijn naar rechts tot de volgende leeftijd gevolgd door een verticale lijn naar boven tot het punt. In een derde stap (figuur linksonder) tekenen we voor alle leeftijden kleiner dan 16 een horizontale lijn bij een cumulatieve frequentie

Leeftijd	Absolute frequentie	Cumulatieve absolute frequentie
16	1	1
18	8	9
19	6	15
20	7	22
21	5	27
22	5	32
23	1	33
24	5	38
25	3	41
26	5	46
28	1	47
29	5	52
30	3	55
31	2	57
32	3	60
34	1	61
35	1	62
36	2	64
37	5	69
38	1	70
40	1	71
41	1	72
43	1	73
44	1	74
46	1	75
47	1	76
48	1	77
49	1	78
50	1	79
51	1	80
52	1	81
53	2	83
54	1	84
55	1	85
61	1	86
63	1	87
64	1	88
68	1	89
74	1	90

Tabel 2.9: De absolute frequenties en cumulatieve absolute frequenties van Leeftijd. De eerste en derde kolom vormen samen de cumulatieve absolute frequentieverdeling.

van 0: er zijn immers geen personen in de steekproef die jonger zijn dan 16. In een laatste stap (figuur rechtsonder) tekenen we voor alle leeftijden boven 74 jaar een horizontale lijn bij een cumulatieve frequentie van 90 (de steekproefgrootte): alle personen zijn immer 74 jaar of jonger.

Je kan ook op basis van de cumulatieve relatieve frequenties een cumulatieve frequentiecurve opstellen.

De cumulatieve frequentiecurve stelt ons nu in staat om de verdeling van de leeftijden beter te begrijpen. We zien bijvoorbeeld dat meer dan de helft van de personen jonger zijn dan 30: inderdaad, voor een leeftijd van 30 ligt de cumulatieve frequentie boven 45 (herinner je dat de steekproefgrootte 90 is). Uiteraard kan je deze informatie ook uit de cumulatieve frequentieverdeling halen (deze verdeling is zelfs nauwkeuriger dan op het zicht de curve te raadplegen), maar deze tabellen kunnen zeer lang zijn en dan is het interessant om de informatie in een figuur weer te geven.

Illustratie in R

Op basis van de absolute frequenties kunnen we de cumulatieve frequenties bekomen via `cumsum()`.

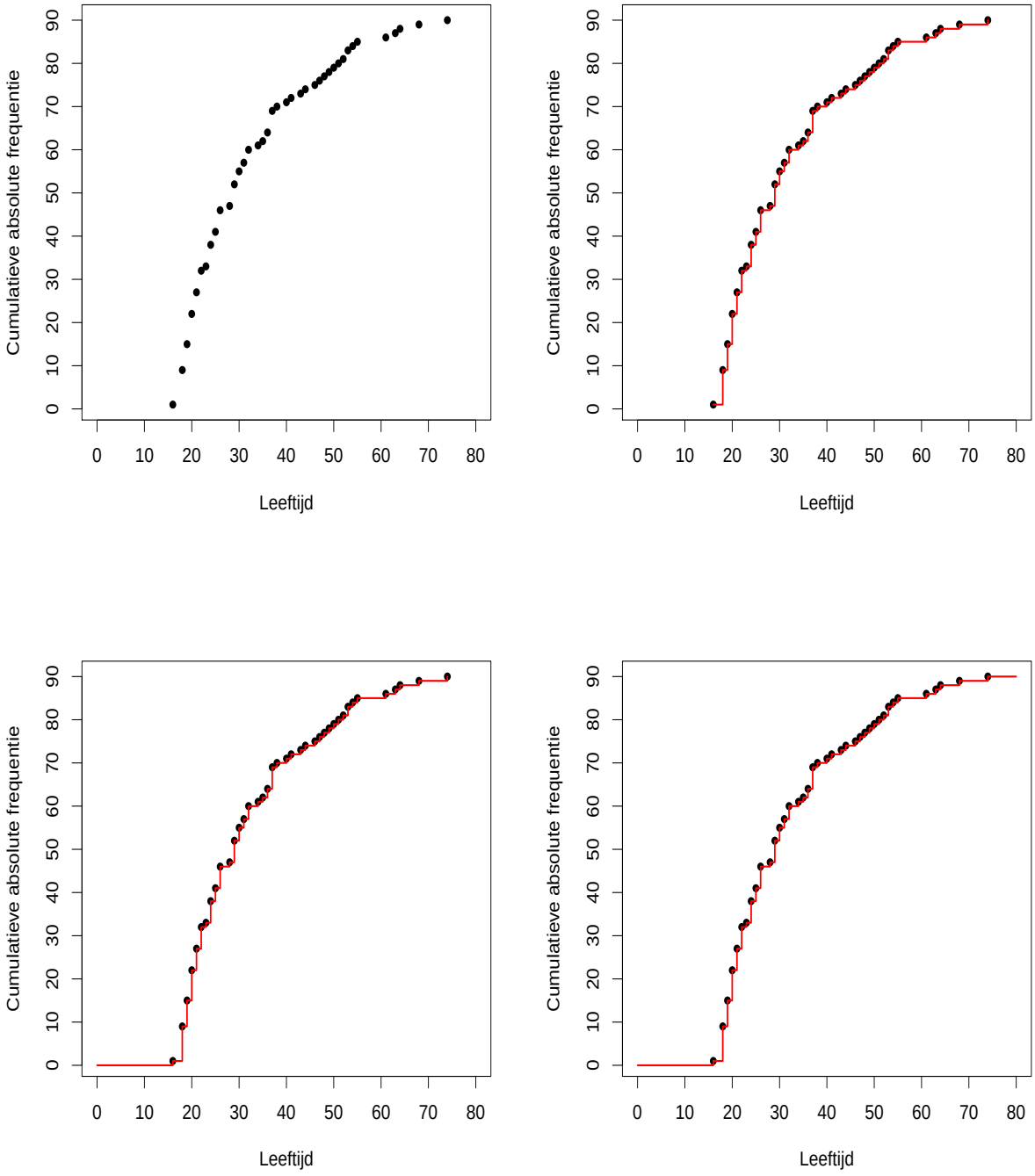
```
> abs.frequentie.leeftijd <- table(DataIAT$Leeftijd)
> cumsum(abs.frequentie.leeftijd)
```

```
16 18 19 20 21 22 23 24 25 26 28 29 30 31 32 34 35 36 37 38 40 41 43
 1  9 15 22 27 32 33 38 41 46 47 52 55 57 60 61 62 64 69 70 71 72 73
44 46 47 48 49 50 51 52 53 54 55 61 63 64 68 74
74 75 76 77 78 79 80 81 83 84 85 86 87 88 89 90
```

De eerste en de derde rij geven de leeftijden weer, de tweede en de vierde de overeenkomstige cumulatieve absolute frequenties.

Een figuur gelijkaardig aan Figuur 2.11 kan bekomen worden via `ecdf()` en `plot()`. Deze figuur geeft de cumulatieve relatieve frequenties in plaats van de cumulatieve absolute frequenties.

```
> plot(ecdf(DataIAT$Leeftijd), verticals = TRUE, pch = 20)
```



Figuur 2.11: De verschillende stappen om een cumulatieve frequentiecurve op te stellen. De figuur rechtsonder is de finale figuur.

2.5.2 Gegroepeerde data

Het is ook mogelijk om op basis van gegroepeerde data cumulatieve frequenties (zowel absolute als relatieve) op te stellen en grafisch weer te geven. We starten van de gegroepeerde absolute frequentieverdeling uit Tabel 2.7. De cumulatieve frequenties worden opnieuw bekomen door de frequenties op te tellen. De cumulatieve frequentieverdeling wordt weergegeven in Tabel 2.10. Er zitten bijvoorbeeld 22 personen in de leeftijdsklasse]15,20] en 19 personen in de leeftijdsklasse]20,25], bijgevolg zijn er 41 personen (22 + 19 = 41) die tot de leeftijdsklasse]20,25] of een lagere klasse behoren.

! De **cumulatieve absolute frequentie** van een klasse is gelijk aan het aantal elementen in die klasse plus het aantal elementen in lagere klassen.

! De **cumulatieve absolute gegroepeerde frequentieverdeling** van X is een tabel met twee kolommen (of rijen) waar de eerste kolom de klassen van X weergeeft en de tweede kolom de overeenkomstige cumulatieve absolute frequenties.

Analoog voor de relatieve frequenties.

Op het eerste zicht lijkt het bepalen van de cumulatieve frequentieverdeling dezelfde als bij de ongegroepeerde data. Echter, bij gegroepeerde data is er een extra moeilijkheid. Stel dat we de exacte leeftijden van de personen in de steekproef niet hebben, maar enkel de klassen van de gegroepeerde data. Vervolgens wil je weten wat de cumulatieve frequentie is horende bij een leeftijd van 45 jaar. We weten uit Tabel 2.10 dat 71 personen 40 jaar zijn of jonger en dat 79 personen 50 jaar zijn of jonger. We kunnen echter geen exacte uitspraak doen over het aantal personen van 45 jaar of jonger. Om hier toch een waarde voor te bekomen, zullen we een rekenregel toepassen. Omdat 45 precies in het midden van het interval]40,50] ligt, zijn er 5 leeftijden kleiner dan of gelijk aan 45 (namelijk 41, 42, 43, 44 en 45 zelf) en 5 leeftijden groter dan 45 (namelijk 46, 47, 48, 49 en 50). Er zitten 8 personen in deze klasse en we zullen veronderstellen[m] dat de helft van die personen tussen de 41 en de 45 jaar oud zijn en de andere helft tussen de 46 en 50 jaar oud zijn. Dus volgens deze rekenregel is het aantal personen dat 45 jaar is of jonger gelijk aan het aantal dat 40 jaar is of jonger (dit zijn er 71) plus de helft van de personen die tussen de 40 en 50 jaar oud zijn (er zijn er 8 in totaal, dus de helft is 4). In totaal zijn er dus 75 personen 45 jaar of jonger. Symbolisch kunnen we dit als volgt neerschrijven[n]:

$$F(45) = 71 + \frac{5}{10} \times 8 = 71 + \frac{1}{2} \times 8 = 71 + 4 = 75.$$

[m] Deze veronderstelling kan uiteraard fout zijn, het is slechts een rekenregel die we toepassen.

[n] De uitdrukking $\frac{5}{10} \times 8$ staat voor vijf tienden vermenigvuldigd met acht. Dus $\frac{5}{10} \times 8 = 4$.

Analoog kan je cumulatieve frequentie berekenen van 43 jaar. In het interval]40,50] zijn er 3 leeftijden gelijk aan of kleiner dan 43 (namelijk 41, 42 en 43 zelf) en 7 leeftijden groter dan 43 (namelijk 44, 45, 46, 47, 48, 49 en 50). Bijgevolg zullen we het aantal personen in dit interval vermenigvuldigen met 3/10. Dit geeft ons dan:

$$F(43) = 71 + \frac{3}{10} \times 8 = 73.4.$$

Dit laatste voorbeeld illustreert dat de rekenregel aanleiding geeft tot niet-gehele getallen, iets wat in de praktijk niet mogelijk is: er zijn geen 73.4 personen van 43 jaar of jonger.

Klasse	Cumulatieve frequentie
]15,20]	22
]20,25]	41
]25,30]	55
]30,35]	62
]35,40]	71
]40,50]	79
]50,60]	85
]60,90]	90

Tabel 2.10: De gegroepeerde cumulatieve absolute frequentieverdeling van de variabele Leeftijd.

Voor gegroepeerde data kunnen we ook een cumulatieve frequentiecurve maken. Figuur 2.12 geeft de verschillende stappen weer. Voor de horizontale as zullen we ter hoogte van de klassengrenzen (hier 15, 20, 25, 30, 35, 40, 50, 60, 90) punten tekenen (figuur linksboven). Voor de eerste waarde is de cumulatieve frequentie 0: er is niemand jonger dan 15 in de steekproef. We zetten een punt ter hoogte van 15 op de horizontale as en 0 op de verticale as. Bij de andere waarden is de cumulatieve frequentieverdeling diegene van de klasse waarvoor de waarde de bovengrens is. Voor 20 jaar bijvoorbeeld, zijn er 22 personen die tot de klasse]15,20] behoren. We zetten dus een punt ter hoogte van 20 op de horizontale as en 22 op de verticale as. Voor 25 jaar zijn er 41 die tot de klasse]20,25] of een lagere klasse behoren, dus we zetten een punt ter hoogte van 25 op de horizontale as en 41 op de verticale as. Dit doen we voor alle klassen. Vervolgens verbinden we deze punten met rechten (figuur rechtsboven). Dit is verschillend van Figuur 2.11: we tekenen geen trapsgewijze lijnen, maar rechten. Deze rechten komen overeen met de formule die we gebruiken om de cumulatieve frequenties te berekenen van leeftijden die binnen een interval liggen, zoals eerder geïllustreerd voor 45 en 43 jaar ($F(45) = 75$ en $F(43) = 73.4$). In een derde stap (figuur linksonder) tekenen we voor alle leeftijden onder 15 jaar een horizontale lijn bij de cumulatieve frequentie van 0: er zijn immers geen personen in de steekproef die jonger zijn dan 15. In een laatste stap (figuur rechtsonder) tekenen we voor alle leeftijden boven 90 jaar een horizontale

lijn bij een cumulatieve absolute frequentie van 90 (de steekproefgrootte): alle personen zijn immer 90 jaar of jonger.

Er zijn duidelijke verschillen tussen de cumulatieve frequentiecurve op basis van de originele (ongegroepeerde) leeftijden (Figuur 2.11) en die op basis van de gegroepeerde leeftijden (Figuur 2.12). Bij de ongegroepeerde data hebben we meer informatie: we weten bijvoorbeeld dat de oudste persoon 74 jaar is, terwijl we voor de gegroepeerde data enkel weten dat niemand ouder is dan 90. Op basis van onze rekenregel bekomen we bij de gegroepeerde data dat 75 personen 45 zijn of jonger, terwijl dit op basis van de ongegroepeerde data 74 is. Het groeperen van de data leidt tot informatieverlies. Indien de originele data dus beschikbaar zijn, is het beter om de data niet te groeperen om de cumulatieve frequenties te berekenen.

2.6 Een voorbeeld: grafische voorstelling van raciale voorkeur

In deze paragraaf gebruiken we de verschillende tabellen en figuren van voorgaande paragrafen om inzicht te krijgen in de raciale voorkeur van de personen in de steekproef. Uit paragraaf 2.2 weten we dat 60% van de deelnemers vrouwen zijn, terwijl de analyses in paragrafen 2.4 en 2.5 hebben aangetoond dat de meeste deelnemers jonger dan 30 zijn[o].

Tabel 2.11 toont de relatieve frequentieverdeling van de variabele Ras. We zien dat er iets meer blanken hebben deelgenomen aan het experiment. Figuur 2.13 toont de staafdiagrammen van de data afkomstig van de gevoelsthermometer[p]. We zien dat de scores wat verschillen: de laagste score t.o.v. zwarten is 3 terwijl dit 0 is t.o.v. blanken. Ook zijn er iets meer personen met een score van 7 of 8 t.o.v. blanken in vergelijking met de score t.o.v. zwarten.

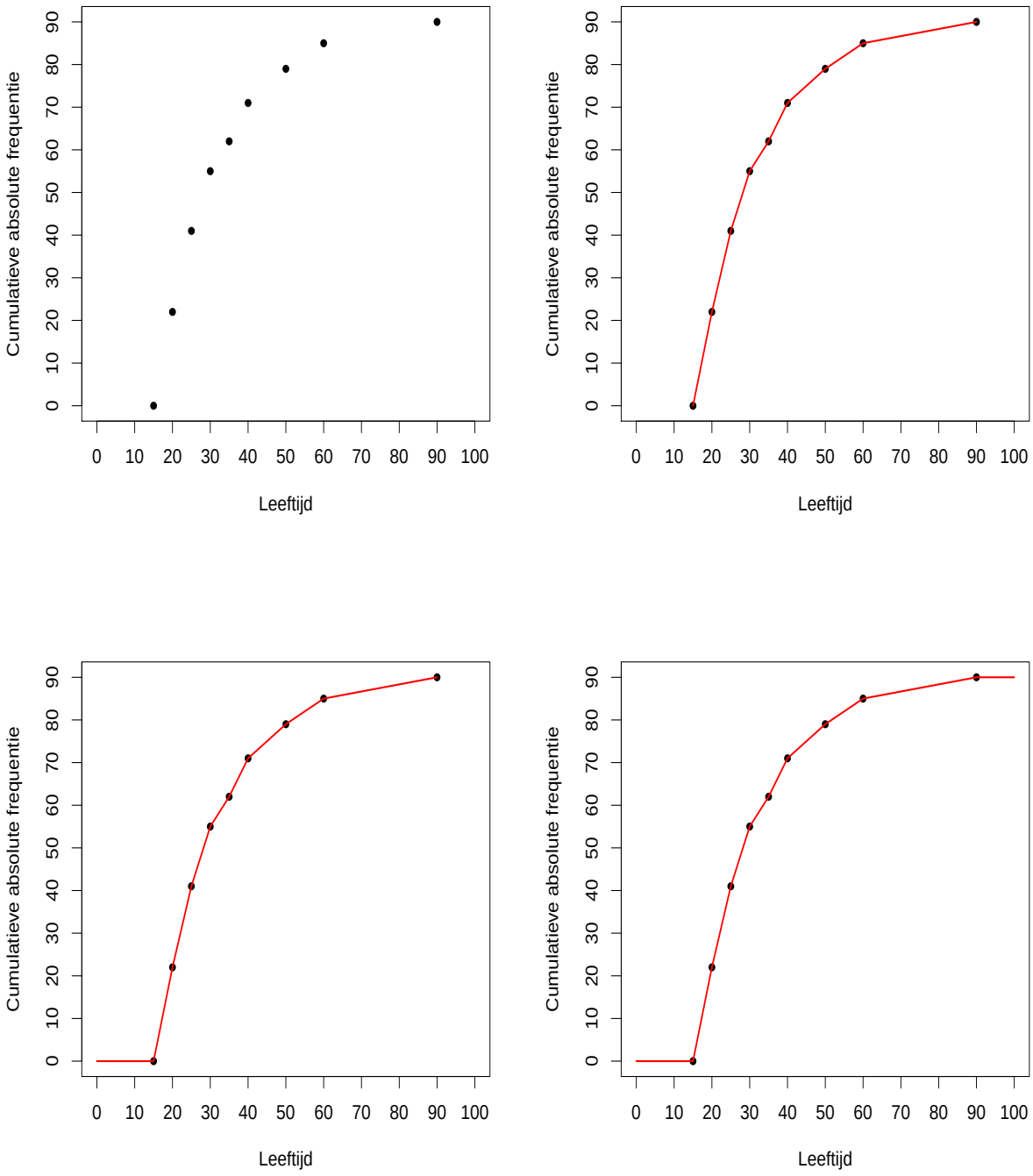
Ras	blank	zwart
Relatieve frequentie	61%	39%

Tabel 2.11: Relatieve frequentieverdeling van de variabele Ras.

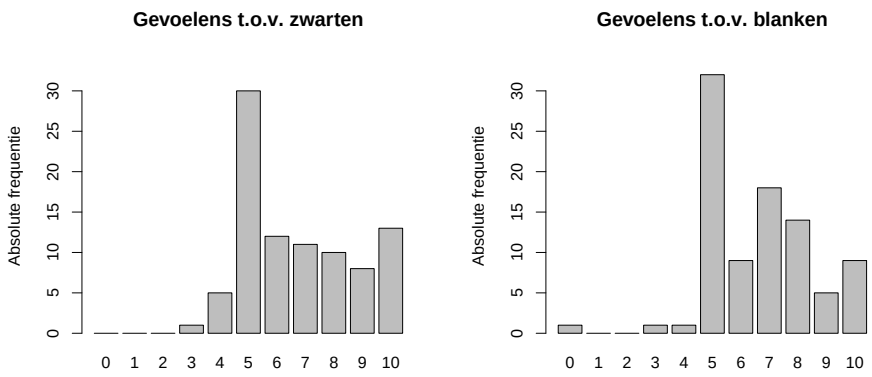
Omdat dit een studie is naar raciale voorkeur, is het interessanter om deze figuren opnieuw te maken, maar nu per ras. We maken dus een staafdiagram voor de zwarten

[o]Uiteraard kunnen er nog vele andere conclusies geformuleerd worden.

[p]Herinner je dat deze variabele een score voorstelt tussen 0 en 10 met 0 = koude gevoelens, 5 = neutraal, 10 = warme gevoelens.



Figuur 2.12: De verschillende stappen om een cumulatieve frequentiecurve op te stellen op basis van gegroepeerde data. De figuur rechtsonder is de finale figuur.



Figuur 2.13: Staafdiagram van de gevoelens t.o.v. zwarten en blanken.

en een staafdiagram voor de blanken. Figuur 2.14 geeft deze staafdiagrammen weer. Voor zowel blanken als zwarten zien we een (bescheiden) verschuiving in de gevoelens als men een uitspraak moet doen over het eigen ras.

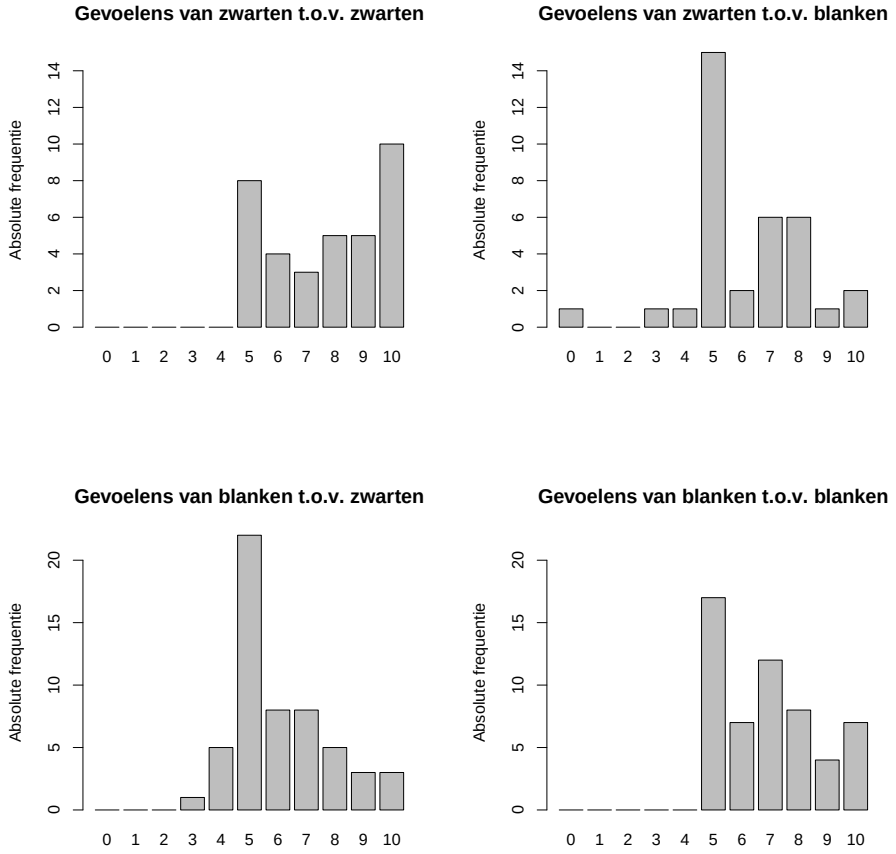
Hoewel deze figuren ons al inzicht verschaffen in de data, geven ze geen informatie over de individuele rapportering van de gevoelens: welke personen hebben warmere gevoelens voor hun eigen ras? Om dit te kunnen bestuderen zullen we een nieuwe variabele moeten aanmaken op basis van de gevoelens voor blanken en de gevoelens voor zwarten. Deze variabele noemen we *Relatieve voorkeur gevoel* en heeft de volgende waarden:

$$\begin{cases} \textit{zwart} & \text{indien Gevoelens t.o.v. zwarten} > \text{Gevoelens t.o.v. blanken} \\ \textit{neutraal} & \text{indien Gevoelens t.o.v. zwarten} = \text{Gevoelens t.o.v. blanken} \\ \textit{blank} & \text{indien Gevoelens t.o.v. zwarten} < \text{Gevoelens t.o.v. blanken} \end{cases}$$

Voor de eerste persoon in de dataset (zie Tabel 2.1 op pagina 36) bijvoorbeeld, zou *Relatieve voorkeur gevoel* de waarde *zwart* aannemen, omdat de gevoelens t.o.v. zwarten warmer (groter) zijn dan die t.o.v. blanken ($10 > 9$).

Tabel 2.12 geeft de absolute frequenties weer volgens ras. Er zijn bijvoorbeeld 2 zwarte personen die aangeven een relatieve voorkeur te hebben voor blanken, terwijl er 11 raciaal neutraal zijn en 22 die een relatieve voorkeur hebben voor zwarten. De meeste blanken rapporteren raciaal neutraal te zijn, terwijl de meeste zwarten rapporteren een voorkeur te hebben voor hun eigen ras. Er zijn echter ook veel zwarten die aangeven raciaal neutraal te zijn en veel blanken die een voorkeur hebben voor blanken.

We kunnen nu de reactietijden analyseren om te onderzoeken of deze conclusies ook



Figuur 2.14: Staafdiagram van de gevoelens t.o.v. zwarten en blanken per ras

Ras = zwart

Relatieve voorkeur gevoel	Frequentie
blank	2
neutraal	11
zwart	22

Ras = blank

Relatieve voorkeur gevoel	Frequentie
blank	24
neutraal	30
zwart	1

Tabel 2.12: De absolute frequenties van de variabele Relatieve voorkeur gevoel *opgedeeld volgens* Ras.

opgaan voor de impliciete raciale voorkeur zoals gemeten door een IAT. We maken een variabele analoog aan de variabele *Relatieve voorkeur gevoel*, maar nu op basis van de reactietijden van de congruente en incongruente opdrachten[q]. De variabele *Relatieve impliciete voorkeur* neemt de volgende waarden aan

$$\begin{cases} \textit{zwart} & \text{indien congruente reactietijd} > \text{incongruente reactietijd} \\ \textit{neutraal} & \text{indien congruente reactietijd} = \text{incongruente reactietijd} \\ \textit{blank} & \text{indien congruente reactietijd} < \text{incongruente reactietijd} \end{cases}$$

Met congruente reactietijd bedoelen we de reactietijd van de congruente opdrachten en analoog voor de incongruente reactietijd. Tabel 2.13 geeft de absolute frequenties weer volgens ras. Geen enkele persoon heeft de waarde 'neutraal'. Dit is niet verwonderlijk, de reactietijden worden in milliseconden gemeten en het is bijna onmogelijk om exact dezelfde tijd te registreren voor de congruente en incongruente opdrachten. Van de zwarte personen in de steekproef zijn er 19 die een impliciete voorkeur hebben voor hun eigen ras, terwijl 16 personen een impliciete voorkeur hebben voor het ander ras. Bij de blanken is het verschil nog kleiner en zijn er net iets meer die een impliciete voorkeur hebben voor het andere ras. We zien dus geen sterke (impliciete) voorkeur voor het eigen ras.

Ras = zwart

Relatieve impliciete voorkeur	Frequentie
blank	16
neutraal	0
zwart	19

Ras = blank

Relatieve impliciete voorkeur	Frequentie
blank	27
neutraal	0
zwart	28

Tabel 2.13: De absolute frequenties van de variabele Relatieve impliciete voorkeur *opgedeeld volgens* Ras.

We kunnen dit verder in detail bekijken door histogrammen op te stellen op basis van het verschil in reactietijd tussen de congruente en incongruente opdrachten:

Verschil in reactietijd = congruente reactietijd - incongruente reactietijd

Als dit verschil positief is, is de reactietijd bij de congruente opdrachten groter (dus een tragere reactie) dan bij de incongruente opdrachten. Het omgekeerde geldt als het

[q]Bij de congruente opdrachten horen het concept 'Blank' en de evaluatie 'Goed' samen. Bij de incongruente opdrachten horen het concept 'Zwart' en de evaluatie 'Goed' samen. Zie pagina 33 voor meer informatie.

verschil negatief is. Door het verschil in tijd te visualiseren kunnen we ook een beeld krijgen van *hoeveel* tijdsverschil er zit tussen de types van opdrachten. De variabele *Relatieve impliciete voorkeur* geeft enkel aan voor welk type opdracht men de snelste tijd heeft, zonder informatie te geven over de grootte van het tijdsverschil.

Figuur 2.15 toont de histogrammen van dit tijdsverschil volgens beide rassen[r]. Bij de zwarten (figuur links) kunnen de verschillen in reactietijd iets groter zijn wanneer men voor de incongruente opdrachten de snelste tijd had (de positieve verschillen). Bij de blanken (figuur midden) is er een persoon waarvoor de reactietijd bij de incongruente opdracht ongeveer 2000 milliseconden trager was. Dit extreem verschil zorgt ervoor dat het rechterdeel van het histogram moeilijk leesbaar is. Daarom geeft de figuur rechts een ingezoomde versie weer, waarbij de horizontale as loopt van -400 tot 600. We zien dat de verschillen in reactietijd iets groter kunnen zijn wanneer men voor de incongruente opdrachten de snelste tijd had.

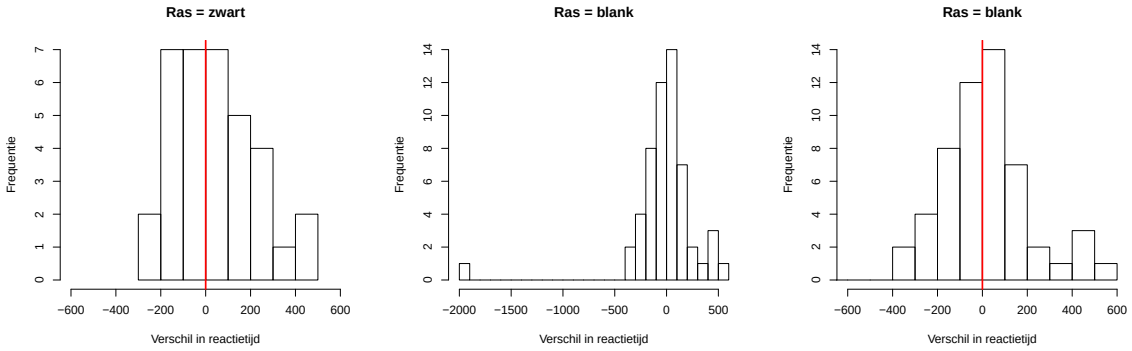
Op deze kleine verschillen na, ziet de verdeling van de positieve verschillen er min of meer gelijk uit als de verdeling van de negatieve verschillen. We hebben dus geen overtuigend bewijs gevonden dat er sterke impliciete raciale voorkeuren zijn. Figuren en frequentieverdeling alleen zijn echter niet voldoende om de data te analyseren. Het interpreteren van figuren is bijvoorbeeld subjectief en verschillende personen kunnen tot verschillende besluiten komen. Niettegenstaande figuren zeer nuttig zijn, hebben we nog nood aan extra statistische technieken. Het numeriek samenvatten van data is hiervan een belangrijk voorbeeld en wordt besproken in Hoofdstuk 3.

2.7 Samenvatting

In dit hoofdstuk hebben we verschillende statistische methodes besproken die ons in staat stellen data te visualiseren. We hebben gebruik gemaakt van:

- frequentietabellen.
- cirkeldiagrammen en staafdiagrammen.
- histogrammen.

[r]Herinner je dat we de data eerst moeten groeperen alvorens we een histogram kunnen opstellen en dat het groeperen van data subjectief is. Verschillende personen kunnen andere keuzes maken van klassen en daardoor kan het histogram er wat anders uitzien. Hier hebben we gekozen om de histogrammen op te delen in gelijke klassen waarbij op de verticale as de absolute frequenties worden weergegeven.



Figuur 2.15: Histogram van het verschil tussen de reactietijd van de congruente en incongruente opdrachten (bij een positief verschil is er een snellere reactietijd voor de incongruente opdrachten). Links: het verschil voor de zwarten. Midden en rechts: het verschil voor de blanken. De figuur rechts is een zoom-in van de figuur in het midden om de leesbaarheid te vergroten. De rode verticale lijn bij de figuren links en rechts komt overeen met een verschil van nul.

- cumulatieve frequenties en cumulatieve frequentiecurves.

Het maken van klassen (data groeperen) kan soms handig zijn om data overzichtelijk weer te geven (zoals bij een histogram). Het groeperen resulteert echter ook in informatieverlies waardoor we soms gebruik moeten maken van rekenregels.

De meerwaarde van deze methodes hebben we geïllustreerd aan de hand van een studie rond raciale voorkeur. Om deze studie te kunnen begrijpen, is het noodzakelijk te weten hoe een IAT werkt. Dit maakt echter geen deel uit van de examenstof. Het dient louter om aan te tonen hoe statistiek in de praktijk wordt toegepast en hoe je statistische methoden kan gebruiken om een antwoord te formuleren op een onderzoeksvraag.

Hoofdstuk 3

Samenvatten van data

Naast het visualiseren van data, kan het numeriek samenvatten ook het inzicht in de gegevens vergroten. Het *gemiddelde* is veruit de populairste methode om data samen te vatten. In dit hoofdstuk bespreken we het gemiddelde samen met andere samenvattingsmaten. Paragrafen 3.1-3.3 bevatten de essentie. Door deze methodes toe te passen op de studie rond raciale voorkeur uit Hoofdstuk 2, illustreren we in paragraaf 3.4 hoe samenvattingsmaten in de praktijk kunnen worden gebruikt.

3.1 Centrummaten

Een *centrummaat* is een maat voor het 'centrum' van een verdeling en ze laat toe om de waarden van een variabele samen te vatten in één getal. Omdat het 'centrum' van een verdeling niet altijd eenduidig vastligt, zal men soms ook spreken over 'locatiematen' in plaats van 'centrummaten'.

3.1.1 Het gemiddelde

Het gemiddelde is een belangrijke centrummaat die we kunnen berekenen op basis van de waarden van een variabele, op basis van een frequentieverdeling of op basis van gegroepeerde data.

Het gemiddelde op basis van de waarden van een variabele

Het rekenkundig gemiddelde is een zeer populaire centrummaat. Je kan dit getal berekenen door alle waarden van een variabele op te tellen en te delen door de steekproefgrootte. Door gebruik te maken van de notatie die we hebben ingevoerd in paragraaf 2.2 op pagina 39, kunnen we voor een variabele X in een steekproef met 3 elementen het rekenkundig gemiddelde symbolisch schrijven als:

$$\frac{x_1 + x_2 + x_3}{3}.$$

Voor alle duidelijkheid: het laatste puntje naast de voorgaande formule is het leesteken en heeft dus niets te maken met de symbolische schrijfwijze van het rekenkundig gemiddelde.

In een steekproef met 4 element wordt dit:

$$\frac{x_1 + x_2 + x_3 + x_4}{4},$$

en in een steekproef met 5 elementen:

$$\frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}.$$

We gebruiken 3 puntjes '...' om lange wiskundige bewerkingen te verkorten. Het rekenkundig gemiddelde voor een steekproef met 5 elementen kan verkort geschreven worden als:

$$\frac{x_1 + \ldots + x_5}{5}.$$

We kunnen dit doen voor steekproefgroottes van 6, 7, 8, etc. Meer algemeen voor een steekproef met n elementen (dus steekproefgrootte n) kunnen we het rekenkundig gemiddelde beknopt schrijven als:

$$\frac{x_1 + \ldots + x_n}{n}.$$

Deze symbolische notatie drukt uit dat je alle waarden van de variabele moet optellen en delen door de steekproefgrootte. Door de volgorde van bewerkingen te wijzigen, kan dit ook geschreven worden als:

$$\frac{1}{n}(x_1 + \ldots + x_n).$$

Niettegenstaande deze notatie al compact is, is het wenselijk om een nog meer compactere notatie te bekomen. Dit zal vooral handig zijn bij meer complexe bewerkingen[a].

[a] Dit zal later in de cursus duidelijk worden.

Deze compacte notatie bekomen we door middel van het symbool Σ (de Griekse letter Sigma). Dit symbool wordt ook het sommatieteken genoemd. Er geldt:

$$\sum_{i=1}^{n} x_i = x_1 + \ldots + x_n.$$

Enkele voorbeelden:

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 \quad \sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 \quad \sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5.$$

Deze notatie laat toe het rekenkundig gemiddelde te schrijven als:

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

Door de volgorde van bewerkingen te wijzigen kan dit ook geschreven worden als:

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

Deze laatste twee uitdrukkingen zijn dus compacte notaties om aan te geven dat je alle waarden moet optellen en delen door de steekproefgrootte. Omdat het rekenkundig gemiddelde zo belangrijk is, hebben we er ook een symbool voor: $\bar{x}$. Formeel kunnen we stellen:

! Het **rekenkundig gemiddelde** van een variabele X in een steekproef wordt gegeven door

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

! **Meetniveau**. Het rekenkundig gemiddelde is enkel zinnig voor interval- en ratiovariabelen.

Bij ordinale variabelen is het niet zinnig om het rekenkundig gemiddelde te berekenen: de numerieke waarden van de variabele dienen enkel om een volgorde weer te geven. Indien we bijvoorbeeld de uitslag van een wedstrijd numeriek coderen als 1 = goud, 2 = zilver en 3 = brons, dan is het 'gemiddelde podium' gelijk aan $\bar{x} = \frac{1+2+3}{3} = 2$. Niettegenstaande je dit getal wel wiskundig kan berekenen is het statistisch onzinnig om dit te doen: het 'gemiddelde podium' heeft geen betekenis.

Omdat het rekenkundig gemiddelde niet zinnig is voor ordinale variabelen is het automatisch ook niet zinnig voor nominale variabelen (omwille van de hiërarchische structuur van overerving, zie paragraaf 1.3.1 op pagina 15). Het houdt geen steek om bijvoorbeeld het gemiddelde van twee rekeningnummers te berekenen.

Naast het rekenkundig gemiddelde bestaan er ook andere soorten gemiddeldes, zoals het *harmonisch gemiddelde* en het *meetkundig gemiddelde*. Het harmonisch gemiddelde wordt gegeven door:

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}.$$

Het meetkundig gemiddelde bekomen we door alle waarden te vermenigvuldigen en dan de n-de machtswortel te nemen:

$$\sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}.$$

Niettegenstaande de formules van het harmonisch en meetkundig gemiddelde er misschien vreemd uitzien, zijn er situaties waar ze een relevante betekenis hebben. Het rekenkundig gemiddelde is echter veruit de populairste van alle gemiddeldes en wordt daarom ook kortweg *het gemiddelde* genoemd.

We gebruiken de data uit Hoofdstuk 2 om het gemiddelde te illustreren. In Tabel 2.6 op pagina 45 worden de leeftijden van alle 90 personen in de steekproef weergegeven. Het rekenkundig gemiddelde is dus[b]:

$$\frac{40 + 21 + 22 + \ldots + 43}{90} = 31.31.$$

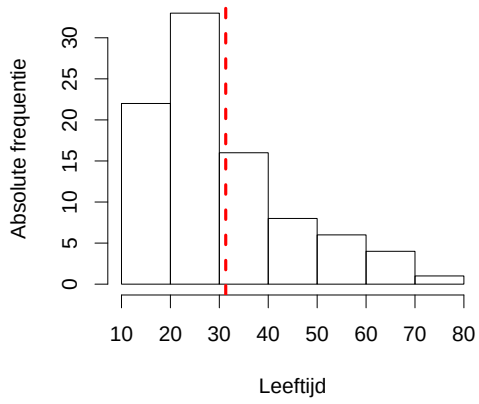
Opgelet: in de syllabus zullen we vaak getallen afronden tot op twee cijfers na de komma (twee decimalen) volgens de klassieke afrondingsregels:

- indien het derde cijfer na de komma kleiner is dan 5, ronden we naar beneden af.
- indien het derde cijfer na de komma gelijk aan of groter is dan 5, ronden we naar boven af.

Volgens deze rekenregels wordt bijvoorbeeld 1.454 afgerond naar 1.45, terwijl 1.455 wordt afgerond naar 1.46. We ronden enkel het finale resultaat af, dus voor alle bewerkingen gebruiken we de originele (niet-afgeronde) getallen.

[b]Herinnering: voor kommagetallen wordt een punt gebruikt. Dus 31.31 lees je als 31 komma 31.

Figuur 3.1 toont het histogram van Leeftijd met een aanduiding van het (rekenkundig) gemiddelde: bij de waarde 31.31 op de horizontale as tekenen we een rode verticale stippellijn. Deze figuur illustreert waarom het gemiddelde een *centrummaat* (ook een *maat van centrale tendentie* genoemd) is: ze ligt vaak (maar niet altijd) in het centrum van de verdeling.



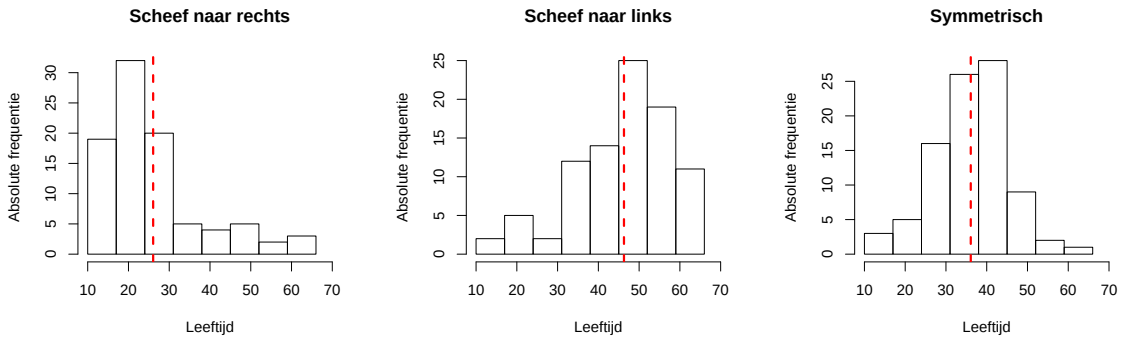
Figuur 3.1: Een histogram van de variabele Leeftijd met aanduiding van het gemiddelde.

Figuur 3.2 toont de histogrammen afkomstig uit Figuur 2.10 op pagina 50 van de verdeling van de leeftijden in drie hypothetische steekproeven. Het gemiddelde ligt iedere keer ongeveer in het centrum van de verdeling. Bij de symmetrische verdeling ligt het gemiddelde mooi centraal, terwijl bij de verdeling scheef naar rechts het gemiddelde wat naar links opschuift. Dit houdt steek: bij een scheve verdeling naar rechts zijn er immers meer jongere mensen en die zullen ervoor zorgen dat de gemiddelde leeftijd lager zal liggen (dus naar links opschuift op de horizontale as). Bij de verdeling scheef naar links, geldt het omgekeerde.

Illustratie in R

Via `mean()` kan je in R het rekenkundig gemiddelde berekenen:

```
> mean(DataIAT$Leeftijd)

[1] 31.31111
```



Figuur 3.2: De histogrammen van de variabele Leeftijd bij drie verschillende steekproeven met aanduiding van het gemiddelde.

Voor het harmonisch en meetkundig gemiddelde zijn er geen standaardfuncties in het basispakket van R. Het is echter eenvoudig om de functionaliteit van R uit te breiden door het laden van zogenaamde *pakketten*. Het pakket `psych` bevat functies om deze gemiddeldes te berekenen en kan geladen worden via het commando[c]:

```
> library("psych")
```

Via `harmonic.mean()` en `geometric.mean()` kunnen we nu het harmonisch en geometrisch gemiddelde berekenen:

```
> harmonic.mean(DataIAT$Leeftijd)
```

```
[1] 27.00577
```

```
> geometric.mean(DataIAT$Leeftijd)
```

```
[1] 28.9237
```

In deze cursus zullen wij vooral het rekenkundig gemiddelde gebruiken en zoals al eerder aangegeven noemen we dit kortweg het gemiddelde.

[c]Als je werkt via Athena dan volstaat `library("psych")` om het pakket te laden. Indien je werkt op een lokale versie van RStudio zal je dit pakket eerst eenmalig moeten downloaden via `install.packages("psych")` om dan vervolgens `library("psych")` uit te voeren.

Het gemiddelde berekenen op basis van de frequentieverdeling

In paragraaf 3.1.1 hebben we het (rekenkundig) gemiddelde berekend op basis van de waarden van de variabele Leeftijd door alle waarden op te tellen en te delen door de steekproefgrootte. We kunnen het gemiddelde ook berekenen op basis van de frequentieverdeling. Voordat we dit kunnen doen, moeten we eerst wat extra **notatie** invoeren.

! Bij een frequentieverdeling duiden we met x_i^u de *unieke* waarden aan van de variabele X in de steekproef.

! De absolute frequentie horende bij de waarde x_i^u wordt aangeduid als f_i.

In Hoofdstuk 2 op pagina 39 hebben we de notatie x_i ingevoerd om waarden aan te duiden van de variabele X voor alle elementen in de steekproef. Voor de variabele Geslacht was dit $x_1 =$ vrouw, $x_2 =$ man, $x_3 =$ man, ..., $x_{90} =$ man. De notatie x hebben we ingevoerd om *één* van de mogelijke waarden van de variabele X aan te duiden. Bij Geslacht kan dit dus zijn $x =$ vrouw of $x =$ man. De nieuwe notatie x_i^u laat nu toe om *alle unieke* waarden van X weer te geven. Voor Geslacht zal dit dus zijn $x_1^u =$ vrouw en $x_2^u =$ man[d]. Via deze notatie kunnen we uit Tabel 2.3 op pagina 38 aflezen dat $f_1 = 54$ en $f_2 = 36$. Dit is dus symbolische notatie om weer geven dat er 54 vrouwen en 36 mannen zijn.

Op dezelfde wijze kunnen we deze nieuwe notatie gebruiken voor de variabele Leeftijd in Tabel 2.9 op pagina 54. Hier geeft de eerste kolom de unieke waarden weer van de leeftijden (x_i^u) en de tweede kolom de absolute frequenties (f_i).

Via deze nieuwe notatie kunnen we het gemiddelde op basis van een frequentieverdeling berekenen via de formule

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} f_i x_i^u, \tag{3.2}$$

waarbij p staat voor het aantal unieke waarden van de variabele X in de steekproef. We illustreren dit voor de variabele Leeftijd[e].

De eerste kolom geeft de unieke leeftijden weer (dus x_i^u) en de tweede kolom de absolute frequenties (dus f_i). Formule (3.2) drukt uit dat we eerst elke frequentie moeten vermenigvuldigen met de overeenkomstige leeftijd, vervolgens moeten we al deze getallen

[d] De volgorde van de cijfers maakt niet uit, dus we kunnen evengoed schrijven $x_1^u =$ man en $x_2^u =$ vrouw.

[e] We hebben zonet de nieuwe notatie in detail geïllustreerd aan de hand van de nominale variabele Geslacht. We kunnen deze variabele echter niet gebruiken om de formule van het gemiddelde te illustreren, want het gemiddelde is enkel relevant voor variabelen gemeten op interval- of ratioschaal.

optellen en delen door de steekproefgrootte. Tabel 3.1 geeft deze verschillende stappen weer. De vierde kolom geeft de vermenigvuldiging weer en de laatste twee rijen de optelling en de optelling gedeeld door de steekproefgrootte.

De overeenkomst tussen de formule (3.2) en de Tabel 3.1 kan expliciet gemaakt worden door de formule uit te schrijven[f]:

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{p} f_i x_i^u \\
&= \frac{1}{n}\left(f_1 x_1^u + f_2 x_2^u + \ldots + f_p x_p^u\right) \\
&= \frac{1}{90}\Big((1\times 16) + (8\times 18) + \ldots + (1\times 74)\Big) \\
&= \frac{1}{90}\left(16 + 144 + \ldots + 74\right) \\
&= \frac{2818}{90} \\
&= 31.31.
\end{aligned}$$

Zoals verwacht komen we hetzelfde gemiddelde uit als de berekening via formule (3.1) op pagina 68. We hebben dus twee verschillende formules om het gemiddelde te berekenen: formule (3.1) op basis van de oorspronkelijke data en formule (3.2) op basis van de frequentieverdeling.

Het gemiddelde voor gegroepeerde data

We kunnen ook het gemiddelde berekenen voor gegroepeerde data door gebruik te maken van de gegroepeerde frequentieverdeling. Bij gegroepeerde data kunnen we de formule (3.2) niet onmiddellijk gebruiken omdat we de exacte waarden van de variabelen niet kennen (x_i^u is ongekend). We kennen enkel de klasse waartoe de waarde van een variabele behoort. Om hiervoor een oplossing te bieden, zullen we gebruik maken van een rekenregel. De rekenregel is de volgende: pas de formule (3.2) toe waarbij de waarden x_i^u vervangen worden door hun *klassenmiddens*:

! **Klassenmidden**. Het klassenmidden van een interval $]a, b]$ wordt gegeven door $\frac{a+b}{2}$.

! De klassenmiddens van de intervallen $]a, b]$, $[a, b]$, $]a, b[$, $[a, b]$ zijn gelijk.

[f]Om de leesbaarheid te vergroten, schrijven we soms het vermenigvuldigingsteken $\times$ expliciet uit.

i	Leeftijd (x_i^u)	Absolute frequentie (f_i)	Absolute frequentie $\times$ Leeftijd ($f_i x_i^u$)
1	16	1	16
2	18	8	144
3	19	6	114
4	20	7	140
5	21	5	105
6	22	5	110
7	23	1	23
8	24	5	120
9	25	3	75
10	26	5	130
11	28	1	28
12	29	5	145
13	30	3	90
14	31	2	62
15	32	3	96
16	34	1	34
17	35	1	35
18	36	2	72
19	37	5	185
20	38	1	38
21	40	1	40
22	41	1	41
23	43	1	43
24	44	1	44
25	46	1	46
26	47	1	47
27	48	1	48
28	49	1	49
29	50	1	50
30	51	1	51
31	52	1	52
32	53	2	106
33	54	1	54
34	55	1	55
35	61	1	61
36	63	1	63
37	64	1	64
38	68	1	68
39	74	1	74
	Som: $\sum_{i=1}^{p} f_i x_i^u$		2818
	Som gedeeld door n: $\frac{1}{n}\sum_{i=1}^{p} f_i x_i^u$		$\frac{2818}{90} = 31.31$

Tabel 3.1: De verschillende stappen om het gemiddelde te berekenen op basis van de frequentieverdeling.

Als we dit toepassen op het voorbeeld van de gegroepeerde data in Tabel 2.7 op pagina 46 dan is het klassenmidden van het interval $]15, 20]$ gelijk aan 17.5 ($\frac{15+20}{2}$), voor $]20, 25]$ is dit 22.5 ($\frac{20+25}{2}$), etc. Als we met a_i en b_i de grenzen aanduiden horende bij klasse i, dan kunnen we de formule voor het gemiddelde van gegroepeerde data schrijven als:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} f_i \frac{(a_i + b_i)}{2}. \tag{3.3}$$

Tabel 3.2 geeft de verschillende stappen weer om het gemiddelde te berekenen op basis van de gegroepeerde leeftijden: voor elke klasse berekenen we het klassenmidden, vervolgens vermenigvuldigen we dit klassenmidden met de absolute frequentie om dan de som te nemen en te delen door n.

De overeenkomst tussen de formule (3.3) en de Tabel 3.2 kan expliciet gemaakt worden door de formule uit te schrijven:

$$\begin{aligned}
\bar{x} &= \frac{1}{n} \sum_{i=1}^{p} f_i \frac{(a_i + b_i)}{2} \\
&= \frac{1}{n} \left(f_1 \frac{(a_1 + b_1)}{2} + f_2 \frac{(a_2 + b_2)}{2} + \ldots + f_p \frac{(a_p + b_p)}{2} \right) \\
&= \frac{1}{90} \left(22 \frac{(15 + 20)}{2} + 19 \frac{(20 + 25)}{2} + \ldots + 5 \frac{(60 + 90)}{2} \right) \\
&= \frac{1}{90} \left((22 \times 17.5) + (19 \times 22.5) + \ldots + (5 \times 75) \right) \\
&= \frac{1}{90} (385 + 427.5 + \ldots + 375) \\
&= \frac{2827.5}{90} \\
&= 31.42.
\end{aligned}$$

Het gemiddelde op basis van deze berekening ($\bar{x} = 31.42$) is verschillend van diegene op basis van de originele leeftijden ($\bar{x} = 31.31$). Dit is niet verwonderlijk: voor de gegroepeerde data hebben we enkel informatie gebruikt van de gegroepeerde frequentieverdeling en hebben we de rekenregel via de klassenmiddens gebruikt om het gemiddelde te bereken. Voor de berekeningen via formules (3.1) en (3.2) hebben we gebruik gemaakt van de ongegroepeerde data. De waarde van het gemiddelde hangt dus af of we gegroepeerde of ongegroepeerde data gebruiken. Dit was ook het geval voor de cumulatieve frequentiecurve uit paragraaf 2.5. Daar hebben we besloten dat de ongegroepeerde data meer informatie bevatten en daardoor onze voorkeur genieten. Bij het gemiddelde is dit ook zo. Indien mogelijk, trachten we dus altijd het gemiddelde te berekenen op basis

van de ongegroepeerde data. Er zullen echter situaties zijn waar enkel gegroepeerde data beschikbaar zijn en dan kunnen we niet anders dan formule (3.3) te gebruiken.

i	Klasse $(]a_i, b_i])$	Absolute frequentie (f_i)	Klassenmidden $(\frac{a_i+b_i}{2})$	Absolute frequentie × Klassenmidden $(f_i\frac{(a_i+b_i)}{2})$
1	]15,20]	22	17.5	385.0
2	]20,25]	19	22.5	427.5
3	]25,30]	14	27.5	385.0
4	]30,35]	7	32.5	227.5
5	]35,40]	9	37.5	337.5
6	]40,50]	8	45.0	360.0
7	]50,60]	6	55.0	330.0
8	]60,90]	5	75.0	375.0
		Som: $\sum_{i=1}^{p} f_i\frac{(a_i+b_i)}{2}$		2827.5
	Som gedeeld door n: $\frac{1}{n}\sum_{i=1}^{p} f_i\frac{(a_i+b_i)}{2}$			$\frac{2827.5}{90} = 31.42$

Tabel 3.2: De verschillende stappen om het gemiddelde te berekenen van gegroepeerde data.

3.1.2 De mediaan

Naast het gemiddelde is *de mediaan* ook een populaire centrummaat. Symbolisch wordt de mediaan van een variabele X geschreven als md_X.

Voor de mediaan is het niet eenvoudig om de maat uit te drukken als een wiskundige formule. Informeel is de mediaan de middelste waarde nadat we de waarden van een variabele van klein naar groot geordend hebben. Meer formeel kunnen we ze omschrijven als volgt:

! **De mediaan** van een variabele X in een steekproef is de waarde md_X waarvoor geldt dat:

- niet meer dan de helft van de elementen in de steekproef een waarde kleiner dan md_X hebben

EN

- niet meer dan de helft van de elementen in de steekproef een waarde groter dan md_X hebben.

We starten met een eenvoudig voorbeeld ter illustratie. We hebben in een steekproef de lengte van vijf personen gemeten in centimeter. De waarden zijn 151, 174, 183, 152, 168. Om de mediaan te vinden, ordenen we eerst de waarden van klein naar groot: 151, 152, 168, 174, 183. Nu bekijken we de waarde van de middelste persoon: er zijn vijf personen dus de derde is de middelste. De waarde van de derde persoon is 168 en vormt de mediaan. Er zijn twee personen die kleiner zijn dan 168 cm en twee personen die groter zijn. De formele definitie klopt: niet meer dan de helft van de personen heeft een waarde kleiner dan de mediaan en niet meer dan de helft heeft een waarde groter dan de mediaan.

! **Meetniveau**. Omdat de mediaan gebruik maakt van een ordening, is ze enkel zinnig voor ordinale, interval- en ratiovariabelen.

Stel nu dat we een zesde persoon meten met een lengte van 172 cm. De geordende waarden zijn nu 151, 152, 168, 172, 174, 183. Nu is er geen middelste persoon: voor persoon 3 zijn er 2 kleiner en 3 groter, terwijl voor persoon 4 er 3 kleiner zijn en 2 groter. Dit komt omdat we een even aantal personen hebben. Het volgt dat er nu twee waarden zijn die voldoen aan de definitie van de mediaan, namelijk 168 (lengte persoon 3) en 172 (lengte persoon 4). Inderdaad, er zijn twee waarden (dus niet meer dan de helft) kleiner dan 168 en 3 waarden (dus niet meer dan de helft) groter. Idem voor 172: er zijn 3 waarden kleiner en 2 waarden groter. Om nu toch een unieke waarde te hebben voor de mediaan, wordt volgende regel toegepast:

! Indien verschillende waarden voldoen aan de definitie van de mediaan, wordt de mediaan gelijkgesteld aan het rekenkundig gemiddelde van deze waarden.

Voor de steekproef met 6 personen is de mediaan volgens deze rekenregel gelijk aan $\frac{168+172}{2} = 170$ en schrijven we $md_X = 170$.

! **Meetniveau**. Indien de mediaan bekomen wordt door het rekenkundig gemiddelde te nemen, is ze enkel zinnig voor interval- en ratiovariabelen.

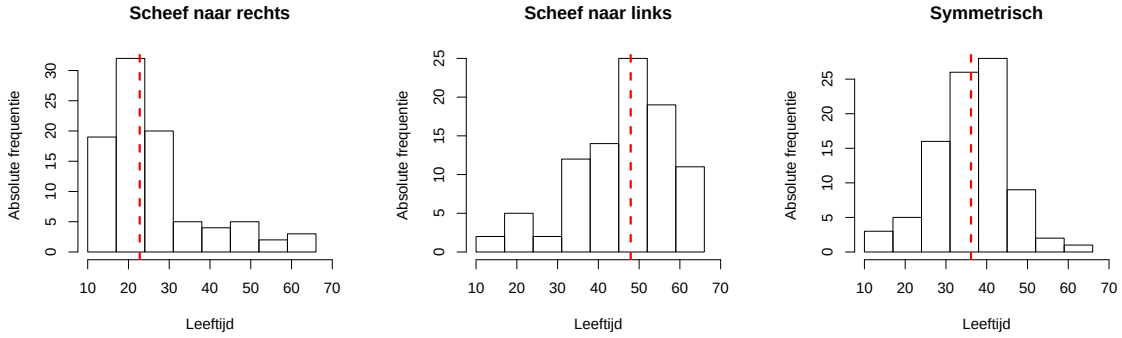
Om de mediaan te bepalen van de variabele Leeftijd uit Tabel 2.6 op pagina 45, moeten we eerst de data ordenen van klein naar groot. Dit wordt weergegeven in Tabel 3.3. Omdat er 90 personen in de steekproef zijn en dit getal even is, zullen we de leeftijd moeten bekijken van de persoon op de 45e plaats en de persoon op de 46e plaats. Ze zijn beiden 26 jaar. Aangezien dit geen verschillende waarden zijn, moeten we het rekenkundig gemiddelde niet berekenen en is de mediaan gelijk aan 26 jaar: $md_X = 26$.

16	18	18	18	18	18	18	18	18	19
19	19	19	19	19	20	20	20	20	20
20	20	21	21	21	21	21	22	22	22
22	22	23	24	24	24	24	24	25	25
25	26	26	26	26	26	28	29	29	29
29	29	30	30	30	31	31	32	32	32
34	35	36	36	37	37	37	37	37	38
40	41	43	44	46	47	48	49	50	51
52	53	53	54	55	61	63	64	68	74

Tabel 3.3: De geordende leeftijden van de 90 personen in de steekproef.

De mediaan kan je ook afleiden uit de cumulatieve frequentieverdeling uit Tabel 2.9 op pagina 54. Aangezien er 41 personen jonger zijn dan 26 en 46 personen jonger of gelijk aan 26 zijn, zijn personen op de 45e en 46e plaats beiden 26 jaar. De mediaan is dus 26.

Figuur 3.3 toont de histogrammen van de drie hypothetische steekproeven met aanduiding van de mediaan. Analoog zoals bij het gemiddelde, ligt de mediaan ongeveer centraal.



Figuur 3.3: De histogrammen van de variabele Leeftijd bij drie verschillende steekproeven met aanduiding van de mediaan.

Voor gegroepeerde data kunnen we de mediaan berekenen door gebruik te maken van een rekenregel. Eerst moeten we de klasse bepalen waartoe de mediaan behoort. Uit de cumulatieve frequentieverdeling van de gegroepeerde leeftijd in Tabel 2.10 op pagina 58 zien we dat personen 45 en 46 (indien we de leeftijden ordenen) behoren tot de leeftijdsklasse]25, 30]. De mediaan ligt dus tussen 25 jaar (25 niet inbegrepen) en 30 jaar (30 inbegrepen). De klasse]25, 30] wordt ook de *mediane klasse* genoemd. De

mediaan kan nu berekend worden met volgende formule:[g]

$$md_X = a + \frac{(\frac{n}{2} - c)(b - a)}{d},$$

waar

- a: de ondergrens van de mediane klasse.
- b: de bovengrens van de mediane klasse.
- c: de cumulatieve absolute frequentie van de klasse net kleiner dan de mediane klasse.
- d: de absolute frequentie van de mediane klasse.
- n: de steekproefgrootte.

Uit Tabel 2.10 lezen we dus af: $a = 25$, $b = 30$, $c = 41$ (dit is de cumulatieve frequentie van de klasse]20, 25]). Uit de tabel van de absolute frequenties (Tabel 2.7 op pagina 46) lezen we af dat $d = 14$. De steekproefgrootte is 90 zodat $n = 90$. Nu we alle getallen bepaald hebben, kunnen we de mediaan berekenen:

$$md_X = 25 + \frac{(\frac{90}{2} - 41)(30 - 25)}{14} = 26.43.$$

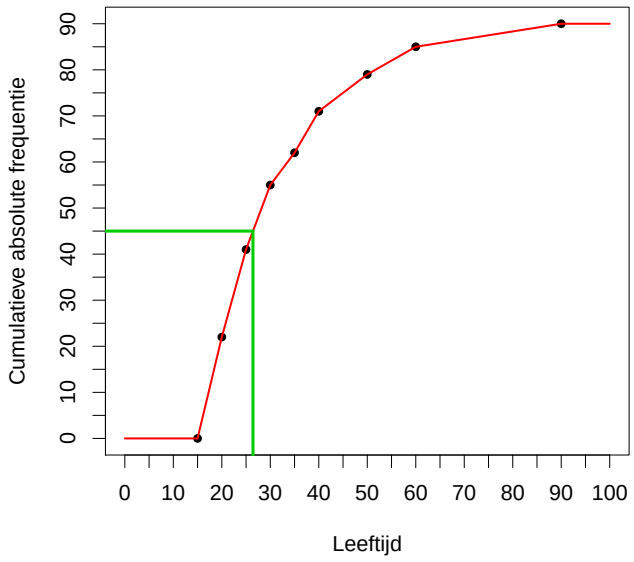
Figuur 3.4 toont het cumulatief frequentiediagram met aanduiding van de mediaan berekend volgens voorgaande formule. De groene verticale lijn wordt getrokken bij de waarde 26.43. De horizontale groene lijn toont het snijpunt tussen de verticale as en de curve en komt overeen met cumulatieve absolute frequentie van 45.

De mediaan berekend op basis van de gegroepeerde leeftijden ($md_X = 26.43$) is verschillend van de mediaan op basis van de originele leeftijden ($md_X = 26$). Dit was ook het geval voor het gemiddelde en kan op dezelfde wijze verklaard worden: bij de gegroepeerde data hebben we minder informatie en moeten we rekenregels gebruiken om de mediaan te bekomen.

Illustratie in R

De mediaan kan in R bekomen worden via `median()`.

[g]Het aantonen hoe deze formule tot stand komt, vormt geen onderdeel van de cursus.



Figuur 3.4: De cumulatieve frequentiecurve voor gegroepeerde data met aanduiding van de mediaan. Op de grafiek komt de mediaan overeen met een cumulatieve frequentie van 45.

```
> median(DataIAT$Leeftijd)
```

```
[1] 26
```

3.1.3 De modus

De modus is een derde centrummaat.

! De **modus** (symbool *mo*) is de klasse of de waarde met de grootste frequentie. Als er meerdere dergelijke klassen of waarden zijn, dan zijn er meerdere *modi*.

Als we kijken naar de absolute frequenties van de variabele Leeftijd in Tabel 2.9 op pagina 54, zien we dat 18 jaar de modus is: er zijn 8 personen van 18 jaar en dit is het hoogste aantal per leeftijd. Omdat er hier maar 1 modus is, wordt de verdeling *unimodaal* genoemd. Indien een verdeling twee modi heeft, wordt ze *bimodaal* genoemd.

Voor de frequentieverdeling van de gegroepeerde data in Tabel 2.7 op pagina 46 vormt de klasse]15, 20] de *modale klasse*: ze bevat de meeste personen. Via een histogram

met gelijke klassenbreedtes kunnen we ook op het zicht de modus bepalen: ze komt overeen met de klasse van de hoogste rechthoek. Voor Figuur 2.9 op pagina 50 zien we dat de modale klasse]20, 30] is.

! **Meetniveau**. De modus is enkel afhankelijk van de waarden van de variabelen en het aantal keer dat een waarde voorkomt. Er wordt dus geen gebruik gemaakt van een ordening. De modus is daarom zinnig voor nominale, ordinale, interval- en ratiovariabelen.

Omdat de modus zinnig is voor nominale variabelen, kunnen we ze ook toepassen op de variabele Geslacht. Uit Tabel 2.3 op pagina 38 lezen we af dat er meer vrouwen zijn dan mannen, dus de modus is 'vrouw'. De modus wordt vooral gebruikt bij discrete of gegroepeerde variabelen.

Illustratie in R

Voor de variabele Geslacht kunnen we via `table()` de frequenties berekenen en zien we onmiddellijk dat 'vrouw' de modus is omdat ze de grootste frequentie heeft:

```
> table(DataIAT$Geslacht)

  man vrouw
   36    54
```

Analoog voor de variabele Leeftijd: hier is de de modus 18 jaar.

```
> table(DataIAT$Leeftijd)

16 18 19 20 21 22 23 24 25 26 28 29 30 31 32 34 35 36 37 38 40 41 43
 1  8  6  7  5  5  1  5  3  5  1  5  3  2  3  1  1  2  5  1  1  1  1
44 46 47 48 49 50 51 52 53 54 55 61 63 64 68 74
 1  1  1  1  1  1  1  1  2  1  1  1  1  1  1  1
```

3.1.4 Gevoeligheid aan outliers

De 3 centrummaten (gemiddelde, mediaan en modus) worden bekomen op basis van de waarden van de variabelen. Nu kan het zijn dat er in een steekproef waarden voorkomen die zeer groot of zeer klein zijn in vergelijking met de meeste andere waarden. Als voorbeeld nemen we het verschil van de reactietijden zoals besproken in Hoofdstuk 2 op pagina 63, waar:

Verschil in reactietijd = congruente reactietijd - incongruente reactietijd

Figuur 3.5 toon het histogram van deze variabele voor de blanke personen. Het is duidelijk dat er een persoon is die een waarde heeft die zeer verschillend is van de waarden van de andere personen. Deze persoon heeft een verschil in reactietijd van ongeveer -2000 milliseconden, terwijl dit voor de andere tussen -500 en 600 ligt. De extreme waarde wordt een *outlier* genoemd.

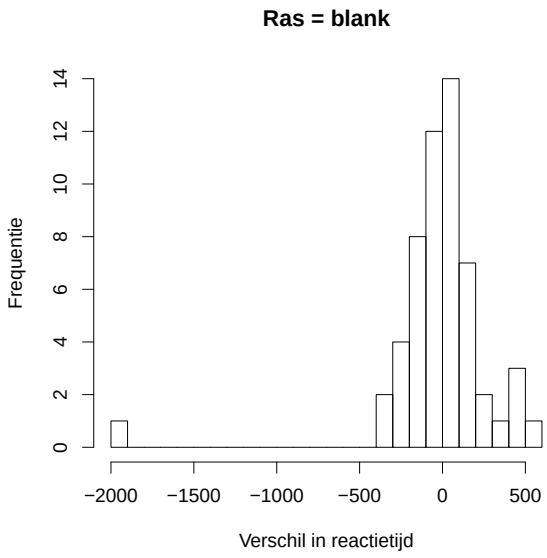
! **Outliers** (ook uitschieters genoemd) zijn waarden die ver verwijderd zijn van de overige waarden van een variabele.

Het bepalen of een waarde een outlier is, is subjectief: er wordt in de voorgaande definitie niet gespecificeerd wat *ver verwijderd* precies is. Verder in de cursus zullen we vuistregels zien die ons zullen helpen om outliers te detecteren.

Outliers kunnen bepaalde centrummaten sterk beïnvloeden. Het (rekenkundig) gemiddeld verschil in reactietijd is gelijk aan $\bar{x} = -17$ milliseconden. Dit geeft aan dat de blanken in de steekproef gemiddeld een snellere reactietijd hebben bij de congruente opdrachten[h]. Indien we de persoon met de outlier echter weglaten uit de data en het gemiddelde opnieuw berekenen, bekomen we $\bar{x} = 18$. Dit verschil is nu positief: na het verwijderen van de outlier kunnen we dus besluiten dat de blanken in de steekproef gemiddeld een *tragere* reactietijd hebben bij de congruente opdrachten. Het verwijderen van deze ene extreme observatie zorgt ervoor dat het gemiddelde redelijk wat wijzigt: hier van een (kleine) negatieve waarde naar een (kleine) positieve waarde. Dit geeft aan dat het *gemiddelde gevoelig is aan outliers*. Deze gevoeligheid wordt vaak als een nadeel ervaren, want 1 persoon in de steekproef kan de waarde van het gemiddelde sterk beïnvloeden.

Opgelet: om de impact van een outlier op het gemiddelde te illustreren, hebben we het gemiddelde berekend na het verwijderen van de outlier. Dit is louter ter illustratie

[h] De congruente opdrachten zijn diegene waar *Blanke mensen* en *Goed* samenhoren.



Figuur 3.5: Histogram van het verschil tussen de reactietijd van de congruente en incongruente opdrachten (bij een positief verschil is er een snellere reactietijd voor de incongruente opdrachten) voor de blanke personen. Uiterst links is er een outlier.

en dit wil niet zeggen dat je outliers moet verwijderen. Enkel indien outliers duidelijk foutief zijn (bijvoorbeeld een negatieve leeftijd), kan je ze verwijderen. Indien de outlier een correct gemeten waarde is, laat je ze best in de dataset.

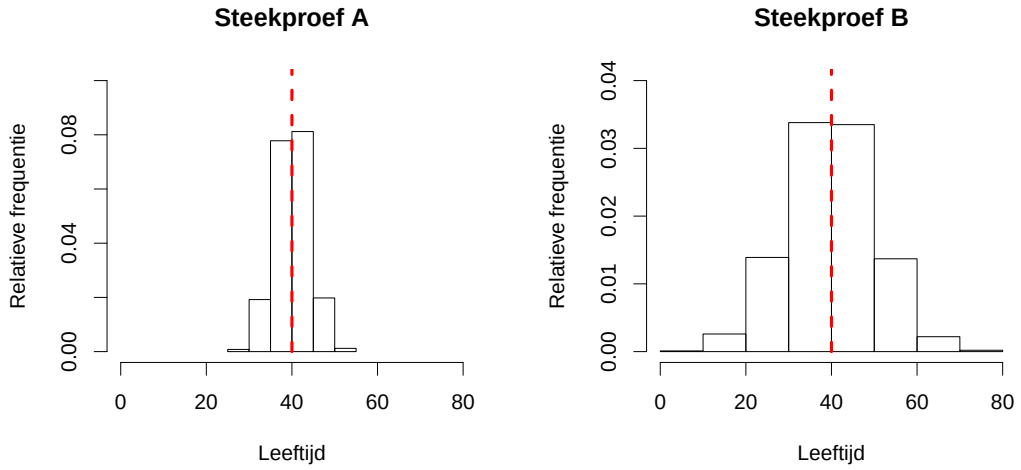
De mediaan van het verschil in reactietijd is $md_X = 2.15$ milliseconden. Indien we de outlier weglaten is dit $md_X = 4.2$ milliseconden. Bij de mediaan heeft de outlier dus veel minder invloed: *de mediaan is niet gevoelig aan outliers.*

De modus is ook niet gevoelig aan outliers. Dit komt omdat de modus overeenkomt met de waarde of klasse met de grootste frequentie. Outliers hebben typisch een lage frequentie en zullen dus de modus niet beïnvloeden.

3.2 Spreidingsmaten

Centrummaten laten ons toe data samen te vatten en geven een idee over het centrum van de verdeling. Maar een centrummaat alleen is niet voldoende. Figuur 3.6 toont ter illustratie het histogram van de leeftijden in twee steekproeven. Voor beide steekproeven is de gemiddelde leeftijd 40 jaar. Niettegenstaande de gemiddeldes gelijk zijn, is het

duidelijk dat de verdeling van de leeftijden verschillend is: bij steekproef A *varieert* de leeftijd minder: de meeste personen zijn rond de 40 jaar oud (het histogram is smaller). Bij steekproef B is er meer variatie: nu zijn er ook twintigers en zeventigers aanwezig (het histogram is breder). De breedte van het histogram hangt af van de *spreiding* van de waarden (de mate waarin de waarden onderling verschillen) en spreidingsmaten zullen toelaten deze spreiding numeriek te kwantificeren.



Figuur 3.6: Histogram van de leeftijd gemeten bij twee verschillende steekproeven (steekproef A en B). Voor beide steekproeven is de gemiddelde leeftijd 40 jaar (rode verticale stippellijn).

3.2.1 De variatiebreedte

De variatiebreedte is een zeer intuïtieve maat voor de spreiding: het is de afstand tussen de grootste en de kleinste waarde.

! De **variatiebreedte** v_X is gelijk aan:

- de grootste min de kleinste waarde voor ongegroepeerde data.
- de bovengrens van de laatste klasse min de ondergrens van de eerste klasse voor gegroepeerde data (wanneer de klassen van klein naar groot geordend zijn).

Als de variatiebreedte gelijk is aan nul, wil dit zeggen dat de grootste en kleinste waarde gelijk zijn. Dit is het geval wanneer er geen spreiding is. Een variatiebreedte groter

dan nul[i], wil zeggen dat er ten minste twee waarden van elkaar verschillen: er is dus spreiding. Dit geeft aan dat de variatiebreedte een maat is voor de spreiding.

! **Meetniveau**. De variatiebreedte maakt gebruik van een afstand tussen twee waarden, daardoor is ze enkel zinnig voor interval- en ratiovariabelen.

Tabel 3.3 op pagina 78 laat ons eenvoudig toe om de variatiebreedte te berekenen voor de variabele Leeftijd in de steekproef met 90 personen. De jongste persoon is 16 jaar en de oudste 74 jaar bijgevolg is $v_X = 74 - 16 = 58$. De variatiebreedte voor de gegroepeerde data in Tabel 2.10 op pagina 58 is gelijk aan $v_X = 90 - 15 = 75$.

De variatiebreedte voor steekproef A in Figuur 3.6 is 26, terwijl dit voor steekproef B gelijk is aan 63. De variatiebreedte geeft dus weer dat de spreiding niet gelijk is - iets wat we al visueel hadden vastgesteld, maar nu ook kunnen kwantificeren.

Illustratie in R

Om de variatiebreedte te berekenen moeten we het minimum en maximum van de variabele Leeftijd bepalen. In R kan dit via `min()` en `max()`:

```
> min(DataIAT$Leeftijd)

[1] 16

> max(DataIAT$Leeftijd)

[1] 74
```

De variatiebreedte is nu gelijk aan het verschil van deze waarden

```
> max(DataIAT$Leeftijd) - min(DataIAT$Leeftijd)

[1] 58
```

[i] De variatiebreedte kan nooit negatief zijn omdat het maximum nooit kleiner kan zijn dan het minimum.

3.2.2 De gemiddelde absolute afwijking

Naast de variatiebreedte bestaan er nog andere spreidingsmaten: de *gemiddelde absolute afwijking* bijvoorbeeld. We leggen eerst de intuïtie uit van deze spreidingsmaat.

Indien er spreiding is, zullen er waarden zijn die verschillen van het gemiddelde. Als we de notatie van pagina 39 gebruiken (waar x_i de waarde voorstelt van het i^{de} element in de steekproef) dan kunnen we het verschil tussen de i^{de} waarde en het gemiddelde schrijven als:

$$x_i - \bar{x}.$$

Voor elk van de n elementen in de steekproef kunnen we dit verschil berekenen:

$$x_1 - \bar{x}, \quad x_2 - \bar{x}, \quad \ldots \quad , \quad x_n - \bar{x}.$$

Hoe groter deze verschillen, hoe meer spreiding. Om op basis van deze n verschillen één getal te bekomen, is het verleidelijk om de som te nemen:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}).$$

Via het sommatieteken (zie pagina 68) kunnen we dit verkort schrijven als

$$\sum_{i=1}^{n} (x_i - \bar{x}).$$

Dit is echter geen goede maat voor spreiding omdat positieve verschillen (waarden waarvoor x_i groter is dan het gemiddelde) en negatieve verschillen (waarden waarvoor x_i kleiner is dan het gemiddelde) elkaar zullen opheffen. Je kan zelfs aantonen dat de som *altijd* nul zal zijn, ongeacht hoeveel spreiding er is.

Dit probleem kan eenvoudig worden opgelost door de absolute waarde te nemen van de verschillen[j]:

$$|x_1 - \bar{x}|, \quad |x_2 - \bar{x}|, \quad \ldots \quad , \quad |x_n - \bar{x}|.$$

Vervolgens kunnen we de som nemen om één getal te bekomen

$$\sum_{i=1}^{n} |x_i - \bar{x}|.$$

Dit is al een betere maat voor de spreiding, maar ze staat nog niet volledig op punt. Er wordt een som genomen van n positieve getallen (namelijk de getallen $|x_i - \bar{x}|$). Hoe

[j]De absolute waarde van een getal a wordt aangeduid als $|a|$ en is gelijk aan a als a positief is en gelijk aan $-a$ als a negatief is. Bijvoorbeeld $|3| = 3$, $|-3| = 3$, $|-18.7| = 18.7$, etc.

groter n, hoe groter deze som zal zijn, ook als de spreiding dezelfde blijft. Een goede spreidingsmaat zou enkel maar groter mogen worden als er meer spreiding is. Om dit op te lossen, zullen we de som delen door n:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

Deze spreidingsmaat wordt *de gemiddelde absolute afwijking* genoemd en heeft als symbool ga_X.

! De **gemiddelde absolute afwijking** van een variabele X in een steekproef wordt gegeven door:

$$ga_X = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

! **Meetniveau**. De gemiddelde absolute afwijking is enkel zinnig voor variabelen op interval- of ratioschaal.

De gemiddelde absolute afwijking maakt gebruik van het gemiddelde en van verschillen tussen waarden. Beide bewerkingen zijn zinloos voor ordinale en nominale variabelen, vandaar dat we ze enkel gebruiken bij interval- of ratiovariabelen.

Tabel 3.4 illustreert hoe je handmatig de gemiddelde absolute afwijking kan berekenen voor de variabele Leeftijd. In de eerste kolom staan de leeftijden van de 90 personen. In de tweede kolom berekenen we het verschil tussen de leeftijd en het gemiddelde (het gemiddelde is $\bar{x} = 31.31$). In de derde kolom nemen we de absolute waarde van dit verschil en ten slotte tellen we al deze getallen op en delen we door de steekproefgrootte $n = 90$. De eerste persoon is 40 jaar oud zodat $x_i - \bar{x} = 40 - 31.31 = 8.69$. Analoog voor de andere leeftijden.

Illustratie in R

Via het pakket `lsr` kunnen we `aad()` (wat staat voor *average absolute deviation*) gebruiken om de gemiddelde absolute afwijking te berekenen:

```
> library('lsr')
> aad(DataIAT$Leeftijd)

[1] 10.72741
```

i	Leeftijd x_i	Verschil $x_i - \bar{x}$	Absolute waarde van het verschil $\lvert x_i - \bar{x} \rvert$
1	40	8.69	8.69
2	21	-10.31	10.31
3	22	-9.31	9.31
4	22	-9.31	9.31
5	18	-13.31	13.31
6	29	-2.31	2.31
7	20	-11.31	11.31
8	38	6.69	6.69
9	21	-10.31	10.31
10	24	-7.31	7.31
11	25	-6.31	6.31
12	24	-7.31	7.31
13	37	5.69	5.69
14	20	-11.31	11.31
15	20	-11.31	11.31
16	19	-12.31	12.31
17	20	-11.31	11.31
18	24	-7.31	7.31
19	31	-0.31	0.31
20	47	15.69	15.69
21	20	-11.31	11.31
22	18	-13.31	13.31
23	44	12.69	12.69
24	49	17.69	17.69
25	48	16.69	16.69
26	22	-9.31	9.31
27	29	-2.31	2.31
28	32	0.69	0.69
29	20	-11.31	11.31
30	37	5.69	5.69
31	32	0.69	0.69
32	19	-12.31	12.31
33	26	-5.31	5.31
34	23	-8.31	8.31
35	32	0.69	0.69
36	18	-13.31	13.31
37	36	4.69	4.69
38	35	3.69	3.69
39	24	-7.31	7.31
40	28	-3.31	3.31
41	16	-15.31	15.31
42	53	21.69	21.69
43	29	-2.31	2.31
44	22	-9.31	9.31
45	53	21.69	21.69
46	18	-13.31	13.31
47	20	-11.31	11.31
48	55	23.69	23.69
49	41	9.69	9.69
50	31	-0.31	0.31
51	37	5.69	5.69
52	61	29.69	29.69
53	26	-5.31	5.31
54	21	-10.31	10.31
55	19	-12.31	12.31
56	74	42.69	42.69
57	18	-13.31	13.31
58	21	-10.31	10.31
59	68	36.69	36.69
60	36	4.69	4.69
61	34	2.69	2.69
62	19	-12.31	12.31
63	19	-12.31	12.31
64	25	-6.31	6.31
65	29	-2.31	2.31
66	52	20.69	20.69
67	18	-13.31	13.31
68	46	14.69	14.69
69	24	-7.31	7.31
70	29	-2.31	2.31
71	64	32.69	32.69
72	18	-13.31	13.31
73	30	-1.31	1.31
74	25	-6.31	6.31
75	51	19.69	19.69
76	18	-13.31	13.31
77	21	-10.31	10.31
78	22	-9.31	9.31
79	26	-5.31	5.31
80	37	5.69	5.69
81	26	-5.31	5.31
82	30	-1.31	1.31
83	26	-5.31	5.31
84	30	-1.31	1.31
85	50	18.69	18.69
86	37	5.69	5.69
87	54	22.69	22.69
88	63	31.69	31.69
89	19	-12.31	12.31
90	43	11.69	11.69
	$\frac{1}{n}\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert$		10.73

Tabel 3.4: De verschillende stappen om de gemiddelde absolute afwijking te berekenen.

3.2.3 De variantie en de standaarddeviatie

De gemiddelde absolute afwijking heeft wiskundig gezien één nadeel: er worden absolute waarden gebruikt. Deze absolute waarden zorgen ervoor dat bepaalde wiskundige eigenschappen van de gemiddelde absolute afwijking moeilijk te bestuderen zijn. Dit kan eenvoudig verholpen worden door de absolute waarden te vervangen door kwadraten: net als de absolute waarde zorgt het kwadraat ervoor dat negatieve getallen positief worden, zodat de verschillen tussen de waarden en het gemiddelde $(x_i - \bar{x})$ elkaar niet opheffen. Verschillend van de absolute waarde, zorgt het kwadrateren ervoor dat de eigenschappen van de spreidingsmaat eenvoudig te bestuderen te zijn[k].

De naam van deze spreidingsmaat is de *variantie*. Ze is veruit de meest populaire spreidingsmaat en heeft als symbool sn_X^2[l].

! De **variantie** van een variabele X in een steekproef wordt gegeven door:

$$sn_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.4}$$

Analoog als bij de andere spreidingsmaten wordt de variantie groter als er meer spreiding is.

! **Meetniveau**. De variantie is enkel zinnig voor variabelen gemeten op interval- of ratioschaal.

De *standaarddeviatie* is een spreidingsmaat die bekomen wordt door de vierkantswortel te nemen van de variantie:

! De **standaarddeviatie** van een variabele X in een steekproef wordt gegeven door:

$$sn_X = \sqrt{sn_X^2}. \tag{3.5}$$

De standaarddeviatie heeft dezelfde meeteenheid als de variabele, terwijl dit niet zo is voor de variantie. Voor de variabele Leeftijd is de variantie uitgedrukt in 'jaren

[k] Dit is vooral belangrijk voor de vervolgcursussen Statistiek, in deze cursus zullen we de wiskundige eigenschappen niet bestuderen.

[l] Het kwadraat in het symbool zal verder duidelijk worden alsook waarom we de afkorting *sn* gebruiken.

in het kwadraat' (omdat we de verschillen kwadrateren), terwijl de standaarddeviatie uitgedrukt is in 'jaren' (omdat we de vierkantswortel nemen).

Tabel 3.5 illustreert hoe je de variantie voor de variabele Leeftijd kan berekenen. In de eerste kolom staan de leeftijden van de 90 personen. In de tweede kolom berekenen we het verschil tussen de leeftijd en het gemiddelde (het gemiddelde niet afgerond is $\bar{x} = 31.31111$). In de derde kolom nemen we het kwadraat van dit verschil en tenslotte tellen we al deze getallen op en delen we door de steekproefgrootte $n = 90$. De eerste persoon is 40 jaar oud zodat $x_1 - \bar{x} = 40 - 31.31111 = 8.68889$. Vervolgens is $(x_1 - \bar{x})^2 = 8.68889^2 = 75.49681$. Dit ronden we af naar 75.50. Merk op dat we hier expliciet de niet afgeronde getallen hebben gebruikt bij de berekeningen. Analoog voor de andere leeftijden. We bekomen $sn_X^2 = 179.55$ en $sn_X = \sqrt{179.55} = 13.40$.

Net als bij het steekproefgemiddelde kunnen we de variantie ook berekenen op basis van een frequentieverdeling (zie pagina 72) door gebruik te maken van de formule:

$$sn_X^2 = \frac{1}{n} \sum_{i=1}^{p} f_i (x_i^u - \bar{x})^2. \tag{3.6}$$

Als oefeningen kan je zelf nagaan dat deze formule ook de waarde $sn_X^2 = 179.55$ zal geven.

Naast formule (3.4) bestaat er nog een andere formule om de variantie te berekenen met als symbool s_X^2. Bij deze formule zullen we delen door $n - 1$ i.p.v. door n, dus

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

Men deelt door $n - 1$ omdat deze formule gunstige eigenschappen heeft. Verder in de cursus komen we hierop terug (op pagina 208 staat de uitleg). Het belangrijkste is te onthouden dat er 2 formules zijn om de variantie te berekenen, elk met hun eigen symbool namelijk s_X^2 en sn_X^2 en analoog zijn er twee formules voor de standaarddeviatie $s_X = \sqrt{s_X^2}$ en $sn_X = \sqrt{sn_X^2}$. In de cursus zullen we vaak s_X^2 en s_X gebruiken en in mindere mate sn_X^2 en sn_X.

Toegepast op de variabele Leeftijd uit Tabel 3.5 kunnen we afleiden dat $s_X^2 = 181.57$ en $s_X = 13.47$.

Illustratie in R

Via `var()` kunnen de variantie berekenen. Let wel, R zal steeds s_X^2 berekenen i.p.v. sn_X^2.

i	Leeftijd x_i	Verschil $x_i - \bar{x}$	Kwadraat van het verschil $(x_i - \bar{x})^2$
1	40	8.69	75.50
2	21	-10.31	106.32
3	22	-9.31	86.70
4	22	-9.31	86.70
5	18	-13.31	177.19
6	29	-2.31	5.34
7	20	-11.31	127.94
8	38	6.69	44.74
9	21	-10.31	106.32
10	24	-7.31	53.45
11	25	-6.31	39.83
12	24	-7.31	53.45
13	37	5.69	32.36
14	20	-11.31	127.94
15	20	-11.31	127.94
16	19	-12.31	151.56
17	20	-11.31	127.94
18	24	-7.31	53.45
19	31	-0.31	0.10
20	47	15.69	246.14
21	20	-11.31	127.94
22	18	-13.31	177.19
23	44	12.69	161.01
24	49	17.69	312.90
25	48	16.69	278.52
26	22	-9.31	86.70
27	29	-2.31	5.34
28	32	0.69	0.47
29	20	-11.31	127.94
30	37	5.69	32.36
31	32	0.69	0.47
32	19	-12.31	151.56
33	26	-5.31	28.21
34	23	-8.31	69.07
35	32	0.69	0.47
36	18	-13.31	177.19
37	36	4.69	21.99
38	35	3.69	13.61
39	24	-7.31	53.45
40	28	-3.31	10.96
41	16	-15.31	234.43
42	53	21.69	470.41
43	29	-2.31	5.34
44	22	-9.31	86.70
45	53	21.69	470.41
46	18	-13.31	177.19
47	20	-11.31	127.94
48	55	23.69	561.16
49	41	9.69	93.87
50	31	-0.31	0.10
51	37	5.69	32.36
52	61	29.69	881.43
53	26	-5.31	28.21
54	21	-10.31	106.32
55	19	-12.31	151.56
56	74	42.69	1822.34
57	18	-13.31	177.19
58	21	-10.31	106.32
59	68	36.69	1346.07
60	36	4.69	21.99
61	34	2.69	7.23
62	19	-12.31	151.56
63	19	-12.31	151.56
64	25	-6.31	39.83
65	29	-2.31	5.34
66	52	20.69	428.03
67	18	-13.31	177.19
68	46	14.69	215.76
69	24	-7.31	53.45
70	29	-2.31	5.34
71	64	32.69	1068.56
72	18	-13.31	177.19
73	30	-1.31	1.72
74	25	-6.31	39.83
75	51	19.69	387.65
76	18	-13.31	177.19
77	21	-10.31	106.32
78	22	-9.31	86.70
79	26	-5.31	28.21
80	37	5.69	32.36
81	26	-5.31	28.21
82	30	-1.31	1.72
83	26	-5.31	28.21
84	30	-1.31	1.72
85	50	18.69	349.27
86	37	5.69	32.36
87	54	22.69	514.79
88	63	31.69	1004.19
89	19	-12.31	151.56
90	43	11.69	136.63
	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$		179.55

Tabel 3.5: Verschillende stappen om de variantie te berekenen.

```
> var(DataIAT$Leeftijd)

[1] 181.565
```

Via `sd()` kan de standaarddeviatie s_X bekomen worden:

```
> sd(DataIAT$Leeftijd)

[1] 13.47461
```

Je kan eenvoudig controleren dat de standaarddeviatie inderdaad gelijk is aan de vierkantswortel van de variantie via `sqrt()` (wat staat voor *square root*, wat Engels is voor vierkantswortel):

```
> sqrt(var(DataIAT$Leeftijd))

[1] 13.47461
```

3.2.4 De interkwartielafstand

De interkwartielafstand is een maat voor de spreiding op basis van *percentielen*.

! Voor een geheel getal k tussen 0 en 100, is het k-de percentiel (symbool P_k) het getal P_k waarvoor geldt dat

$$\frac{F(P_k)}{n} = \frac{k}{100}.$$

Dit is misschien een abstracte definitie, maar de betekenis van een percentiel is relatief eenvoudig. Het 10e percentiel P_{10}, bijvoorbeeld, is de waarde van een variabele, waarvoor 10% van de waarden hetzelfde of kleiner zijn. Inderdaad, $F(x)$ staat voor de cumulatieve absolute frequentie van x. Bijgevolg staat $F(P_{10})$ voor de cumulatieve absolute frequentie van P_{10}. Indien we delen door n, bekomen we de cumulatieve relatieve frequentie. Het 10e percentiel is dus de waarde van de variabele waarvoor 10% van de waarden hetzelfde of kleiner zijn.

Een bijzonder percentiel is de mediaan: ze is gelijk aan het 50e percentiel P_{50}. Inderdaad, de mediaan is de waarde waarvoor de helft (dus 50%) van de observaties hetzelfde of kleiner zijn, dus $P_{50} = md_X$.

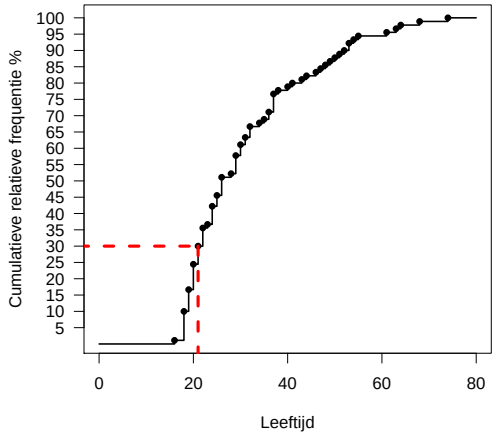
Via de cumulatieve frequentiecurve kunnen we de percentielen aflezen. Figuur 3.7 toont de cumulatieve relatieve frequentiecurve. Om het 30e percentiel te berekenen, trekken we eerst bij een waarde van 30 op de verticale as een horizontale lijn tot aan de curve, om vervolgens een verticale lijn te trekken naar de horizontale as. De waarde op de horizontale as ter hoogte van die lijn is het 30e percentiel. Voor Figuur 3.7 is $P_{30} = 21$: 30% van de mensen in de steekproef zijn 21 jaar of jonger. Het 30e percentiel P_{30} is dus het punt waarvoor geldt dat:

$$\frac{F(P_{30})}{n} = \frac{30}{100},$$

of

$$\frac{F(P_{30})}{n} \times 100\% = 30\%.$$

Analoog voor de andere percentielen.



Figuur 3.7: De cumulatieve relatieve frequentiecurve met aanduiding van het 30e percentiel (rode stippellijn): het snijpunt van de rode stippellijn met de horizontale as geeft het 30e percentiel P_{30}. Hier is dit 21 jaar, dus $P_{30} = 21$.

Er zijn drie bijzondere percentielen die vaak gebruikt worden: het *eerste kwartiel* P_{25}, het *tweede kwartiel* P_{50} en het *derde kwartiel* P_{75}. Het eerste kwartiel is de waarde zodat een kwart (25%) van alle waarden hetzelfde of kleiner zijn. Het tweede kwartiel is de waarde waarvoor de helft (twee keer een kwart) van alle waarden hetzelfde of kleiner

zijn, zoals eerder aangegeven (dus het tweede kwartiel is een andere benaming voor de mediaan). Het derde kwartiel is de waarde waarvoor driekwart (75%) van alle waarden hetzelfde of kleiner zijn.

De *interkwartielafstand* is een spreidingsmaat die gebruik maakt van de kwartielen.

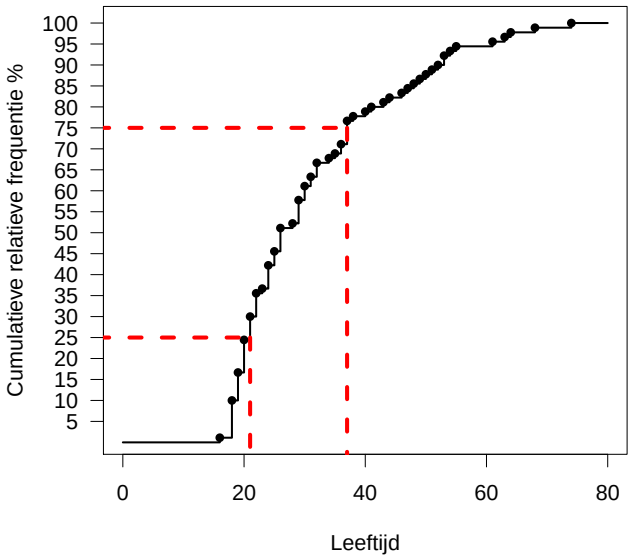
! De **interkwartielafstand** (symbool Q) is gelijk aan $P_{75} - P_{25}$, met P_{25} het eerste kwartiel en P_{75} het derde kwartiel.

De interkwartielafstand is dus het verschil tussen het derde kwartiel en het eerste kwartiel. Op basis van de kwartielen kan men ook het *interkwartielinterval* definiëren: dit is het interval $[P_{25}, P_{75}]$. Dit interval bevat 50% van alle waarden.

! **Meetniveau**. Omdat de interkwartielafstand gebruik maakt van een verschil tussen twee waarden, is ze enkel zinnig voor interval- en ratiovariabelen. Het interkwartielinterval is zinnig voor ordinale, interval- en ratiovariabelen.

Het afleiden van het eerste en derde kwartiel kan gebeuren aan de hand van de cumulatieve relatieve frequentiecurve. Vertrekkende van de verticale as trekken we een horizontale lijn ter hoogte van 25% (dit komt overeen met het eerste kwartiel) en 75% (dit komt overeen met het derde kwartiel) naar de frequentiecurve. Op de horizontale as kunnen we de bijhorende leeftijden aflezen, respectievelijk 21 en 37 jaar. De interkwartielafstand is bijgevolg gelijk aan $37 - 21 = 16$ en het interkwartielinterval is $[21, 37]$. Figuur 3.8 toont de cumulatieve relatieve frequentiecurve en hierbij zijn ook het eerste en derde kwartiel aangeduid.

Stel dat we de kwartielen willen berekenen van volgende lengtes: 151, 187, 164, 190, 172. We starten met de lengtes te ordenen: 151, 164, 172, 187, 190. Wat is nu het eerste kwart van 5 personen? Uit $5/4 = 1.25$ volgt dat het eerste kwartiel tussen 151 (lengte eerste persoon) en 164 (lengte tweede persoon) ligt, maar ze is niet gelijk aan het gemiddelde van deze waarden (omdat 1.25 niet in het midden van 1 en 2 ligt). Er bestaan verschillende rekenregels om het kwartiel te berekenen. Deze rekenregels behoren niet tot de inhoud van de cursus en wij zullen ons beperken tot het gebruik van de cumulatieve frequentieverdeling om de kwartielen (bij benadering) af te lezen, zoals geïllustreerd in Figuur 3.8 of we gaan gebruik maken van R om de kwartielen (en percentielen) te berekenen.



Figuur 3.8: De cumulatieve relatieve frequentiecurve met aanduiding van het eerste en derde kwartiel.

Illustratie in R

Via `quantile()` kunnen verschillende percentielen berekend worden, waaronder de kwartielen. Bij de berekening van de percentielen worden verschillende rekenregels gebruikt die we niet verder bespreken.

```
> quantile(DataIAT$Leeftijd)

  0%  25%  50%  75% 100%
  16   21   26   37   74
```

Via `IQR()` kunnen we dan de interkwartielafstand bekomen (IQR staat voor interquartile range):

```
> IQR(DataIAT$Leeftijd)

[1] 16
```

3.2.5 De spreidingsmaat d

De spreidingsmaat d wordt vooral gebruikt met nominale variabelen. De letter p stelt het aantal unieke waarden voor dat een variabele kan aannemen. Bij Geslacht is $p = 2$ bijvoorbeeld, omdat ze de waarden 'man' of 'vrouw' kan aannemen. Laat f_{mo} de frequentie van de modus zijn (een waarde of een klasse).

! De **spreidingsmaat** d wordt gedefinieerd door:

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

Stel nu dat de frequentie van de modus gelijk is aan de steekproefgrootte: $f_{mo} = n$. Dit wil zeggen dat er geen spreiding is: alle waarden van de variabele zijn gelijk aan de modus. Als we nu d berekenen, volgt:

$$d = \frac{1 - \frac{f_{mo}}{n}}{1 - \frac{1}{p}} = \frac{1 - \frac{n}{n}}{1 - \frac{1}{p}} = \frac{0}{1 - \frac{1}{p}} = 0.$$

Indien er dus geen spreiding is, is $d = 0$. Dit is wat je verwacht van een goede spreidingsmaat.

Stel nu dat er veel spreiding is, zodat de frequentie van de modus gelijk is aan n/p (de frequentie van de modus kan nooit kleiner zijn dan n/p, want dan zou ze geen modus meer zijn). Als we nu d berekenen, volgt:

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

Bij een maximale spreiding is $d = 1$.

! **Meetniveau**. Voor de berekening van de spreidingsmaat d wordt enkel gebruik gemaakt van de frequentie van de modus, de steekproefgrootte en het aantal unieke waarden. We mogen bijgevolg de modus gebruiken voor nominale, ordinale, interval- en ratiovariabelen.

Voor de variabele Geslacht in Tabel 2.3 op pagina 38, zien we dat $p = 2$, $f_{mo} = 54$ en $n = 90$ zodat:

$$d = \frac{1 - \frac{54}{90}}{1 - \frac{1}{2}} = \frac{0.4}{0.5} = 0.8.$$

3.2.6 Gevoeligheid aan outliers

Om de gevoeligheid aan outliers van de verschillende spreidingsmaten te onderzoeken, bekijken we opnieuw het verschil in reactietijd tussen de congruente en incongruente opdrachten van de blanken, zoals weergegeven in Figuur 3.5 op pagina 83.

We berekenen eerst de spreidingsmaten op basis van alle waarden (inclusief de outlier), en herhalen dit dan voor de waarden zonder de outlier. Indien er een groot verschil is tussen het resultaat met of zonder de outlier, besluiten we dat de maat gevoelig is aan outliers.

Tabel 3.6 geeft deze berekeningen weer en illustreert dat de variatiebreedte v_X, de gemiddelde absolute afwijking ga_X, de variantie s_X^2 en de standaarddeviatie s_X gevoelig zijn aan outliers. Van deze spreidingsmaten zijn voornamelijk de variatiebreedte en variantie zeer gevoelig aan outliers.

De interkwartielafstand Q anderzijds is niet gevoelig aan outliers.

	v_X	ga_X	s_X^2	s_X	Q
met outlier	2492.40	181.90	108724.76	329.73	201.46
zonder outlier	974.82	150.13	41527.03	203.78	199.79

Tabel 3.6: De spreidingsmaten voor het verschil in reactietijd bij de blanken op basis van alle waarden (inclusief de outlier) en op basis van de waarden zonder de outlier.

De spreidingsmaat d is vooral nuttig voor nominale en ordinale variabelen. We zullen dit niet berekenen voor het verschil in reactietijd omdat elke blanke persoon een unieke waarde heeft, waardoor d dicht bij 1 zal liggen.

Omdat de spreidingsmaat d afhangt van de frequentie van de modus, het aantal unieke waarden en de steekproefgrootte, is ze niet gevoelig aan outliers.

3.3 Boxplot

Op basis van de kwartielen en de interkwartielafstand, zoals besproken op pagina 92, kunnen we een nieuwe figuur maken: de *boxplot*. Een boxplot kan opgesteld worden zonder data te groeperen en is bijgevolg niet gebruikersafhankelijk. Dit is verschillend van het histogram.

Hoewel een boxplot er op het eerste zicht wat vreemd uitziet, is het een zeer bruikbare

figuur eenmaal je ze gewoon bent. Het zal ons in staat stellen om een idee te krijgen over de verdeling van de data en om *outliers* visueel vast te stellen. We gebruiken de volgende rekenregel om te bepalen of een waarde van een variabele een outlier is. We berekenen eerst de interkwartielafstand Q en vervolgens het verschil:

$$P_{25} - 1.5 \times Q,$$

dus het eerste kwartiel min 1.5 keer de interkwartielafstand. Alle waarden die *kleiner* zijn dan dit verschil, zijn outliers (volgens deze rekenregel). Vervolgens berekenen we de som:

$$P_{75} + 1.5 \times Q,$$

dus het derde kwartiel plus 1.5 keer de interkwartielafstand. Alle waarden die *groter* zijn dan deze som, zijn ook outliers. Outliers kunnen dus zowel grote als kleine waarden zijn. Het is uiteraard ook mogelijk dat een variabele *geen* outliers heeft.

In Figuur 3.9 construeren we stap voor stap een boxplot op basis van de variabele Leeftijd zoals gegeven in Tabel 2.6 op pagina 45. We starten door een verticale as te tekenen die de leeftijd voorstelt. Vervolgens zetten we naast de as een stip voor de leeftijd van elk van de 90 personen in de steekproef (Figuur A). De eerste persoon in Tabel 2.6 is 40 jaar, dus zetten we een stip ter hoogte van 40 naast de verticale as. De tweede persoon is 21, dus zetten we een stip ter hoogte van 21 naast de verticale as, etc.

Vervolgens gebruiken we de rekenregels om outliers te bepalen. Voor Leeftijd is het eerste kwartiel $P_{25} = 21$, het derde kwartiel $P_{75} = 37$ en de interkwartielafstand $Q = 37 - 21 = 16$. Bijgevolg is $P_{25} - 1.5 \times Q = 21 - 1.5 \times 16 = -3$. Dus alle personen jonger dan -3 jaar zijn outliers volgens de rekenregel. Het is evident dat er geen personen zijn met een negatieve leeftijd, dus zijn er geen outliers bij de kleine waarden. Voor de grote waarden berekenen we $P_{75} + 1.5 \times Q = 37 + 1.5 \times 16 = 61$. Alle personen die ouder zijn dan 61 jaar, zijn dus outliers. Als we kijken naar Tabel 2.6 zien we dat er 4 personen ouder zijn dan 61 jaar. In de Figuur 3.9 zullen we de stippen horende bij deze leeftijden rood kleuren (Figuur B).

In een volgende stap (Figuur C) tekenen we een horizontale lijn[m] bij de laagste stip die niet rood is (dus de kleinste waarde die geen outlier is). Aangezien er geen outliers zijn bij de kleine waarden, komt dit overeen met de minimumleeftijd (hier 16 jaar). We tekenen ook een horizontale lijn bij de hoogste stip die niet rood is (dus de grootste die geen outlier is).

Voor de kwartielen doen we iets gelijkaardigs: we tekenen een horizontale lijn ter hoogte van het eerste kwartiel $P_{25} = 21$, en ter hoogte van het derde kwartiel $P_{75} = 37$ (Figuur

[m] De breedte van de horizontale lijn mag je zelf kiezen.

D). De horizontale lijnen ter hoogte van de kwartielen verbinden we nu met elkaar met verticale lijnen zodat we een rechthoek bekomen (Figuur E).

In een volgende stap (Figuur F) verwijderen we alle stippen die niet rood zijn (dus alle waarden die geen outliers zijn). In een voorlaatste stap (Figuur G) tekenen we een verticale stippellijn van de onderste horizontale lijn tot het eerste kwartiel, en van het derde kwartiel tot de bovenste horizontale lijn. Deze verticale stippellijnen worden ook de *whiskers* of *snorharen* genoemd.

De laatste stap bestaat uit het tekenen van een horizontale lijn in de rechthoek ter hoogte van de mediaan $P_{50} = md_X = 26$. Deze laatste figuur stelt de boxplot voor van de variabele Leeftijd.

Een boxplot is handig omdat het visueel volgende informatie bevat:

- de mediaan (centrummaat): de horizontale lijn in de rechthoek.
- de interkwartielafstand (spreidingsmaat): de hoogte van de rechthoek.
- de outliers: de observaties die door bolletjes zijn aangeduid.

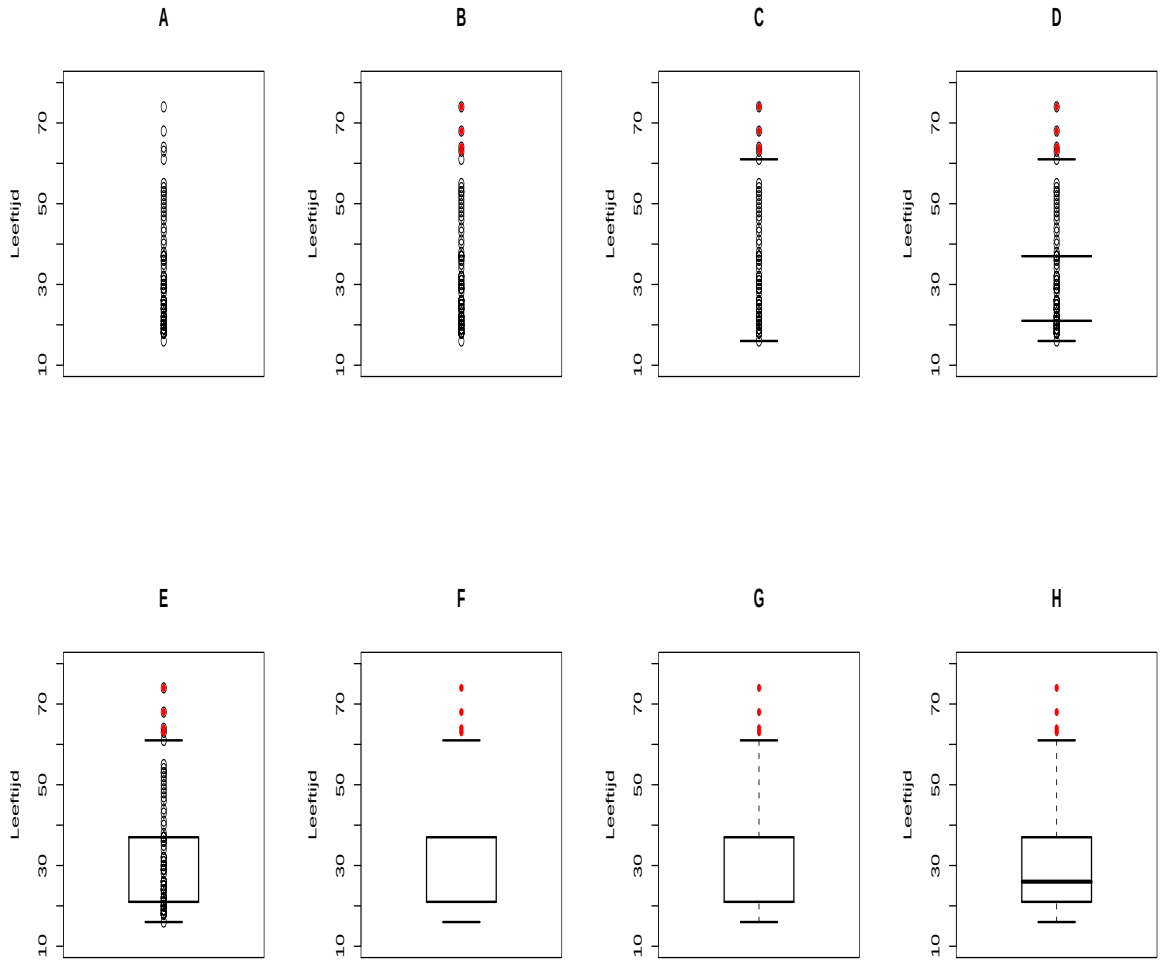
Figuur 3.10 toont de boxplots van Leeftijd in de 3 hypothetische steekproeven uit Figuur 2.10 op pagina 50. De kleine cirkels duiden hier de outliers aan.

Het is ook mogelijk om een boxplot *horizontaal* te tekenen i.p.v. *verticaal*. Figuur 3.11 geeft de horizontale variant van Figuur 3.10 die ontstaat door de verticale boxplots een kwartslag te draaien. Voor de scheve verdeling naar rechts (Figuur 3.11 links) zien we dat er redelijk wat outliers rechts zijn. Dit houdt steek: er zijn vooral jongeren in deze steekproef, dus de enkele personen die ouder zijn, worden als outliers beschouwd. Bij de scheve verdeling naar links is het net omgekeerd, terwijl er bij de symmetrische verdeling zowel enkele jongeren als ouderen outliers zijn.

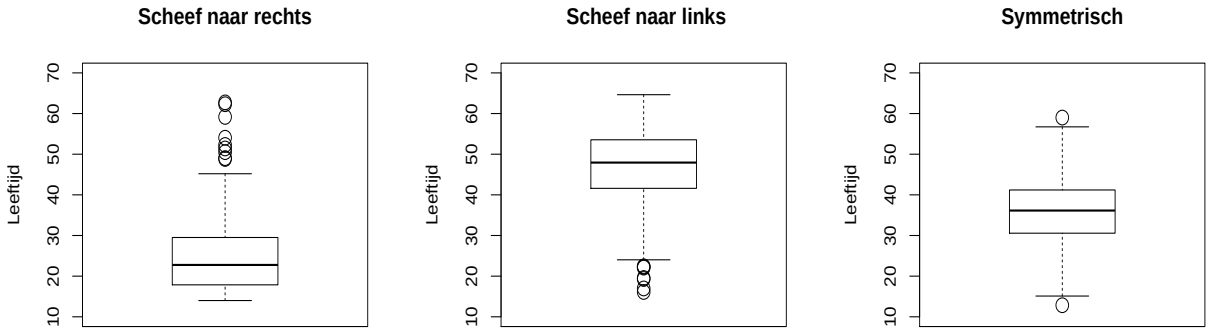
Illustratie in R

Via `boxplot()` bekomen we een boxplot in R.

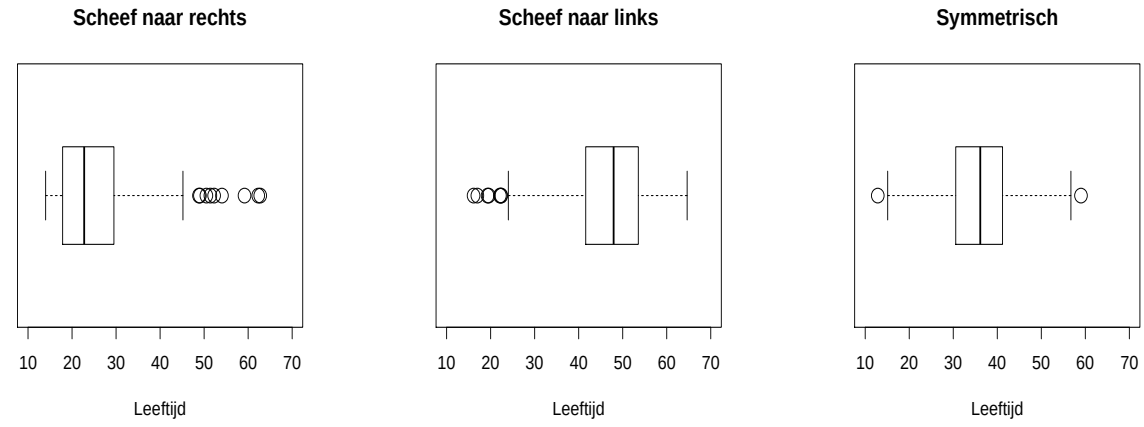
```
> boxplot(DataIAT$Leeftijd)
```



Figuur 3.9: Constructie van een boxplot voor de variabele Leeftijd. We verwijzen naar de tekst voor de uitleg.



Figuur 3.10: De boxplots van Leeftijd voor 3 verschillende (hypothetische) steekproeven.



Figuur 3.11: De horizontale weergave van de boxplots voor 3 verschillende (hypothetische) steekproeven.

3.4 Een voorbeeld: samenvatten van raciale voorkeur

We gebruiken nu de centrum- en spreidingsmaten alsook een boxplot om meer inzicht te krijgen in het onderzoek naar raciale voorkeur uit Hoofdstuk 2.

Omdat de gevoelens t.o.v. zwarten en blanken ordinale variabelen zijn, gebruiken we de mediaan en de modus als centrummaat, en d_X als spreidingsmaat. Op basis van het staafdiagram op pagina 61 zien we dat de modus van de gevoelens t.o.v. zwarten 5 is. Hetzelfde geldt voor de gevoelens t.o.v. blanken. De spreidingsmaat d voor de gevoelens t.o.v. zwarten is $d_X = 0.73$ terwijl dit t.o.v. blanken $d_X = 0.71$ is. De spreiding van de gevoelens t.o.v. blanken is dus iets lager. Dit komt overeen met wat we zien in Figuur 2.13 op pagina 61: de meeste scores bij de rechterfiguur gaan van 5 t.e.m. 10, terwijl de meeste scores voor de linkerfiguur 4 of hoger zijn. De mediaan van de gevoelens t.o.v. zwarten is 6 terwijl dit t.o.v. blanken 7 is.

Tabel 3.7 geeft deze waarden weer nadat we de data opgesplitst hebben per ras. Behalve voor de modus bij de blanken, zijn de centrummaten altijd hoger voor de gevoelens t.o.v. het eigen ras. Dit kan opgevat worden als een indicatie van een voorkeur voor eigen ras. Let wel, zowel de modus als de mediaan zijn minimum 5 (neutrale voorkeur). Een voorkeur voor het eigen ras impliceert dus geen afkeur voor het ander ras.

De spreidingsmaten zijn groter bij de gevoelens voor het eigen ras, wat wil zeggen dat deze gevoelens bij de personen in de steekproef meer variëren dan de gevoelens voor het ander ras.

	Ras = zwart		Ras = blank	
	Gevoelens t.o.v. zwarten	Gevoelens t.o.v. blanken	Gevoelens t.o.v. zwarten	Gevoelens t.o.v. blanken
modus	10	5	5	5
d_X	0.79	0.63	0.66	0.76
mediaan	8	5	5	7

Tabel 3.7: Samenvattende maten voor de gevoelens t.o.v. zwarten en blanken per ras.

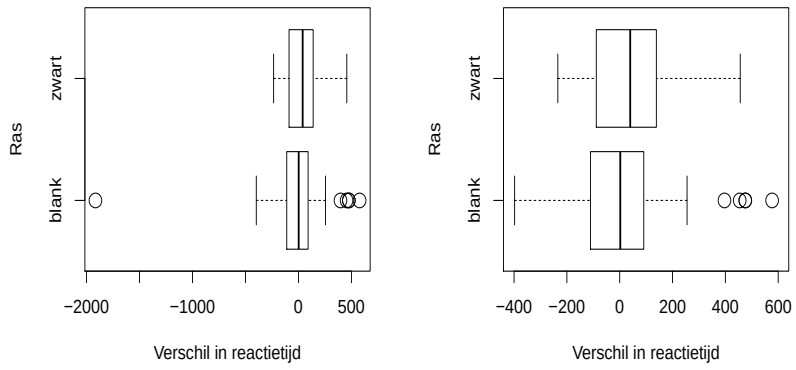
Figuur 3.12 toont de boxplots van het verschil in reactietijd van congruente en incongruente opdrachten, zoals besproken op pagina 65. Bij de blanken zijn er enkele personen voor wie hun reactietijd positieve outliers zijn, terwijl er één persoon is die als reactietijd een negatieve outlier heeft.

Zoals besproken op pagina 82, heeft deze grote negatieve waarde een substantieel effect

op het gemiddelde. Het gemiddeld verschil in reactietijd voor de blanken is -17.1 milliseconden, terwijl de mediaan 2.2 milliseconden bedraagt. Voor de zwarten liggen het gemiddelde en de mediaan dichter bij elkaar: $\bar{x} = 44.2$ en $md_X = 39.8$.

De blanken hebben dus gemiddeld een snellere reactietijd bij de congruente opdrachten. Dit komt voornamelijk door die negatieve outlier. De mediaan bedraagt 2.2 en geeft aan dat er geen sterke verschillen zijn tussen de reactietijden van de congruente en incongruente opdrachten.

Bij de zwarten is het iets anders: ze hebben gemiddeld een snellere reactietijd bij de incongruente opdrachten. Dit geldt ook voor de mediaan. Het verschil is echter klein (ongeveer 40 milliseconden). We hebben dus geen overtuigend bewijs gevonden dat er een sterke impliciete raciale voorkeur zou zijn.



Figuur 3.12: Boxplot van het verschil tussen de reactietijd van de congruente en incongruente opdrachten (bij een positief verschil is er een snellere reactietijd voor de incongruente opdrachten). Figuur rechts is een ingezoomde versie van de figuur links.

3.5 Samenvatting

In dit hoofdstuk hebben we verschillende samenvattingsmaten besproken. We hebben maten gezien die informatie geven over het centrum (of de locatie) van de verdeling:

- het gemiddelde.
- de mediaan.

- de modus.

Andere maten geven dan weer informatie over de spreiding van de verdeling:

- de variatiebreedte.
- de gemiddelde absolute afwijking.
- de variantie en standaarddeviatie.
- de interkwartielafstand.
- de spreidingsmaat d.

Voor de verschillende samenvattingsmaten hebben we het meetniveau besproken en de gevoeligheid aan outliers. Op basis van de kwartielen hebben we een nieuwe figuur opgesteld: de boxplot.

De boxplot en sommige samenvattingsmaten hebben we toegepast op de data afkomstig van de studie rond raciale voorkeur. Dit illustreert hoe we de methodes in de praktijk kunnen gebruiken.

Hoofdstuk 4

Samenhang tussen twee variabelen

In Hoofdstukken 2 en 3 hebben we statistische technieken (figuren, frequentietabellen, centrummaten en spreidingsmaten) gezien die ons toelaten de verdeling van een variabele beter te begrijpen. We hebben deze technieken toegepast op de variabelen *afzonderlijk*: we hebben één variabele per keer bekeken. Dit wordt *univariate statistiek* genoemd.

In dit hoofdstuk zullen we twee variabelen *gezamenlijk* bekijken. Dit zal ons toelaten om een mogelijke *samenhang* te bestuderen. Het gezamenlijk bestuderen van twee variabelen wordt *bivariate statistiek* genoemd.

De statistische technieken die aan bod komen in dit hoofdstuk worden geïllustreerd op data afkomstig uit onderzoek naar de samenhang tussen intelligentie en hersengrootte. Paragraaf 4.1 geeft meer informatie over deze studie. In paragrafen 4.2-4.6 bespreken we verschillende statistische methoden om de samenhang te onderzoeken en deze vormen de essentie van dit hoofdstuk. In paragraaf 4.7 gebruiken we bepaalde van deze methoden om inzicht te krijgen in de samenhang tussen intelligentie en hersengrootte.

4.1 Onderzoek naar intelligentie en hersengrootte

4.1.1 De onderzoeksvraag

Figuur 4.1 toont de titel van een artikel waarin men een antwoord tracht te geven op de vraag:

Zijn mensen met grotere hersenen slimmer?

Figuur 4.1: Zijn mensen met grotere hersenen slimmer? Bron: Bryner, J. (2012, 5 september). Are Big Brains Smarter? Geraadpleegd op 2 september, 2015, van `http://www.livescience.com/32142-are-big-brains-smarter.html`.

In dit hoofdstuk bepreken wij een studie uitgevoerd eind de jaren 1980 aan de Universiteit van Texas onder leiding van Professor Lee Willerman (Willerman et al., 1991).

We zullen de onderzoeksvraag eerst herformuleren, op het einde van het hoofdstuk zal duidelijk worden waarom:

Is er een verband tussen intelligentie en hersengrootte?

4.1.2 De populatie en de steekproef

In het wetenschappelijk artikel waar Willerman en zijn collega's hun bevinding rapporteren (Willerman et al., 1991), wordt de populatie niet expliciet beschreven. Hoewel het omschrijven van de populatie zeer waardevol is, wordt dit niet altijd gedaan in de wetenschappelijke literatuur.

De wijze waarop de steekproef is genomen, wordt wel uitvoerig besproken. Er werden 40 studenten geselecteerd uit een groep van eerste bachelorstudenten die een cursus "Introductie tot de Psychologie" volgden. Deze 40 studenten gaven aan geen voorgeschiedenis te hebben van alcoholisme, hersenschade, epilepsie of hartziekten. De 40

studenten waren zo gekozen zodat er evenveel mannen als vrouwen waren en evenveel met een 'gemiddeld' IQ (wat in de studie van Willerman gedefinieerd is als een totaal IQ van 103 of lager op de Wechsler Adult Intelligence Scale-Revised) als met een 'hoog' IQ (een totaal IQ van minstens 130)[a]. Deze onderverdeling volgens IQ laat de onderzoekers toe om de hersengrootte te vergelijken van de studenten met een gemiddeld IQ t.o.v. de studenten met een hoog IQ. De hersengrootte werd bepaald via een MRI-scan en is uitgedrukt in duizend pixel.

4.1.3 De data

Tabel 4.1 geeft de data weer van alle 40 studenten in de steekproef. Voor elke student hebben we de volgende variabelen gemeten (met het meetniveau tussen haakjes):

- Geslacht (nominaal).
- IQ: dit is het totaal IQ (interval).
- Verbaal IQ (interval).
- Performaal IQ (interval).
- Hersengrootte (ratio).
- IQ Groep (ordinaal).

De eerste student bijvoorbeeld, is een vrouw met een totaal IQ van 132, een verbaal IQ van 129, een performaal IQ van 124 en een hersengrootte van 833.9 duizend pixel. De variabele IQ Groep verdeelt de studenten in twee groepen op basis van hun totaal IQ:

$$\text{IQ Groep} = \begin{cases} \text{hoog IQ} & \text{als IQ} \geq 130 \\ \text{gemiddeld IQ} & \text{als IQ} \leq 103. \end{cases}$$

Illustratie in R

Analoog zoals op pagina 36 kunnen we de data inlezen via `read.table()`:

[a]Er werden geen studenten geselecteerd met een totaal IQ tussen 103 en 130.

Student	Geslacht	IQ	Verbaal IQ	Performaal IQ	Hersengrootte	IQ Groep
1	vrouw	132	129	124	833.90	hoog IQ
2	vrouw	96	100	90	878.90	gemiddeld IQ
3	man	97	107	84	905.90	gemiddeld IQ
4	man	90	96	86	880.00	gemiddeld IQ
5	vrouw	92	90	98	854.30	gemiddeld IQ
6	man	144	145	137	949.60	hoog IQ
7	man	103	96	110	997.90	gemiddeld IQ
8	vrouw	88	86	94	894.00	gemiddeld IQ
9	man	83	83	86	892.40	gemiddeld IQ
10	man	140	150	124	1001.10	hoog IQ
11	vrouw	138	136	131	991.30	hoog IQ
12	vrouw	99	90	110	928.80	gemiddeld IQ
13	man	80	77	86	889.10	gemiddeld IQ
14	man	89	91	89	935.90	gemiddeld IQ
15	vrouw	83	90	81	834.30	gemiddeld IQ
16	man	135	129	124	924.10	hoog IQ
17	man	100	96	102	945.10	gemiddeld IQ
18	vrouw	135	129	134	790.60	hoog IQ
19	man	89	93	84	904.90	gemiddeld IQ
20	vrouw	132	132	120	852.20	hoog IQ
21	vrouw	101	112	84	808.00	gemiddeld IQ
22	vrouw	137	132	134	951.50	hoog IQ
23	man	141	150	128	1079.50	hoog IQ
24	man	139	123	150	1038.40	hoog IQ
25	man	103	96	110	1062.50	gemiddeld IQ
26	vrouw	77	83	72	793.50	gemiddeld IQ
27	vrouw	133	132	124	816.90	hoog IQ
28	vrouw	83	71	96	865.40	gemiddeld IQ
29	vrouw	85	90	84	798.60	gemiddeld IQ
30	man	133	129	128	965.40	hoog IQ
31	vrouw	91	86	102	831.80	gemiddeld IQ
32	man	81	90	74	930.00	gemiddeld IQ
33	man	139	145	128	955.00	hoog IQ
34	vrouw	133	126	132	857.80	hoog IQ
35	man	133	114	147	955.50	hoog IQ
36	vrouw	133	129	128	948.10	hoog IQ
37	man	140	150	124	949.40	hoog IQ
38	vrouw	140	120	147	856.50	hoog IQ
39	vrouw	130	126	124	866.70	hoog IQ
40	man	141	145	131	935.50	hoog IQ

Tabel 4.1: De waarden van de variabelen van de leerlingen in de steekproef.

```
> # Eerst geven we de url waar de data staan een korte naam
> url.IQ <- "https://users.ugent.be/~jrdeneve/DataStatistiek1/DataIQ.txt"
> # Nu kunnen we via de url de data inlezen in R en de naam DataIQ geven
> DataIQ <- read.table(file = url.IQ)
```

Via `dim()` bekijken we het aantal rijen (= het aantal personen in de steekproef) en het aantal kolommen (= het aantal variabelen):

```
> dim(DataIQ)
```

```
[1] 40  6
```

Vervolgens tonen we de waarden van alle zes de variabelen voor de eerste zes studenten:

```
> head(DataIQ)
```

```
  Geslacht  IQ VIQ PIQ Hersengrootte      IQGroep
1    vrouw 132 129 124         833.9      hoog IQ
2    vrouw  96 100  90         878.9 gemiddeld IQ
3      man  97 107  84         905.9 gemiddeld IQ
4      man  90  96  86         880.0 gemiddeld IQ
5    vrouw  92  90  98         854.3 gemiddeld IQ
6      man 144 145 137         949.6      hoog IQ
```

Hier staat `VIQ` voor Verbaal IQ en `PIQ` voor Performaal IQ.

4.2 Bivariate frequentieverdeling

Om de bivariate frequentieverdeling te illustreren, delen we de hersengrootte op in drie groepen:]790,886],]886, 982] en]982, 1080]. Tabel 4.2 geeft de absolute frequentieverdeling van deze gegroepeerde variabele. Aangezien deze tabel enkel informatie bevat over één variabele (namelijk de gegroepeerde hersengrootte) noemen we ze ook de *univariate absolute frequentieverdeling*. Er zijn 16 studenten met een hersengrootte tussen 790 en 886, 18 met een hersengrootte tussen 886 en 982 en 6 met een hersengrootte tussen 982 en 1080.

Klasse	Absolute frequentie
]790, 886]	16
]886, 982]	18
]982, 1080]	6

Tabel 4.2: Absolute frequentieverdeling van de gegroepeerde hersengrootte.

We kunnen ook de univariate absolute frequentieverdeling opstellen voor de variabele IQ Groep, zie Tabel 4.3. Er zijn 20 studenten met een gemiddeld IQ en 20 studenten met een hoog IQ.

IQ Groep	Absolute frequentie
gemiddeld IQ	20
hoog IQ	20

Tabel 4.3: Absolute frequentieverdeling van de variabele IQ Groep.

Beide univariate frequentieverdelingen laten ons toe conclusies te formuleren over iedere variabele afzonderlijk: bv. hoeveel studenten er een hoog IQ hebben, hoeveel studenten een hersengrootte hebben tussen 886 en 982, etc. Ze laten ons echter niet toe om conclusies te formuleren over de *gezamenlijke* verdeling: bv. hoeveel studenten een hoog IQ én een hersengrootte tussen 886 en 982 hebben. De *bivariate absolute frequentieverdeling* zal toelaten deze twee variabelen gezamenlijk te bestuderen.

Tabel 4.4 geeft deze bivariate frequentieverdeling. Als we willen weten hoeveel studenten een hoog IQ hebben én een hersengrootte tussen 886 en 982, moeten we kijken naar de rij 'hoog IQ' en de kolom ']886, 982]'. De plaats waar deze rij en kolom elkaar kruisen, geeft de frequentie weer: hier 9. Er zijn dus 9 studenten met een hoog IQ en een hersengrootte tussen 886 en 982. Op een gelijkaardige manier kan je de frequenties bekomen van de andere combinaties.

Uit de bivariate verdeling kunnen we steeds de univariate verdeling afleiden. Inderdaad, de univariate verdeling voor IQ Groep kan je bekomen door per rij alle getallen op te tellen. Voor de waarde 'gemiddeld IQ' bekomen we $9 + 9 + 2 = 20$ wat inderdaad overeenstemt met de waarde in Tabel 4.3. Analoog voor de waarde 'hoog IQ': $7+9+4 = 20$. De univariate verdeling van de gegroepeerde hersengrootte bekomen we door de frequenties per kolom op te tellen, vb. voor ']790, 886]' is dit $9+7 = 16$ wat overeenkomt met de waarde in Tabel 4.2. Analoog voor de andere waarden. Als we de univariate verdelingen bepalen op basis van de bivariate verdeling, spreken we soms ook over de *marginale verdelingen*.

De bivariate verdeling bevat meer informatie dan de univariate (of marginale) verde-

	Klassen Hersengrootte		
IQ Groep	]790, 886]	]886, 982]	]982, 1080]
gemiddeld IQ	9	9	2
hoog IQ	7	9	4

Tabel 4.4: Bivariate absolute frequentieverdeling van IQ Groep en de gegroepeerde Hersengrootte.

lingen omdat we de univariate verdelingen kunnen afleiden uit de bivariate, maar niet omgekeerd.

De bivariate frequentieverdeling kan ons ook inzicht verschaffen in de samenhang tussen hersengrootte en intelligentie. Er zijn 9 studenten met een gemiddeld IQ en een 'kleine' hersengrootte (met 'kleine' bedoelen we de klasse]790, 886]), terwijl er maar 2 studenten zijn met een gemiddeld IQ en een 'grote' hersengrootte (met 'grote' bedoelen we de klasse]982, 1080]). Voor de studenten met een hoog IQ zijn deze aantallen respectievelijk 7 en 4.

Van de studenten met een kleine hersengrootte zijn er dus meer met een gemiddeld IQ, terwijl er van de studenten met een grote hersengrootte meer zijn met een hoog IQ. Er is bijgevolg een (zwakke) aanwijzing dat er een samenhang is tussen hersengrootte en IQ. De getallen zijn echter niet overtuigend en hangen ook af van het groeperen van de variabelen. Indien IQ en/of hersengrootte in andere klassen worden onderverdeeld, kunnen de conclusies wijzigen. Tabel 4.5 illustreert dit: hier hebben we geopteerd om de hersengrootte op te delen in 2 klassen:]790, 860] en]860, 1080]. De bivariate frequentieverdeling toont dat er evenveel studenten met een gemiddeld IQ zijn als met een hoog IQ en dit voor beide klassen. Op basis van deze frequentieverdeling zou men bijgevolg besluiten dat er geen samenhang is. De conclusies kunnen dus wijzigen door de data te hergroeperen. Deze subjectiviteit is vaak onwenselijk en kan vermeden worden door de samenhang te bestuderen via bijvoorbeeld een spreidingsdiagram en correlatiecoëfficiënten. We bespreken dit in de volgende paragrafen.

	Klassen hersengrootte	
IQ Groep	]790,860]	]860,1080]
gemiddeld IQ	6	14
hoog IQ	6	14

Tabel 4.5: Bivariate absolute frequentieverdeling van IQ Groep en de gegroepeerde Hersengrootte met twee klassen.

4.3 Spreidingsdiagram

Om het spreidingsdiagram te illustreren, zullen we eerst het Verbaal IQ onderzoeken. Later komen we dan terug op het (totaal) IQ en het Performaal IQ.

Tabel 4.6 geeft voor de 40 studenten het Verbaal IQ en Hersengrootte weer. Het spreidingsdiagram is een figuur die ons zal toelaten de samenhang tussen deze twee variabelen te visualiseren.

Figuur 4.2 toont aan hoe we een spreidingsdiagram kunnen bekomen. Op de horizontale as zetten we Hersengrootte uit en op de verticale as Verbaal IQ[b]. Uit Tabel 4.6 lezen we af dat de eerste student een hersengrootte van 833.9 heeft en een verbaal IQ van 129. We tekenen nu een punt ter hoogte van 833.9 op de horizontale as en ter hoogte van 129 op de verticale as (zoals aangegeven door de rode stippellijn in Figuur 4.2 linksboven). Dit herhalen we nu voor de tweede persoon in de steekproef (Figuur 4.2 rechtsboven). Dit doen we tot we alle 40 personen in de steekproef afgelopen hebben (Figuur 4.2 linksonder). Figuur 4.2 rechtsonder geeft tenslotte het finale spreidingsdiagram.

Om voeling te krijgen met hoe een spreidingsdiagram er kan uitzien, bekijken we volgende 3 eenvoudige experimenten om de verschillende soorten samenhang te illustreren:

- *Perfecte positieve samenhang.* Bij dit experiment noteren we voor verschillende personen de schoenmaat van de linkervoet (variabele 1) en de rechtervoet (variabele 2). Het is duidelijk dat er een samenhang is tussen beiden variabelen: bv. personen met een kleine schoenmaat links hebben ook een kleine schoenmaat rechts. Figuur 4.3 (links) toont het bijhorend spreidingsdiagram: de punten gaan van linksonder tot rechtsboven en liggen op een rechte. Dit wordt een perfecte positieve[c] samenhang genoemd.

- *Perfecte negatieve samenhang.* In een andere experiment tanken we een auto vol en rijden we aan een constante snelheid. Om de 20 kilometer noteren we de afgelegde afstand (variabele 1) en het aantal liter benzine in de tank (variabele 2). Tussen beide variabelen is er een samenhang: hoe meer afstand we afleggen, hoe minder benzine er in de tank zal zitten. Figuur 4.3 (midden) toont het bijhorend spreidingsdiagram: de punten gaan van linksboven naar rechtsonder en liggen op een rechte. Dit wordt een perfecte negatieve samenhang genoemd.

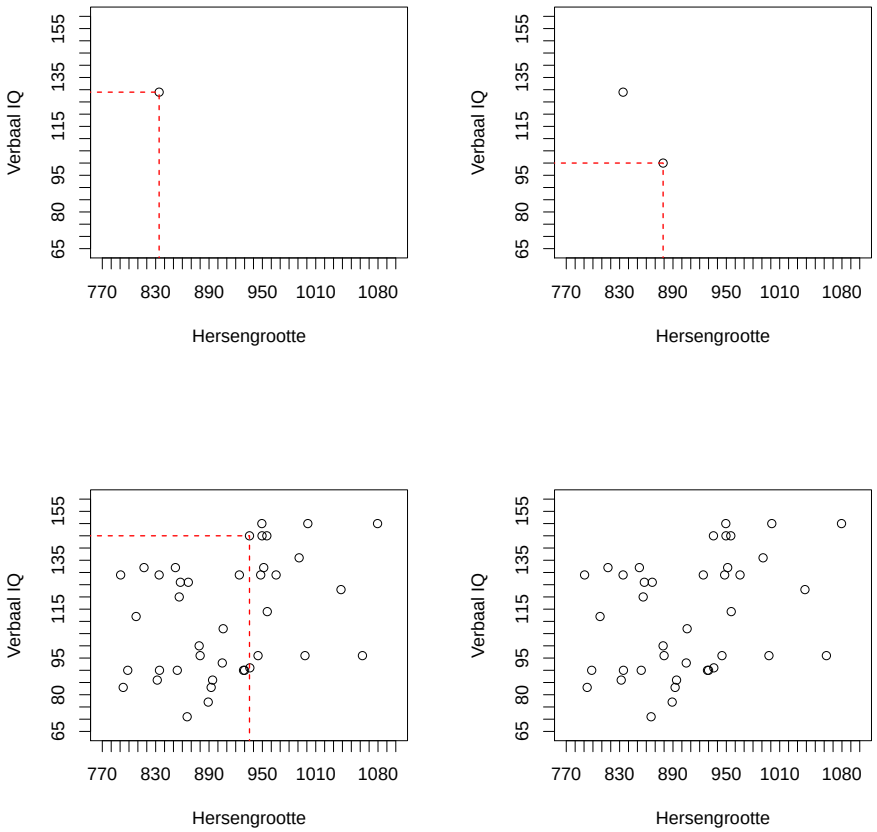
- *Geen samenhang.* Bij dit laatste experiment meten we gedurende een jaar de

[b] Je kan de keuze van variabelen ook omwisselen: Verbaal IQ op de horizontale as en Hersengrootte op de verticale as.

[c] Wat verder in de syllabus zal het duidelijk worden waarom we ze 'positief' noemen.

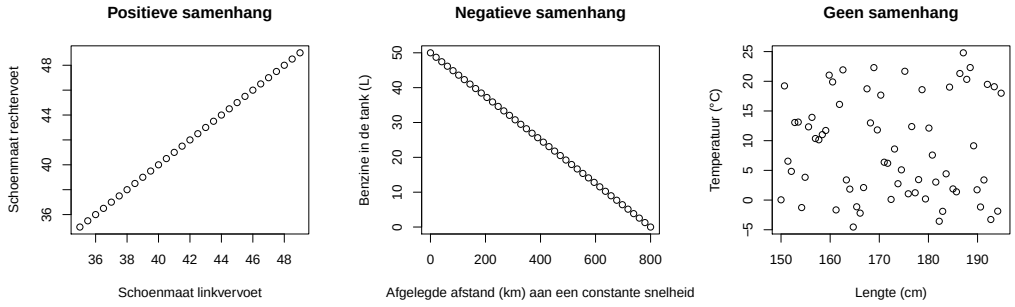
Student	Verbaal IQ	Hersengrootte
1	129	833.9
2	100	878.9
3	107	905.9
4	96	880.0
5	90	854.3
6	145	949.6
7	96	997.9
8	86	894.0
9	83	892.4
10	150	1001.1
11	136	991.3
12	90	928.8
13	77	889.1
14	91	935.9
15	90	834.3
16	129	924.1
17	96	945.1
18	129	790.6
19	93	904.9
20	132	852.2
21	112	808.0
22	132	951.5
23	150	1079.5
24	123	1038.4
25	96	1062.5
26	83	793.5
27	132	816.9
28	71	865.4
29	90	798.6
30	129	965.4
31	86	831.8
32	90	930.0
33	145	955.0
34	126	857.8
35	114	955.5
36	129	948.1
37	150	949.4
38	120	856.5
39	126	866.7
40	145	935.5

Tabel 4.6: Verbaal IQ en Hersengrootte voor alle studenten in de steekproef.



Figuur 4.2: Spreidingsdiagram van Verbaal IQ en Hersengrootte. Linksboven: toevoegen van persoon 1. Rechtsboven: toevoegen van persoon 2. Linksonder: toevoegen van persoon 40. Rechtsonder: finaal spreidingsdiagram.

lengte van willekeurige personen (variabele 1) alsook de buitentemperatuur (variabele 2). Beide variabelen hebben geen samenhang: op koude dagen komen zowel grotere als kleinere mensen naar buiten, net als op warme dagen. Figuur 4.3 (rechts) toont het spreidingsdiagram: hier zien we geen patroon, de punten (ook wel *puntenwolk* genoemd) zijn willekeurig verspreid.

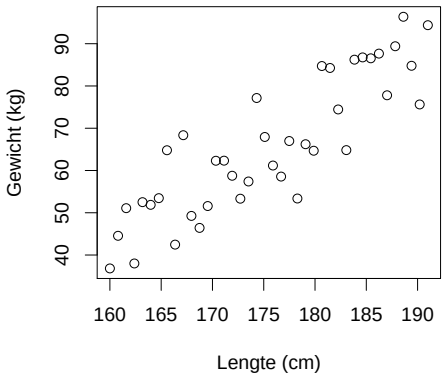


Figuur 4.3: Spreidingsdiagram voor variabelen met een positieve, negatieve en geen samenhang.

Deze drie voorbeelden zijn zeer extreem: er is ofwel een perfecte positieve of perfecte negatieve samenhang ofwel totaal geen samenhang. In de praktijk zal de samenhang van twee variabelen niet zo extreem zijn. Figuur 4.4 illustreert dit: we hebben voor verschillende personen de lengte gemeten in centimeter (variabele 1) en het gewicht bepaald in kilogram (variabele 2). We zien een positieve samenhang: grotere mensen zijn doorgaans zwaarder. De samenhang is echter niet perfect (niet alle observaties liggen op een rechte): sommige kleinere mensen kunnen toch zwaarder zijn dan bepaalde grotere mensen. Bij een realistische positieve samenhang verwachten we dus een puntenwolk van linksonder tot rechtsboven. De punten hoeven echter niet op een rechte te liggen, er kan *spreiding* zijn. Analoog verwachten we bij een negatieve samenhang een puntenwolk van linksboven naar rechtsonder.

Nu we meer inzicht hebben in hoe de spreidingsdiagrammen er kunnen uitzien voor verschillende types van samenhang, kunnen we terugkijken naar het spreidingsdiagram in Figuur 4.2 (rechtsonder). Er is een positieve trend waarneembaar: de puntenwolk gaat van linksonder naar rechtsboven. De trend is echter niet heel uitgesproken: de puntenwolk vertoont redelijk wat spreiding (de punten liggen niet op een rechte). Dit suggereert dat er eerder een 'zwakke' positieve samenhang is tussen Hersengrootte en Verbaal IQ.

Het is duidelijk dat het *interpreteren* van een spreidingsdiagram subjectief is: verschillende personen kunnen andere conclusies trekken op basis van dezelfde figuur: misschien zullen sommigen de samenhang in Figuur 4.2 interpreteren als een 'sterke' samenhang.



Figuur 4.4: Spreidingsdiagram voor het gewicht en de lengte.

Daarom is het handig om de samenhang te *kwantificeren* via *maten van samenhang*.

Illustratie in R

Een spreidingsdiagram kan bekomen worden via `plot()`, met de variabelen als argumenten, gescheiden door een komma:

```
> plot(DataIQ$Hersengrootte, DataIQ$VIQ)
```

4.4 Maten van samenhang

Er bestaan verschillende maten van samenhang. Wij beperken ons tot 3 maten: *de covariantie, de correlatiecoëfficiënt* en *Kendall's* τ. De covariantie en de correlatiecoëfficiënt zijn maten van *lineaire* samenhang, terwijl Kendall's τ een maat is voor *monotone* samenhang - we komen hier uitgebreid op terug in paragraaf 4.4.4 op pagina 125. Om deze maten te kunnen beschrijven hebben we wat extra **notatie** nodig.

Zoals in Hoofdstukken 2 en 3 duiden we een variabele (hier Hersengrootte) aan met X en de waarden met x_1, x_2, ..., x_n met n de steekproefgrootte. We hebben nu ook een tweede variabele (hier Verbaal IQ). Deze tweede variabele zullen we aanduiden met Y en kan de waarden y_1, y_2, ..., y_n aannemen. Het gemiddelde van de waarden van

variabele X wordt voorgesteld door $\bar{x}$ en het gemiddelde van de waarden van variabele Y wordt voorgesteld door $\bar{y}$. De standaarddeviatie van X wordt weergegeven door s_X (zie pagina 90) en van Y door s_Y.

4.4.1 De covariantie

We starten met de formule van de *covariantie* en omschrijven dan hoe deze de samenhang kwantificeert.

! De **covariantie** wordt gegeven door:

$$cov_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{4.1}$$

Als we het sommatieteken uitwerken, kunnen we de covariantie schrijven als

$$cov_{XY} = \frac{1}{n-1}\Big((x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})\Big).$$

Voor de variabele X berekenen we dus het gemiddelde $\bar{x}$ en vervolgens de verschillen $(x_1 - \bar{x})$, $(x_2 - \bar{x})$, $\ldots$, $(x_n - \bar{x})$. Voor de variabele Y is het gelijkaardig: we berekenen het gemiddelde $\bar{y}$ en vervolgens de verschillen $(y_1 - \bar{y})$, $(y_2 - \bar{y})$, $\ldots$, $(y_n - \bar{y})$. In een volgende stap vermenigvuldigen we die verschillen met elkaar $(x_1 - \bar{x})(y_1 - \bar{y})$, $(x_2 - \bar{x})(y_2 - \bar{y})$, $\ldots$, $(x_n - \bar{x})(y_n - \bar{y})$. Tenslotte tellen we deze getallen op en delen we door $n - 1$. De reden waarom we delen door $n - 1$ en niet door n is dezelfde als bij de variantie s_X^2 (zie pagina 90) en we gaan hier niet dieper op in.

! **Meetniveau**. De covariantie maakt gebruik van verschillen en gemiddelden en is bijgevolg enkel maar zinnig als *beide* variabelen van tenminste intervalniveau zijn.

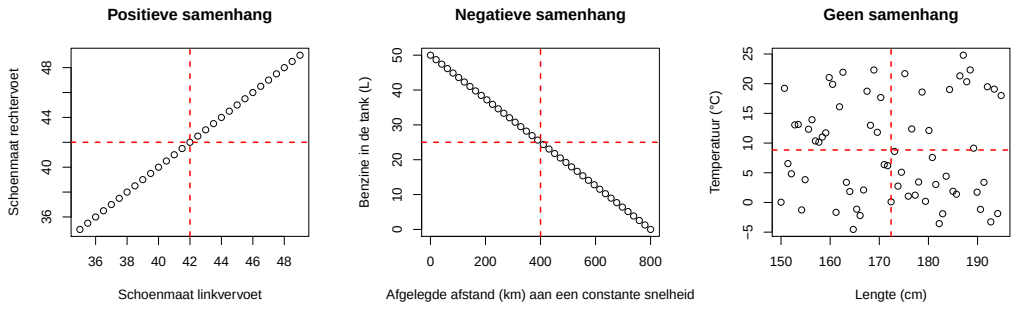
De covarianties horende bij de drie spreidingsdiagrammen in Figuur 4.3 zijn gelijk aan[d] $cov_{XY} = 18.1$ voor de figuur links (een perfecte positieve samenhang), $cov_{XY} = -3594$ voor de figuur in het midden (een perfecte negatieve samenhang) en $cov_{XY} = -0.9$ voor de figuur rechts (geen samenhang).

Er geldt dat:

[d] Omdat de data horende bij deze drie spreidingsdiagrammen niet op Minerva staan, kan je deze covarianties niet zelf narekenen. Verder in de cursus zullen we alle berekeningen illustreren op basis van de Hersengrootte en Verbaal IQ. Dit kan je dan wel zelf narekenen.

- $cov_{XY} > 0$ bij een positieve (lineaire) samenhang.
- $cov_{XY} < 0$ bij een negatieve (lineaire) samenhang.
- $cov_{XY} \approx 0$ indien er geen samenhang is[e].

Om te begrijpen hoe dit komt, moeten we het spreidingsdiagram opsplitsen in vier stukken op basis van de gemiddelden (deze vier stukken worden ook 'kwadranten' genoemd). Dit wordt weergegeven in Figuur 4.5. We kijken eerst naar de figuur links. We berekenen de gemiddelde linkerschoenmaat, deze is hier 42. Vervolgens tekenen we een verticale lijn ter hoogte van 42 op de horizontale as (dus de verticale rode stippellijn). Analoog berekenen we het gemiddelde van de rechterschoenmaten. Dit is hier ook 42 en we tekenen een horizontale lijn ter hoogte van 42 op de verticale as (de horizontale rode stippellijn). We zien dat alle observaties (de punten op de figuur) in de vakken linksonder en rechtsboven liggen. Voor het vak linksonder geldt dat dit punten zijn van personen die kleinere linkervoeten hebben dan gemiddeld ($x_i < \bar{x}$) en kleinere rechtervoeten hebben dan gemiddeld ($y_i < \bar{y}$). Bijgevolg geldt voor deze mensen dat de verschillen $(x_i - \bar{x})$ en $(y_i - \bar{y})$ beiden negatief zijn. Als we deze verschillen vermenigvuldigen worden ze positief (dus $(x_i - \bar{x})(y_i - \bar{y}) > 0$). Analoog zijn voor de personen in het vak rechtsboven de schoenmaten aan beide voeten groter dan gemiddeld: dus $x_i > \bar{x}$ en $y_i > \bar{y}$. Bijgevolg zijn de verschillen $(x_i - \bar{x})$ en $(y_i - \bar{y})$ beiden positief, zodat de vermenigvuldiging ook positief zijn (dus $(x_i - \bar{x})(y_i - \bar{y}) > 0$). De covariantie in formule (4.1) is dus een som van positieve getallen. Bijgevolg is de covariantie ook positief.



Figuur 4.5: Spreidingsdiagram voor variabelen met een positieve (links), negatieve (midden) en geen (rechts) samenhang opgedeeld op basis van de gemiddelden.

Een gelijkaardige redenering gaat op voor de middelste figuur (Figuur 4.5): we verdelen het spreidingsdiagram in vier kwadranten volgens de gemiddelden. Nu liggen alle

[e] $\approx$ staat voor 'ongeveer'.

punten linksboven en rechtsonder. De punten linksboven komen overeen met afstanden kleiner dan de gemiddelde afstand (dus $x_i < \bar{x}$) wanneer er nog meer benzine is dan gemiddeld (dus $y_i > \bar{y}$). Bijgevolg is het verschil $(x_i - \bar{x})$ negatief terwijl het verschil $(y_i - \bar{y})$ positief is, zodat de vermenigvuldiging negatief is (dus $(x_i - \bar{x})(y_i - \bar{y}) < 0$). Voor de punten rechtsonder is het omgekeerd: $x_i > \bar{x}$ en $y_i < \bar{y}$ zodat de vermenigvuldiging van de verschillen ook negatief is (dus $(x_i - \bar{x})(y_i - \bar{y}) < 0$). De covariantie in formule (4.1) is dus een som van negatieve getallen en bijgevolg is de covariantie ook negatief.

Bij de figuur rechts (Figuur 4.5) zien we dat er punten liggen in elk van de vier kwadranten. Dit wil zeggen dat $(x_i - \bar{x})(y_i - \bar{y})$ soms negatief en soms positief kan zijn. De covariantie in formule (4.1) is dus een som van positieve en negatieve getallen die elkaar opheffen zodat de covariantie rond nul zal liggen.

We berekenen nu de covariantie voor Hersengrootte en Verbaal IQ. Tabel 4.7 geeft de verschillende stappen weer.

Opgelet: de waarden in de tabel zijn afgerond tot twee cijfers na de komma, terwijl voor alle berekeningen de originele waarden (zonder afronding) zijn gebruikt. Indien je de covariantie wil narekenen op basis van de gegevens in de tabel, kan de uitkomst wat verschillen. De originele waarden (zonder afronding) zijn beschikbaar via Minerva.

De covariantie is $cov_{XY} = 575.97$, positief wat dus wijst op een positieve samenhang. Dit komt overeen met wat we zien in het spreidingsdiagram in Figuur 4.2.

Nu rest ons de vraag of dit getal groot genoeg is om te besluiten of er een *sterke* samenhang is. De covariantie zal ons hier echter geen antwoord op kunnen geven: de grootte van de covariantie hangt niet enkel af van de sterkte van de samenhang, maar ook van de meeteenheid. Indien we de hersengrootte uitdrukken in pixels in plaats van duizend pixels (we vermenigvuldigen Hersengrootte met duizend), dan is de covariantie ook duizend maal groter: $cov_{XY} = 575971.7$. Het zou daarom interessant zijn om een maat te hebben voor de samenhang die niet afhankelijk is van de meeteenheid. De *correlatiecoëfficiënt* is hiervan een voorbeeld. Alvorens we deze maat bespreken, illustreren we in R hoe je de covariantie kan berekenen.

Illustratie in R

Via `cov()` kunnen we de covariantie tussen twee variabelen berekenen:

```
> cov(DataIQ$Hersengrootte, DataIQ$VIQ)
```

i	Hersengrootte (x_i)	Verbaal IQ (y_i)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	833.9	129	-74.86	16.65	-1246.38
2	878.9	100	-29.86	-12.35	368.74
3	905.9	107	-2.86	-5.35	15.29
4	880.0	96	-28.76	-16.35	470.19
5	854.3	90	-54.46	-22.35	1217.13
6	949.6	145	40.84	32.65	1333.51
7	997.9	96	89.14	-16.35	-1457.48
8	894.0	86	-14.76	-26.35	388.86
9	892.4	83	-16.36	-29.35	480.09
10	1001.1	150	92.34	37.65	3476.70
11	991.3	136	82.54	23.65	1952.13
12	928.8	90	20.04	-22.35	-447.95
13	889.1	77	-19.66	-35.35	694.89
14	935.9	91	27.14	-21.35	-579.49
15	834.3	90	-74.46	-22.35	1664.13
16	924.1	129	15.34	16.65	255.45
17	945.1	96	36.34	-16.35	-594.20
18	790.6	129	-118.16	16.65	-1967.32
19	904.9	93	-3.86	-19.35	74.64
20	852.2	132	-56.56	19.65	-1111.35
21	808.0	112	-100.76	-0.35	35.27
22	951.5	132	42.74	19.65	839.89
23	1079.5	150	170.74	37.65	6428.46
24	1038.4	123	129.64	10.65	1380.69
25	1062.5	96	153.74	-16.35	-2513.69
26	793.5	83	-115.26	-29.35	3382.81
27	816.9	132	-91.86	19.65	-1805.00
28	865.4	71	-43.36	-41.35	1792.83
29	798.6	90	-110.16	-22.35	2462.02
30	965.4	129	56.64	16.65	943.10
31	831.8	86	-76.96	-26.35	2027.83
32	930.0	90	21.24	-22.35	-474.77
33	955.0	145	46.24	32.65	1509.82
34	857.8	126	-50.96	13.65	-695.57
35	955.5	114	46.74	1.65	77.13
36	948.1	129	39.34	16.65	655.05
37	949.4	150	40.64	37.65	1530.19
38	856.5	120	-52.26	7.65	-399.77
39	866.7	126	-42.06	13.65	-574.08
40	935.5	145	26.74	32.65	873.14
	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$				$\frac{22462.9}{39} = 575.97$

Tabel 4.7: De verschillende stappen om de covariantie te berekenen voor Hersengrootte en Verbaal IQ, met $\bar{x} = 908.7575$ *en* $\bar{y} = 112.35$.

```
[1] 575.9717
```

4.4.2 De correlatiecoëfficiënt

De (Pearson) correlatiecoëfficiënt bekomen we door de covariantie te delen door de standaarddeviaties.

! De **correlatiecoëfficiënt** wordt gegeven door:

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

Het delen door de standaarddeviaties zorgt ervoor dat de correlatiecoëfficiënt tussen -1 en 1 ligt:

$$-1 \leq r_{XY} \leq 1.$$

We zullen deze eigenschap niet bewijzen.

Omdat de standaarddeviaties altijd positief zijn, heeft de correlatiecoëfficiënt hetzelfde teken als de covariantie. De correlatiecoëfficiënt voor het voorbeeld met de schoenmaten (Figuur 4.3 op pagina 115) is $r_{XY} = 1$ (perfecte positieve samenhang), voor de middelste figuur is dit $r_{XY} = -1$ (perfecte negatieve samenhang) en voor de figuur rechts geldt $r_{XY} \approx 0$ (geen samenhang).

Samengevat:

- Bij een perfecte positieve (lineaire) samenhang: $r_{XY} = 1$.
- Bij een perfecte negatieve (lineaire) samenhang: $r_{XY} = -1$.
- Indien er geen samenhang is: $r_{XY} \approx 0$.

Voor het voorbeeld rond gewicht en lengte (Figuur 4.4 op pagina 116) is $r_{XY} = 0.87$. Indien er een positieve samenhang is, maar ze is niet perfect (dus niet alle punten liggen op een rechte), zal de correlatie waarden aannemen kleiner dan 1, maar groter dan 0. Voor een negatieve samenhang die niet perfect is, zal ze negatieve waarden aannemen groter dan -1, maar kleiner dan 0.

We keren nu terug naar het voorbeeld rond Hersengrootte en Verbaal IQ. De standaarddeviaties zijn $s_X = 72.3$ voor Hersengrootte, en $s_Y = 23.6$ voor Verbaal IQ. De

covariantie is $cov_{XY} = 575.97$ (zie Tabel 4.7) zodat de correlatiecoëfficiënt gelijk is aan

$$r_{XY} = \frac{575.97}{72.3 \times 23.6} = 0.34.$$

De correlatiecoëfficiënt is positief, wat opnieuw wijst op een positieve samenhang, maar omdat ze relatief klein is, zeggen we dat de samenhang eerder 'zwak' is.

Illustratie in R

De correlatiecoëfficiënt kunnen we berekenen via `cor()`:

```
> cor(DataIQ$Hersengrootte, DataIQ$VIQ)
```

```
[1] 0.3374119
```

4.4.3 Kendall's τ

Naast de correlatiecoëfficiënt bestaan er nog verschillende andere maten van samenhang, zoals bijvoorbeeld *Kendall's* τ (*tau*). Ze wordt berekend door *concordante* en *discordante* paren te tellen.

! Een paar (x_i, y_i) en (x_j, y_j) wordt **concordant** genoemd indien:

$$\frac{y_j - y_i}{x_j - x_i} > 0.$$

! Een paar (x_i, y_i) en (x_j, y_j) wordt **discordant** genoemd indien:

$$\frac{y_j - y_i}{x_j - x_i} < 0.$$

Als voor een paar $x_i = x_j$ of $y_i = y_j$ dan is het paar niet concordant en niet discordant.

! **Kendall's** τ wordt gegeven door:

$$\tau = \frac{2(\text{aantal concordante paren} - \text{aantal discordante paren})}{n(n-1)}.$$

Merk op dat $\frac{y_j - y_i}{x_j - x_i} > 0$ wanneer ($x_i < x_j$ én $y_i < y_j$) of wanneer ($x_i > x_j$ én $y_i > y_j$). Terwijl $\frac{y_j - y_i}{x_j - x_i} < 0$ wanneer ($x_i < x_j$ én $y_i > y_j$) of wanneer ($x_i > x_j$ én $y_i < y_j$). Dus Kendall's τ maakt enkel gebruik van de volgorde van de waarden.

Analoog aan de correlatiecoëfficiënt is ook Kendall's τ begrensd door:

$$-1 \leq \tau \leq 1.$$

Verschillend van de correlatiecoëfficiënt, kan Kendall's τ ook gebruikt worden voor ordinale data.

! **Meetniveau**. Bij de berekening van Kendall's τ gebruikt men enkel de volgorde van de variabelen. Ze is bijgevolg zinnig voor ordinale, interval- en ratiovariabelen.

We illustreren deze maat aan de hand van een voorbeeld. Tabel 4.8 geeft de lengte en het gewicht voor 5 personen. We starten met alle personen paarsgewijs te vergelijken:

- Persoon 1 ($i = 1$) en persoon 2 ($j = 2$). Voor deze personen geldt $x_1 = 160$, $y_1 = 53$, $x_2 = 168$ en $y_2 = 55$. Dit paar is concordant omdat $\frac{y_2 - y_1}{x_2 - x_1} = \frac{55-53}{168-160} = \frac{2}{8} = 0.25 > 0$.
- Persoon 1 ($i = 1$) en persoon 3 ($j = 3$). Voor deze personen geldt $x_1 = 160$, $y_1 = 53$, $x_3 = 176$ en $y_3 = 52$. Dit paar is discordant omdat $\frac{y_3 - y_1}{x_3 - x_1} = \frac{52-53}{176-160} = \frac{-1}{16} < 0$.
- Analoog voor de overige paarsgewijze vergelijkingen: persoon 1 en persoon 4 (concordant), persoon 1 en persoon 5 (concordant), persoon 2 en persoon 3 (discordant), persoon 2 en persoon 4 (concordant), persoon 2 en persoon 5 (concordant), persoon 3 en persoon 4 (concordant), persoon 3 en persoon 5 (concordant), persoon 4 en persoon 5 (concordant).

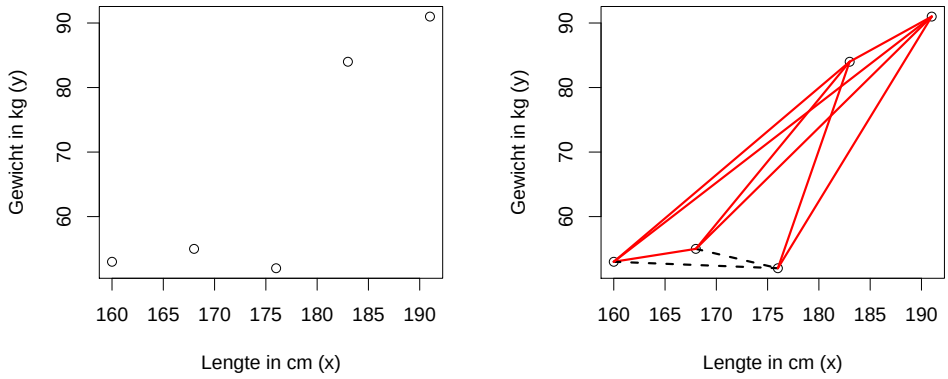
Er zijn dus 8 concordante paren en 2 discordante paren, zodat

$$\tau = \frac{2 \times (8 - 2)}{5 \times (5 - 1)} = \frac{12}{20} = 0.6.$$

De concordante en discordante paren kunnen we ook visueel voorstellen door alle punten in het spreidingsdiagram paarsgewijs te verbinden via rechten, zoals weergegeven in Figuur 4.6. De rechten met een positieve richtingscoëfficiënt (een rechte van linksonder naar rechtsboven) zijn de concordante paren (rode volle lijn op de figuur) en de rechten

Persoon (i)	Lengte in cm (x_i)	Gewicht in kg (y_i)
1	160	53
2	168	55
3	176	52
4	183	84
5	191	91

Tabel 4.8: Dataset om de berekening van Kendall's τ te illustreren.



Figuur 4.6: Links: spreidingsdiagram voor de data uit Tabel 4.8. Rechts: spreidingsdiagram met aanduiding van de concordante paren (rode volle lijn) en discordante paren (zwarte stippellijn).

met een negatieve richtingscoëfficiënt (een rechte van linksboven naar rechtsonder) zijn de discordante paren (zwarte stippellijn op de figuur).

Er bestaan verschillende formules om Kendall's τ te berekenen wanneer een waarde meerdere malen voorkomt ($x_i = x_j$ of $y_i = y_j$), maar deze worden niet besproken in de cursus.

Voor de spreidingsdiagrammen in Figuur 4.3 op pagina 115 geldt dat $\tau = 1$ voor de perfect positieve samenhang (figuur links), $\tau = -1$ voor de perfecte negatieve samenhang (figuur midden) en $\tau \approx 0$ wanneer er geen samenhang is (figuur rechts).

Voor Hersengrootte en Verbaal IQ geldt dat $\tau = 0.28$, wat opnieuw wijst op een (zwakke) positieve samenhang.

Illustratie in R

Door de optie `kendall` te gebruiken, kunnen we via `cor()` Kendall's τ berekenen:

```
> cor(DataIQ$Hersengrootte, DataIQ$VIQ, method = "kendall")
```

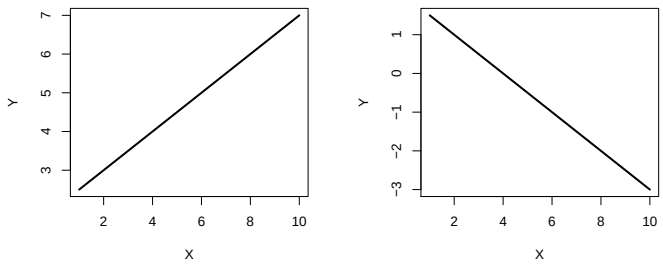
```
[1] 0.2839256
```

Deze berekening maakt gebruik van speciale rekenregels om om te gaan met waarden die meerdere malen voorkomen. Zoals eerder aangegeven gaan we hier niet dieper op in.

4.4.4 Lineaire en niet-lineaire verbanden

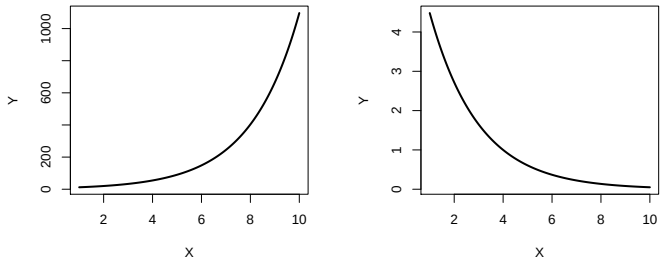
We hebben in de voorgaande paragrafen twee verschillende maten gezien voor samenhang die beiden begrensd zijn tussen -1 en 1, namelijk de correlatiecoëfficiënt en Kendall's τ. Deze laatste maat kan gebruikt worden voor variabelen van tenminste ordinaal meetniveau terwijl de correlatiecoëfficiënt enkel kan gebruikt worden voor variabelen van ten minste interval meetniveau. Er is echter nog een ander belangrijk verschil tussen beiden maten: de correlatiecoëfficiënt (en de covariantie) is een maat voor de *lineaire* samenhang tussen twee variabelen terwijl Kendall's τ een maat is voor een *monotone* samenhang. We bespreken eerst kort wat lineaire en monotone functies zijn.

Een *lineaire functie* is een functie die kan voorgesteld worden door een rechte lijn, zoals geïllustreerd in Figuur 4.7.



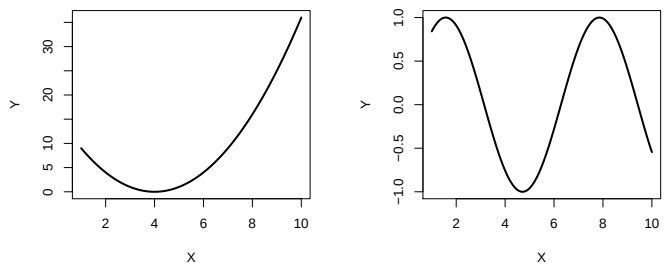
Figuur 4.7: Twee voorbeelden van lineaire functies.

Een monotone functie is een functie die de orde bewaart. Dit wil zeggen dat de functie ofwel moet stijgen ofwel moet dalen, maar niet beiden. De functie moet niet noodzakelijk een rechte lijn zijn. Een lineaire functie is een monotone functie, maar er bestaan ook functies die monotoon zijn zonder lineair te zijn. Figuur 4.8 toont enkele monotone functies die niet lineair zijn. Figuur 4.9 toont tenslotte functies die niet monotoon zijn: de functies kunnen zowel stijgen als dalen.



Figuur 4.8: Twee voorbeelden van monotone functies die niet lineair zijn.

De correlatiecoëfficiënt is dus enkel geschikt als de puntenwolk een lineaire trend vertoont, terwijl Kendall's τ geschikt is voor puntenwolken die een monotone trend vertonen. Figuur 4.10 toont verschillende spreidingsdiagrammen samen met de correlatiecoëfficiënt en Kendall's τ. De figuren bovenaan geven een lineair verband aan en beide maten kunnen gebruikt worden. De middelste rij geeft figuren weer van een monotone niet-lineaire samenhang, hier is Kendall's τ de meeste geschikte maat. De figuren onderaan geven niet-monotone verbanden weer: zowel de correlatiecoëfficiënt



Figuur 4.9: Twee voorbeelden van functies die niet monotoon zijn.

als Kendall's τ zijn hier geen geschikte maten voor de samenhang: op basis van de maten zouden we besluiten dat er een zwakke samenhang is (omdat de getallen klein zijn), terwijl er visueel een sterke (niet-monotone) samenhang is. Het is dus belangrijk om data eerst te *visualiseren* door middel van een spreidingsdiagram en dan pas te beslissen welke maat van samenhang geschikt is.

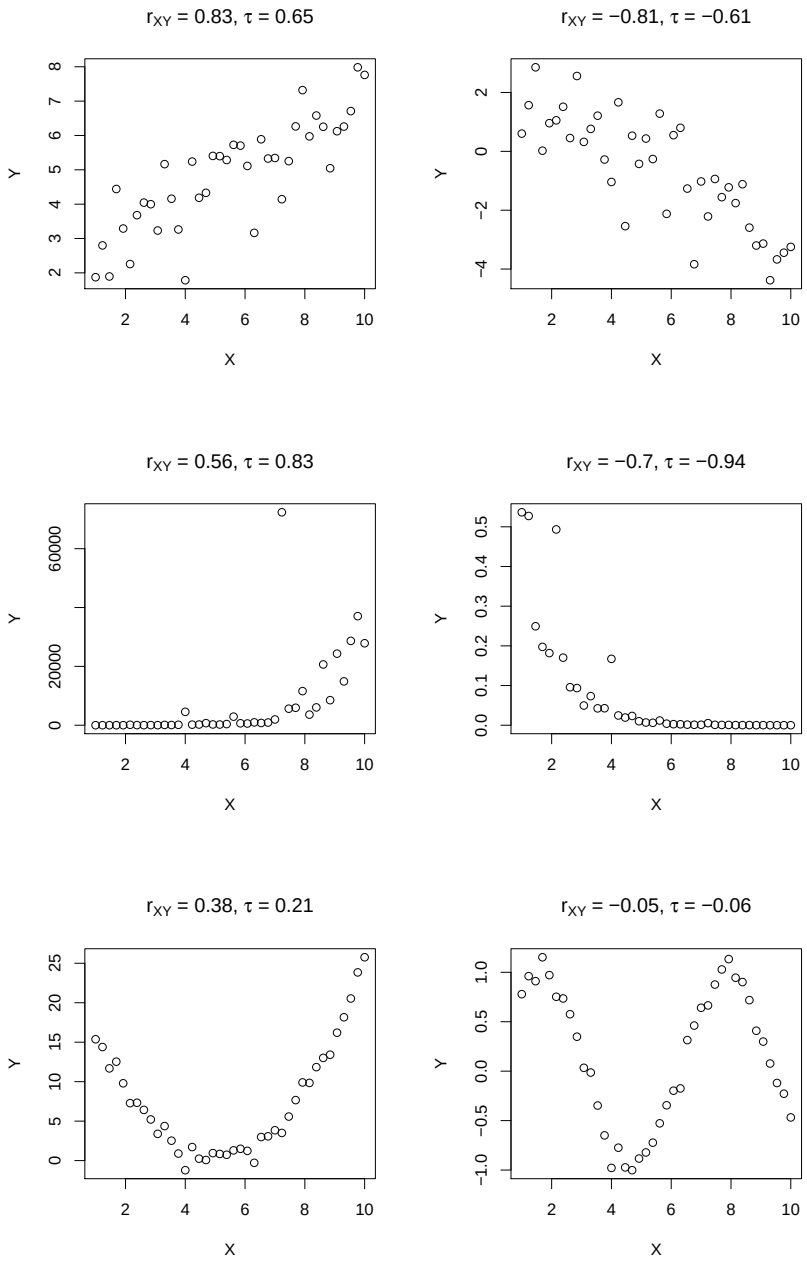
4.4.5 Gevoeligheid aan outliers

De covariantie en de correlatiecoëfficiënt zijn gevoelig aan outliers omdat ze gebruik maken van de waarden van de variabelen. Kendall's τ maakt enkel gebruik van de volgorde van de variabelen en is daardoor niet gevoelig aan outliers. Figuur 4.11 illustreert dit: er is nauwelijks samenhang tussen X en Y op twee outliers na (linksonder en rechtboven op het spreidingsdiagram). Deze outliers zorgen ervoor dat de correlatiecoëfficiënt relatief groot zal zijn ($r_{XY} = 0.69$), wat een vertekend beeld geeft. Kendall's τ ($\tau = 0.06$) wordt niet zo sterk beïnvloed en geeft aan dat er geen samenhang is. Dit komt beter overeen met hetgeen we visueel vaststellen.

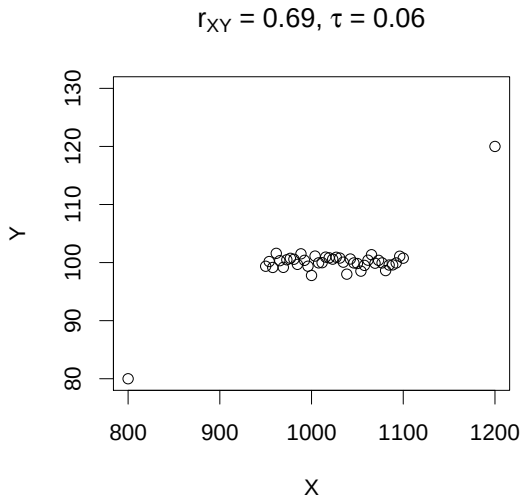
4.5 De regressielijn

De correlatiecoëfficiënt is een veelgebruikte maat voor samenhang en de *regressielijn* zal ons in staat stellen ze te visualiseren op een spreidingsdiagram.

Zoals eerder aangegeven is de correlatiecoëfficiënt geschikt voor een *lineaire* samenhang. Lineair wil zeggen dat het verband tussen Y en X kan beschreven worden door een



Figuur 4.10: Boven: voorbeelden van lineaire samenhang. Midden: voorbeelden van niet-lineaire monotone samenhang. Onder: voorbeelden van niet-monotone samenhang.



Figuur 4.11: Illustratie van een spreidingsdiagram met twee outliers.

rechte[f]:

$$Y = b_0 + b_1 X.$$

We noemen deze rechte *de regressielijn*. Het getal b_1 wordt de *regressiecoëfficiënt* genoemd[g] (dit is de helling van de rechte) en b_0 wordt het *intercept* genoemd (dit is het snijpunt met de verticale as).

4.5.1 Formules indien het lineair verband perfect is

Indien er een perfect lineair verband is, gaat er precies één rechte door alle punten (alle punten liggen dus op 1 rechte). Figuur 4.12 illustreert dit aan de hand van de voorbeelden van Figuur 4.3 op pagina 115.

Bij een perfect lineair verband kunnen we b_0 en b_1 relatief snel berekenen. We kiezen twee *willekeurige* punten (x_i, y_i) en (x_j, y_j) en we passen volgende formule toe voor b_1:

$$b_1 = \frac{y_j - y_i}{x_j - x_i}.$$

[f] In een cursus Wiskunde zal men vaak een rechte voorstellen door $y = ax + b$. In de statistiek gebruiken we echter vaak b_0 en b_1 i.p.v. b en a en gebruiken we hoofdletters voor x en y.

[g] In een cursus Wiskunde zal men b_1 de richtingscoëfficiënt (of rico of helling) noemen.

Eenmaal we b_1 hebben, kunnen we b_0 vinden via:

$$b_0 = y_i - b_1 x_i.$$

We illustreren dit aan de hand van het spreidingsdiagram links in Figuur 4.12. We kiezen twee willekeurige punten, bijvoorbeeld $x_i = 36$, $y_i = 36$ en $x_j = 42$, $y_j = 42$. We vullen vervolgens de formules in:

$$b_1 = \frac{42 - 36}{42 - 36} = \frac{6}{6} = 1,$$

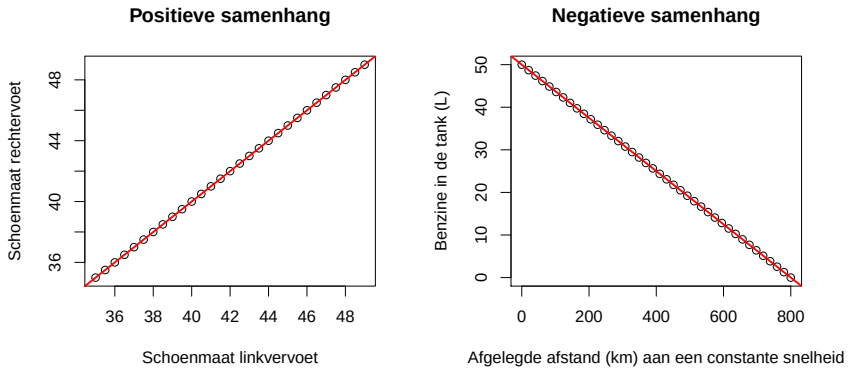
en

$$b_0 = 36 - 1 \times 36 = 0.$$

De vergelijking van de rechte is bijgevolg

$$Y = b_0 + b_1 X = 0 + 1 \times X = X.$$

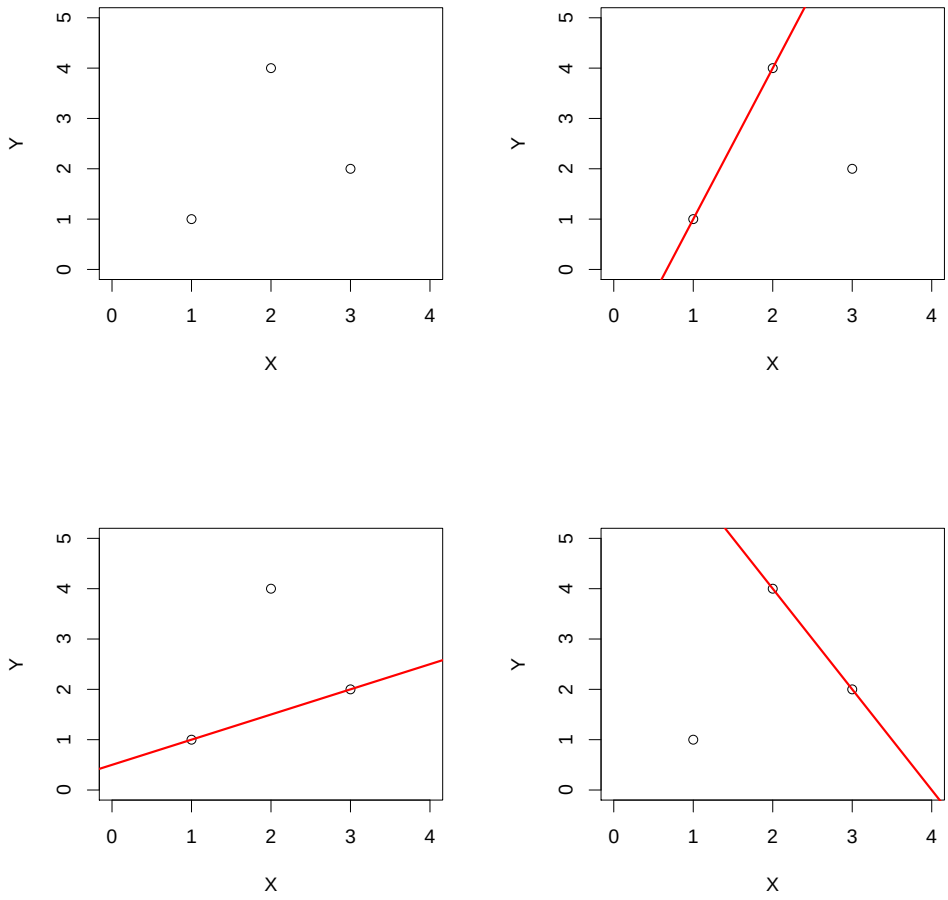
Dit is uiteraard geen verrassing, in deze steekproef hebben alle personen exact dezelfde schoenmaat aan de linkervoet (X) als de rechtervoet (Y), dus $Y = X$.



Figuur 4.12: Perfecte positieve en perfecte negatieve samenhang met aanduiding van de regressielijn (rode volle lijn).

4.5.2 Formules indien het lineair verband niet perfect is

Voor een samenhang die niet perfect is, is het *onmogelijk* een rechte te tekenen die door *alle* punten gaat (niet alle punten liggen op de rechte). Figuur 4.13 illustreert dit aan de hand van een eenvoudig voorbeeld met 3 punten: als we twee punten verbinden met elkaar hebben we een rechte, maar ze gaat niet door het derde punt.



Figuur 4.13: Spreidingsdiagram met 3 punten samen met de 3 mogelijke rechten die we kunnen tekenen door twee punten te verbinden.

Als oplossing zullen we een rechte tekenen die *het best* door de puntenwolk zal gaan via de *kleinste-kwadratenmethode*[h]. De oplossing (die we niet zullen bewijzen) wordt gegeven door:

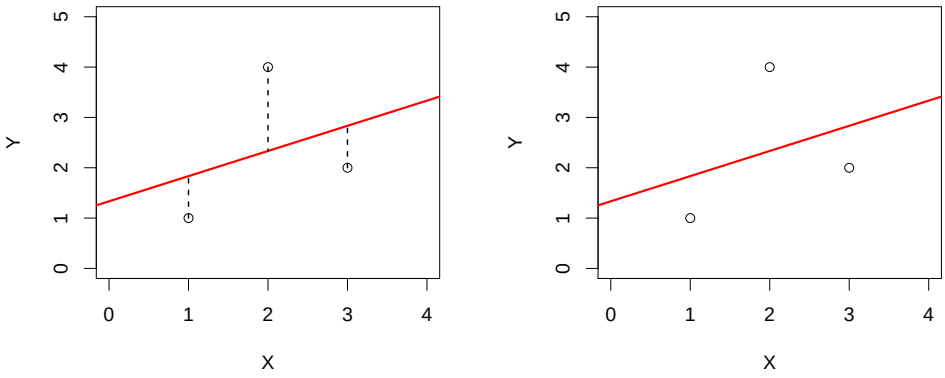
$$b_1 = r_{XY}\frac{s_Y}{s_X},$$

en

$$b_0 = \bar{y} - b_1\bar{x}.$$

De regressiecoëfficiënt wordt dus bekomen door de correlatiecoëfficiënt te vermenigvuldigen met de standaarddeviatie van Y en te delen door de standaarddeviatie van X. Omdat standaarddeviaties nooit negatief kunnen zijn, zal b_1 altijd hetzelfde teken hebben als r_{XY}. In bijna alle gevallen zullen de punten niet op 1 rechte liggen en zullen we gebruik moeten maken van deze formules om b_0 en b_1 te berekenen.

Figuur 4.14 geeft de kleinste-kwadratenmethode grafisch weer: we zullen de rechte (hier de volle rode lijn) zo kiezen zodat de gekwadrateerde lengte van de stippellijnen (dus de afstand tussen een punt en de rode rechte) zo klein mogelijk is. Deze 'best' passende rechte hoeft dus niet door de punten te gaan, maar van alle mogelijke rechten is haar gekwadrateerde afstand tot de punten het kleinst.



Figuur 4.14: Spreidingsdiagram met 3 punten samen met de 'best' passende rechte volgens de kleinste-kwadratenmethode. Links: de lengte van de stippellijn is gelijk aan de afstand tussen de rechte en het punt. Rechts: de finale regressielijn (rode volle lijn).

[h]De vergelijking van deze rechte kunnen we bekomen door volgende uitdrukking te minimaliseren naar b_0 en b_1: $\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$. Dit komt in verschillende vervolgcursussen Statistiek in detail aan bod.

! **Meetniveau**. De vergelijking van de regressielijn maakt gebruik van gemiddelden en de correlatiecoëfficiënt, en is bijgevolg enkel zinnig als beide variabelen van tenminste intervalniveau zijn.

We passen dit toe op het voorbeeld rond Hersengrootte en Verbaal IQ. Op pagina 122 hebben we de correlatiecoëfficiënt berekend $r_{XY} = 0.34$ en de standaarddeviaties zijn $s_X = 72.3$ voor Hersengrootte en $s_Y = 23.6$ voor het Verbaal IQ. Bijgevolg is:

$$b_1 = 0.34 \times \frac{23.6}{72.3} = 0.11.$$

De gemiddelden zijn $\bar{x} = 908.76$ en $\bar{y} = 112.35$ zodat

$$b_0 = 112.35 - 0.11 \times 908.76 = 12.4.$$

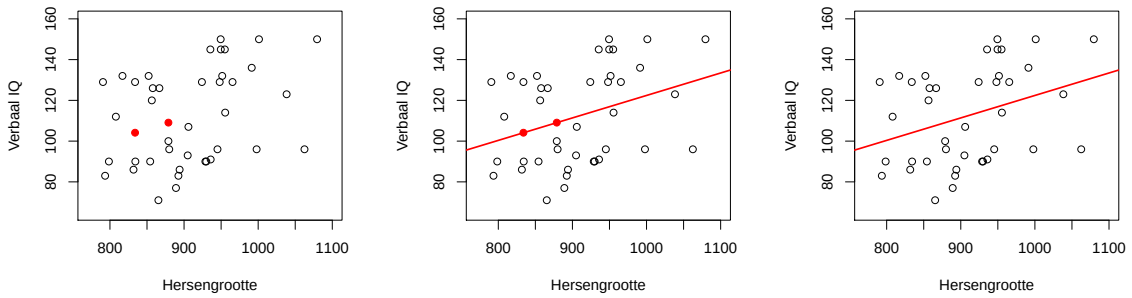
De regressielijn wordt bijgevolg gegeven door

$$Y = 12.4 + 0.11X. \tag{4.2}$$

Deze regressielijn kan je als volgt tekenen op het spreidingsdiagram:

- Neem twee willekeurige waarden voor X, bijvoorbeeld $x_1 = 833.90$ en $x_2 = 878.90$.
- Vul voor elk van deze waarden de formule van de regressielijn in. Voor ons voorbeeld is dit $12.4 + 0.11 \times 833.90 = 104.13$ en $12.4 + 0.11 \times 878.90 = 109.08$.
- Teken deze punten op het spreidingsdiagram: $(833.90, 104.13)$ en $(878.90, 109.08)$, zie Figuur 4.15 links (rode volle punten).
- Als we deze twee punten verbinden met een rechte bekomen we de regressielijn, zie Figuur 4.15 midden (rode volle lijn).

Figuur 4.15 rechts toont het finale spreidingsdiagram samen met de regressielijn. De lijn gaat centraal door de puntenwolk en stelt ons in staat de samenhang beter te beoordelen: er is een stijgende trend (omdat de regressielijn stijgt), maar de punten liggen toch sterk verspreid rond deze rechte. Dit geeft aan dat er een zwakke positieve samenhang is. Dit is dus een bevestiging van onze eerdere conclusies.



Figuur 4.15: Spreidingsdiagram van Hersengrootte en Verbaal IQ met de verschillende stappen om een regressielijn te tekenen.

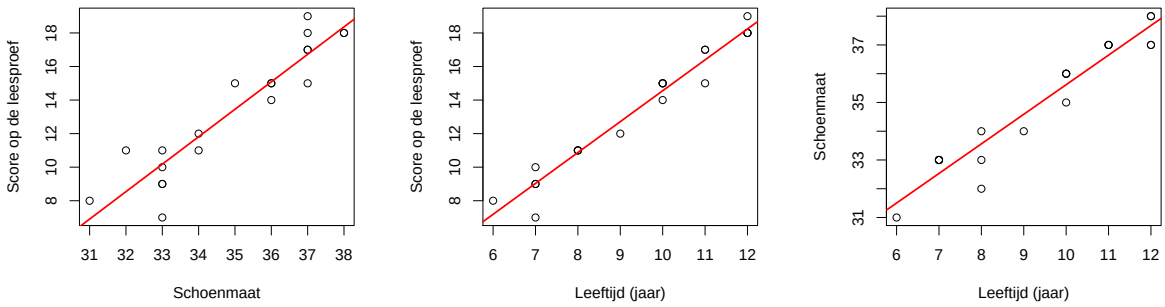
4.6 Samenhang en causaliteit

Indien we besluiten dat er een samenhang is tussen twee variabelen, wil dit niet noodzakelijk zeggen dat er een *causaal* verband is. Zelfs indien er een zeer sterke samenhang is, hoeft dit niet te impliceren dat wijzigingen in de éne variabele veroorzaakt worden door wijzigingen in de andere variabele. Het is bijvoorbeeld mogelijk dat de samenhang veroorzaakt wordt door een derde variabele.

Er is bijvoorbeeld een correlatie tussen de schoenmaat van kinderen en hun score op een leesproef: kinderen met groter voeten behalen vaak betere punten, zie Figuur 4.16 links.

Het is hier echter evident dat de grootte van de voet niet de oorzaak is van de score op de leesproef. Er is namelijk een derde variabele die we over het hoofd zien: de leeftijd van het kind. Kinderen die ouder zijn, scoren beter op de leesproef: zie Figuur 4.16 (midden). Oudere kinderen hebben ook grotere voeten, zie Figuur 4.16 (rechts). Niettegenstaande er een samenhang is tussen schoenmaat en de score, is er dus duidelijk geen oorzakelijk verband.

In dit voorbeeld is het onmiddellijk duidelijk dat er geen causaal verband is tussen de schoenmaat en de leesscore. Bij vele andere voorbeelden is dit echter niet altijd even duidelijk en is het bijgevolg moeilijk om aan te tonen dat de samenhang al dan niet causaal is. We moeten bijgevolg zeer voorzichtig conclusies formuleren: als we besluiten dat er een samenhang is, vermijden we best een formulering die een causaal verband impliceert.



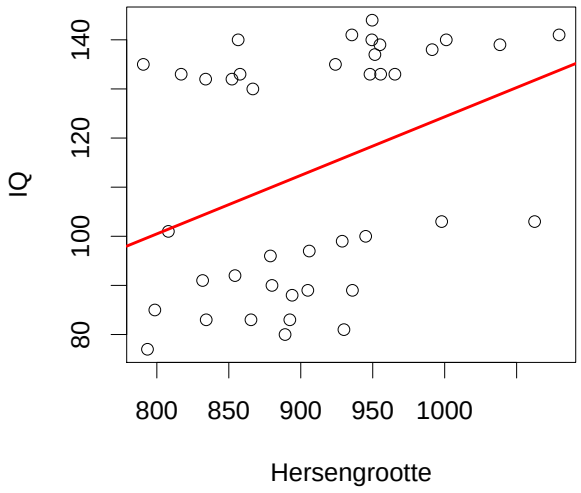
Figuur 4.16: Spreidingsdiagram van de score op een leesproef en de schoenmaat (links), de score op de leesproef en leeftijd (midden) en de schoenmaat en leeftijd (rechts). De rode lijn is de regressielijn.

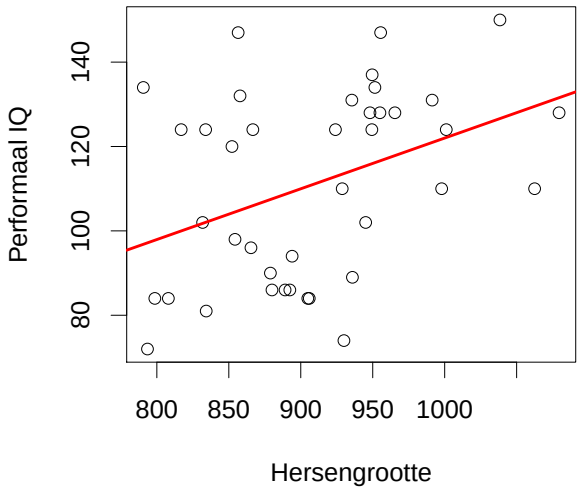
4.7 Een voorbeeld: samenvatten en grafisch voorstellen van onderzoek naar intelligentie en hersengrootte

We sluiten dit hoofdstuk af met de samenhang tussen Hersengrootte en het (totaal) IQ en Performaal IQ te onderzoeken. Figuur 4.17 bovenaan geeft het spreidingsdiagram met de regressielijn voor het totaal IQ. Het spreidingsdiagram toont aan dat er twee puntenwolken zijn, dit komt doordat de onderzoekers 20 studenten met een 'hoog' IQ en 20 met een 'gemiddeld' IQ hebben geselecteerd.

De correlatiecoëfficiënt bedraagt $r_{XY} = 0.36$ en Kendall's $\tau = 0.33$. Analoog als bij het Visueel IQ kunnen we besluiten dat er een zwakke positieve samenhang is tussen Hersengrootte en totaal IQ. We weten nu echter dat samenhang niet noodzakelijk causaliteit impliceert. Uit deze samenhang kan je bijgevolg niet besluiten dat mensen met grotere hersenen ook slimmer zijn, we kunnen enkel besluiten dat er een zwakke positieve samenhang is. Dit is ook de reden waarom we op pagina 105 de onderzoeksvraag hebben geherformuleerd van *Zijn mensen met grotere hersenen slimmer?* naar *Is er een verband tussen hersengrootte en intelligentie?*. Deze laatste formulering laat toe dat het verband niet-causaal kan zijn, terwijl de eerste formulering eerder een causaal verband suggereert.

Op een gelijkaardige wijze kunnen we besluiten formuleren voor de samenhang tussen Hersengrootte en Performaal IQ (Figuur 4.17 onderaan), waarvoor $r_{XY} = 0.39$ en $\tau = 0.28$.





Figuur 4.17: Spreidingsdiagram van Hersengrootte en het (totaal) IQ (boven) en van Hersengrootte en Performaal IQ (onder).

4.8 Samenvatting

In dit hoofdstuk hebben we verschillende methodes besproken om de samenhang tussen twee variabelen te onderzoeken:

- de bivariate frequentieverdeling.
- het spreidingsdiagram.
- maten van samenhang: de covariantie, de correlatiecoëfficiënt en Kendall's τ.
- de regressielijn.

Voor de verschillende maten hebben we het meetniveau besproken samen met de gevoeligheid aan outliers. We hebben aan de hand van een eenvoudig voorbeeld geïllustreerd dat samenhang niet noodzakelijk een causaal verband impliceert.

Deze methodes hebben we toegepast op de gegevens afkomstig uit een studie rond hersengrootte en intelligentie.

Deel II

Kansrekening

Hoofdstuk 5

De populatie en verdelingsfuncties

In Deel I van de syllabus hebben we verschillende technieken besproken om data afkomstig uit een steekproef te visualiseren en samen te vatten. Zoals besproken in paragraaf 1.2 willen we vooral uitspraken doen over de populatie waaruit de steekproef is getrokken. Alvorens we dit kunnen doen, moeten we de populatie eerst formeel wiskundig omschrijven. Veel van de begrippen die we reeds hebben ingevoerd voor de steekproef (vb. gemiddelde, variantie, verdeling) voeren we nu ook in voor de populatie.

De wijze waarop we deze begrippen invoeren, zal iets technischer zijn dan in Deel I. Er zijn hier verschillende redenen voor:

- Een populatie kan *zeer groot* zijn. Denk maar aan het IAT experiment uit Hoofdstuk 2: de populatie bevat ongeveer 258 miljoen elementen. Om een dergelijk grote populatie te beschrijven, is het wiskundig gezien eenvoudiger om ze als *oneindig groot* te beschouwen. Dit vraagt echter wel wat extra wiskundige begrippen.
- Om een uitspraak over een steekproef te veralgemenen naar een populatie zullen we beroep doen op wiskunde, meer bepaald kansrekening.

Het beschrijven van de populatie vraagt wat extra wiskunde. We zullen dit echter beperken tot een minimum en we zullen verschillende definities ook intuïtief verduidelijken.

5.1 Verdelingsfunctie discrete variabelen

Een populatie kan beschreven worden aan de hand van een *verdelingsfunctie*. Een verdelingsfunctie kan gezien worden als de tegenhanger van de frequentieverdeling maar nu gedefinieerd voor een populatie in plaats van een steekproef. De beschrijving van een verdelingsfunctie hangt af van het type variabele: discreet of continu (zie pagina 19 voor meer informatie over deze types). We starten met het discrete type.

Discrete variabelen kunnen maar een *eindig* aantal waarden aannemen. We duiden dit aantal aan met p. Om bijvoorbeeld het visueel geheugen te onderzoeken bij kinderen via de Benton Visual Retention Test (zie paragraaf 1.1.2 op pagina 9), kan men voor tien verschillende opdrachten het aantal correcte antwoorden scoren. De variabele 'Score' kan dus 11 verschillende waarden aannemen (0: alle antwoorden zijn fout, 1: 1 antwoord is correct, ..., 10: alle antwoorden zijn correct), dus $p = 11$ in dit geval.

Afhankelijk van de onderzoeksvraag, kan de populatie groot zijn (bv. alle kinderen van een bepaalde leeftijd in Gent in het jaar 2014, alle kinderen in België in de periode 1995-2015, etc.), en zoals eerder aangegeven zal het eenvoudiger zijn om het aantal elementen in de populatie als oneindig te beschouwen. We hebben dus een populatie van oneindig veel elementen en een variabele die een eindig aantal waarden kan aannemen.

De p verschillende waarden die de variabele X kan aannemen, duiden we aan als x_1, x_2, ..., x_p. Voor de variabele Score is dit: $x_1 = 0, x_2 = 1, \ldots, x_{11} = 10$. Deze notatie is wat verschillend van de notatie uit Deel I en uit de context zal het duidelijk zijn of we verwijzen naar waarden van een variabele uit een steekproef of uit een populatie.

Met $P(X = x_i)$ duiden we de *kans* aan dat de variabele X de waarde x_i aanneemt. De betekenis van deze kans hangt nauw samen met de *frequentieverdeling* in een steekproef. Als we met f_i de absolute frequentie voorstellen van x_i in een steekproef van grootte n, dan kan de kans formeel gedefinieerd worden als:

$$P(X = x_i) = \lim_{n \to \infty} \frac{f_i}{n}. \tag{5.1}$$

Het is de limiet van de relatieve frequentie in de steekproef wanneer de steekproef oneindig groot wordt. Informeel kunnen we de kans $P(X = x_i)$ interpreteren als *relatieve frequentie van x_i in de populatie*. In Hoofdstuk 6 geven we een tweede interpretatie aan $P(X = x_i)$ die duidelijk zal maken waarom we dit een *kans* noemen.

5.1.1 De kansverdeling

! De **kansverdeling** van een discrete variabele X is een tabel met twee kolommen (of rijen) waarbij de eerste kolom (of rij) de waarden x_i weergeeft en de tweede kolom (of rij) de overeenkomstige kansen $P(X = x_i)$.

De kansverdeling kan dus gezien worden als de tegenhanger van de relatieve frequentieverdeling op populatieniveau. Net als de relatieve frequentie ligt de kans in het interval $[0, 1]$.

5.1.2 De cumulatieve verdelingsfunctie

De *cumulatieve verdelingsfunctie* is de tegenhanger van de cumulatieve relatieve frequentie. Soms spreken we kortweg over *de verdelingsfunctie*.

! De **cumulatieve verdelingsfunctie** $F_X(x)$ geeft de kans dat de waarde van een variabele X kleiner dan of gelijk is aan x:

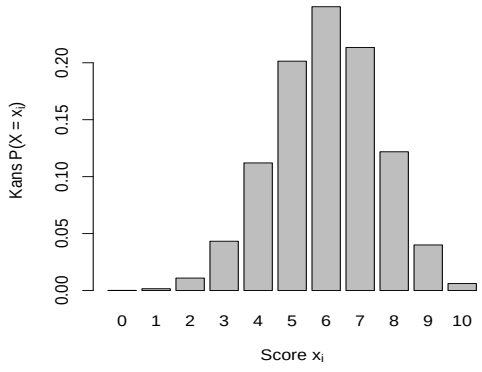
$$F_X(x) = P(X \leq x).$$

We kunnen de kansverdeling en de (cumulatieve) verdelingsfunctie grafisch voorstellen op analoge wijze als de relatieve frequentieverdeling en de cumulatieve frequentieverdeling. We illustreren dit aan de hand van een hypothetisch voorbeeld. Tabel 5.1 geeft de kansverdeling van de scores op de Benton Visual Retention Test. De kans om een score van bv. 5 te halen, is 0.2014, dus ongeveer 20%. Dit wil dus zeggen dat 20% van de kinderen in de populatie een score van 5 hebben. We kunnen deze kansverdeling visualiseren aan de hand van een staafdiagram, zie Figuur 5.1.

De cumulatieve verdelingsfunctie $F_X(x)$ kan je bekomen door de kansen $P(X = x_i)$ uit de kansverdeling waarvoor $x_i \leq x$ op te tellen. Tabel 5.2 en Figuur 5.2 geven deze verdelingsfunctie weer. Ter illustratie: de kans $P(X \leq 3)$ is gelijk aan de som $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$, zodat $P(X \leq 3) = 0.00013 + 0.00158 + 0.01094 + 0.04328 = 0.05593$. Figuur 5.3 geeft dit grafisch weer: de som van de hoogtes van de gekleurde staven geeft de kans $P(X \leq 3)$.

Score x_i	$P(X = x_i)$
0	0.00013
1	0.00158
2	0.01094
3	0.04328
4	0.11205
5	0.20140
6	0.24922
7	0.21341
8	0.12184
9	0.04003
10	0.00612

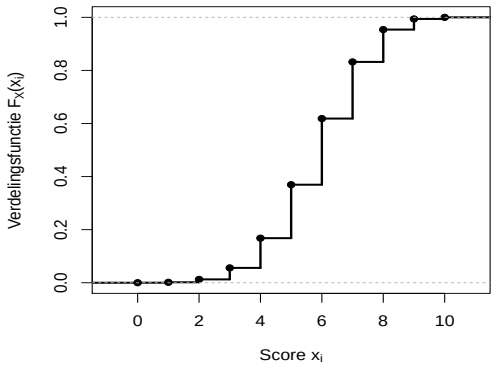
Tabel 5.1: Kansverdeling van de score op de Benton Visual Retention Test.



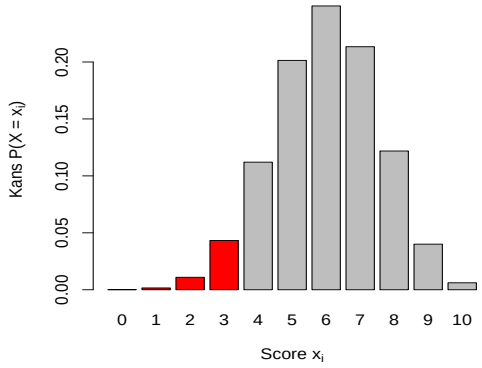
Figuur 5.1: Staafdiagram van de kansverdeling voor de score op de Benton Visual Retention Test.

Score x_i	$F_X(x_i) = P(X \leq x_i)$
0	0.00013
1	0.00171
2	0.01265
3	0.05593
4	0.16798
5	0.36938
6	0.61860
7	0.83201
8	0.95385
9	0.99388
10	1.00000

Tabel 5.2: Cumulatieve verdelingsfunctie van de score op de Benton Visual Retention Test.



Figuur 5.2: Verdelingsfunctie voor de score op de Benton Visual Retention Test.



Figuur 5.3: Staafdiagram van de kansverdeling voor de score op de Benton Visual Retention Test. De som van de hoogtes van de rood gekleurde staven geeft de kans $P(X \leq 3)$.

5.2 Verdelingsfunctie continue variabelen

Een continue variabele kan in theorie oneindig veel verschillende waarden aannemen. Dit impliceert dat de kans $P(X = x) = 0$ voor elke waarde x.

Om dit intuïtief te verduidelijken beschouwen we een analogie waar een computer een willekeurig getal genereert en wij het moeten raden.

In een eerste situatie stellen we de computer zo in dat het enkel gehele getallen tussen 0 en 10 kan genereren. Er zijn dus 11 mogelijke getallen: 0, 1, 2, ..., 10. De kans dat we door te gokken correct het getal kunnen raden is $\frac{1}{11}$.

In een tweede situatie stellen we de computer zo in dat elk *reëel* getal tussen 0 en 10 mogelijk is. De kans dat we het getal kunnen raden is nu gelijk aan 0 omdat er immers oneindig veel mogelijke getallen zijn (vb. $\sqrt{2}$, $\frac{1}{54837}$, 0.125363820272, etc.). Dit wil niet zeggen dat het onmogelijk is om het getal te raden, maar het is wel zéér onwaarschijnlijk (omdat er oneindig veel mogelijkheden zijn). Wiskundig vertaalt zich dit naar een kans van nul.

De eerste situatie is een analogie voor discrete variabelen: er zijn een eindig aantal mogelijke waarden zodat de kansen $P(X = x)$ verschillend van nul kunnen zijn. De tweede situatie komt overeen met continue variabelen: omdat ze een oneindig aantal waarden kunnen aannemen, is de kans $P(X = x)$ gelijk aan nul.

Om kansen te kunnen berekenen bij continue variabelen, zullen we beroep doen op de *dichtheidsfunctie*. Eerst introduceren we de cumulatieve verdelingsfunctie voor continue

variabelen.

5.2.1 De cumulatieve verdelingsfunctie

Niettegenstaande de kans $P(X = x)$ altijd gelijk is aan nul voor continue variabelen, zijn er wel degelijk kansen die verschillend zijn van nul. Eén voorbeeld hiervan is de cumulatieve verdelingsfunctie $P(X \leq x)$.

We illustreren dit opnieuw met de computer die willekeurig reële getallen genereert tussen 0 en 10. We weten dat de kans dat we correct zijn nul is indien we één getal als antwoord geven. Stel nu dat we in plaats van één getal, het volgend antwoord geven: het getal is kleiner dan of gelijk aan 5. De kans dat we nu correct zijn is $\frac{1}{2}$, want de helft van alle mogelijke reële getallen is omvat in ons antwoord.

Analoog als bij discrete variabelen kunnen we dus ook voor continue variabelen de verdelingsfunctie definiëren.

! De **cumulatieve verdelingsfunctie** $F_X(x)$ geeft de kans dat de waarde van een variabele X kleiner dan of gelijk is aan x:

$$F_X(x) = P(X \leq x).$$

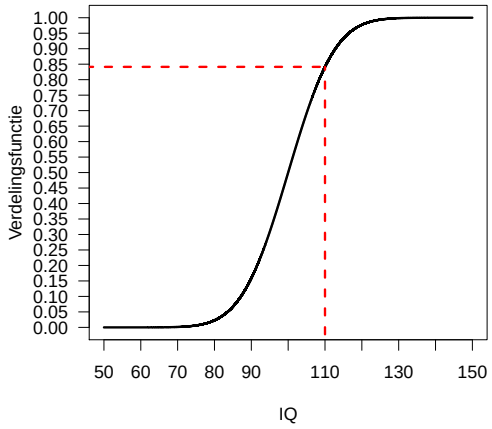
Figuur 5.4 geeft deze verdelingsfunctie voor de variabele IQ in een populatie. De functie is nu continu, terwijl ze voor discrete variabelen trapsgewijs is. Aan de hand van deze grafiek kunnen we de kansen $P(X \leq x)$ aflezen. Ter illustratie: de rode stippellijn geeft aan dat ongeveer 85% van de populatie een IQ kleiner dan of gelijk aan 110 heeft, dus $P(IQ \leq 110) = 0.85$.

Opgelet: Bij continue variabelen maakt het niet uit of we $<$ of $\leq$ gebruiken omdat $P(X = x) = 0$. Het volgt dat[a]:

$$P(X \leq x) = P(X < x \text{ of } X = x) = P(X < x) + P(X = x) = P(X < x) + 0 = P(X < x).$$

Merk op dat bovenstaande eigenschap niet geldig is voor discrete variabelen: omdat $P(X = x)$ verschillend van nul kan zijn, is het mogelijk dat $P(X \leq x) \neq P(X < x)$.

[a] De gelijkheid $P(X < x \text{ of } X = x) = P(X < x) + P(X = x)$ volgt uit rekenregels m.b.t. de kans van de unie van twee gebeurtenissen. In deze syllabus gaan we daar niet dieper op in.



Figuur 5.4: De cumulatieve verdelingsfunctie van de IQ-scores voor een populatie. De rode stippellijn geeft de waarde van de verdelingsfunctie bij een IQ van 110: hier is $P(IQ \leq 110) = 0.85$.

5.2.2 De dichtheidsfunctie

Op basis van de verdelingsfunctie kunnen we de *dichtheidsfunctie* definiëren. Deze zal een zeer belangrijke rol spelen binnen de statistiek. We geven eerst de formele definitie om dan vervolgens op een meer intuïtieve wijze de dichtheidsfunctie te illustreren.

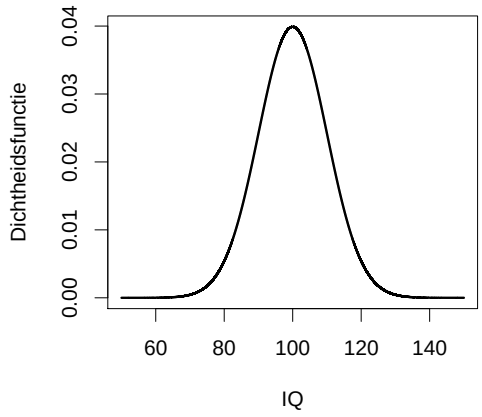
! Voor een variabele X wordt de dichtheidsfunctie $f_X(x)$ (ook wel de kansdichtheid genoemd) gegeven door de afgeleide van de verdelingsfunctie:

$$f_X(x) = \lim_{b \to 0} \frac{F_X(x+b) - F_X(x)}{b}. \tag{5.2}$$

Uitgedrukt in woorden geeft $f_X(x)$ de kans weer dat X valt binnen het interval $[x, x+b]$ *gedeeld* door b, waar b de breedte van het interval voorstelt en naar nul convergeert (dus de breedte van het interval wordt kleiner en kleiner). Omdat we delen door b heeft $f_X(x)$ niet de interpretatie van een kans (als b zeer klein is, kan $f_X(x)$ groter zijn dan 1, dus kan het geen kans zijn).

Opgelet: In deze cursus zullen we zelf geen afgeleiden berekenen, formule (5.2) dient louter om de dichtheidsfunctie formeel te introduceren. Ook zonder kennis van afgeleiden kan je de rest van deze syllabus volgen.

Figuur 5.5 toont de dichtheidsfunctie horende bij de verdelingsfunctie uit Figuur 5.4.



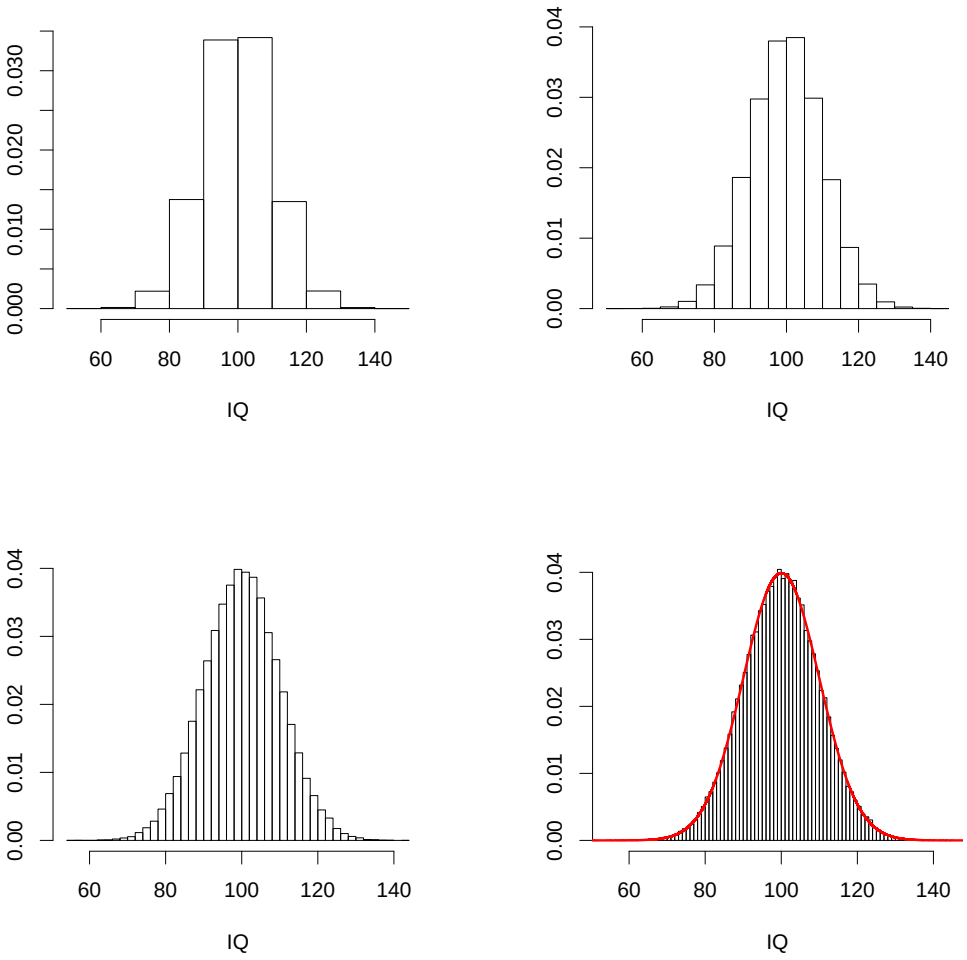
Figuur 5.5: De dichtheidsfunctie van de IQ-scores voor een populatie.

De dichtheidsfunctie kan op een meer intuïtieve wijze bekomen worden via het construeren van een bepaald type histogrammen. Figuur 5.6 illustreert dit. Ze toont een histogram van de IQ-scores voor alle personen in de populatie. In het histogram zijn de *oppervlaktes* van de rechthoeken gelijk aan de relatieve frequenties. In de figuur linksboven verdelen we de IQ-scores in 10 klassen. Voor de figuur rechtsboven zijn 20 klassen gebruikt zodat de klassenbreedtes smaller worden. In de figuur linksonder en rechtsonder zijn de klassenbreedtes nog smaller. We zien dat naarmate het aantal klassen toeneemt (en de breedte van de klassen kleiner wordt), het histogram meer en meer kan benaderd worden door een continue functie (de rode grafiek op het histogram rechtsonder). Deze continue functie is de dichtheidsfunctie en wordt theoretisch bekomen door het histogram op te delen in oneindig veel klassen.

Via de dichtheidsfunctie kunnen we kansen van de vorm $P(x_1 \leq X \leq x_2)$ berekenen. Om deze kansen te bekomen, moeten we de dichtheidsfunctie *integreren*. We illustreren dit voor $x_1 = 90$ en $x_2 = 110$:

$$P(90 \leq X \leq 110) = \int_{90}^{110} f_X(x)dx, \tag{5.3}$$

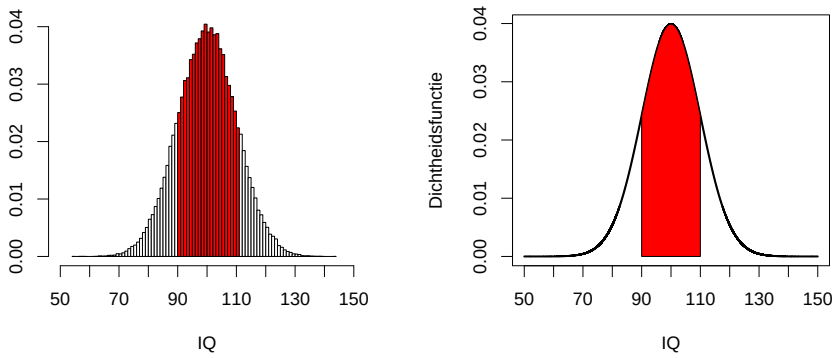
waar X de IQ-score voorstelt. De notatie $\int_{90}^{110} f_X(x)dx$ staat voor 'de integraal van $f_X(x)$ waarbij x gaat van 90 tot 110'. Net als het sommeren, is het *integreren* een wiskundige bewerking. Zoals eerder aangegeven kunnen we de dichtheidsfunctie benaderen door een histogram met zeer veel klassen (zie Figuur 5.6). De integraal $\int_{90}^{110} f_X(x)dx$ kan dan benaderd worden door de oppervlaktes van de rechthoeken tussen de grenzen 90 en 110 op te tellen, zie Figuur 5.7 (figuur links). De exacte integraal $\int_{90}^{110} f_X(x)dx$



Figuur 5.6: Histogrammen voor de IQ-scores van een populatie waarbij de klassen smaller worden. De rode lijn op de figuur rechtsonder geeft de dichtheidsfunctie.

komt overeen met de oppervlakte wanneer we het histogram in oneindig veel klassen verdelen. Figuur 5.7 (figuur rechts) illustreert dit: de oppervlakte van de gekleurde zone stelt de waarde van de integraal $\int_{90}^{110} f_X(x)dx$ voor. Op basis van vergelijking (5.3) is de oppervlakte van dit gekleurde stuk ook gelijk aan de kans $P(90 \leq IQ \leq 110)$.

In deze cursus zullen we integralen niet analytisch oplossen, we gebruiken ze louter om formeel bepaalde kansen in te voeren. In plaats van integralen op te lossen, zullen we ze visueel voorstellen door oppervlaktes.



Figuur 5.7: De dichtheidsfunctie van de IQ-scores. De gekleurde zone komt overeen met de kans $P(90 \leq IQ \leq 110)$. Figuur links: een benadering van deze kans door de oppervlaktes van de gekleurde rechthoeken op te tellen. Figuur rechts: de exacte kans die bekomen wordt door de integraal $\int_{90}^{110} f_X(x)dx$ uit te werken, wat overeenkomt met het berekenen van de oppervlakte van de gekleurde zone.

Opgelet: Het integreren is een wiskundige bewerking en wordt aangeduid met de symbolen $\int$ en dx. Deze notatie kan complex overkomen en het wiskundig berekenen van een integraal is vaak niet eenvoudig. Voor deze cursus volstaat het te weten dat integralen visueel kunnen worden voorgesteld door oppervlaktes. Het *integreren* van een dichtheidsfunctie maakt geen deel uit van deze cursus. Ook als je niet vertrouwd bent met integraalrekening, kan je de rest van deze syllabus volgen.

Algemeen kunnen we stellen dat

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx. \tag{5.4}$$

Zoals eerder aangegeven is deze integraal gelijk aan een oppervlakte:

! De kans dat een variabele X in het interval $[x_1, x_2]$ ligt, is gelijk aan de oppervlakte onder de dichtheidsfunctie $f_X(x)$ tussen de waarden x_1 en x_2. Deze oppervlakte komt overeen met de integraal van $f_X(x)$ over het interval $[x_1, x_2]$: $\int_{x_1}^{x_2} f_X(x)dx$.

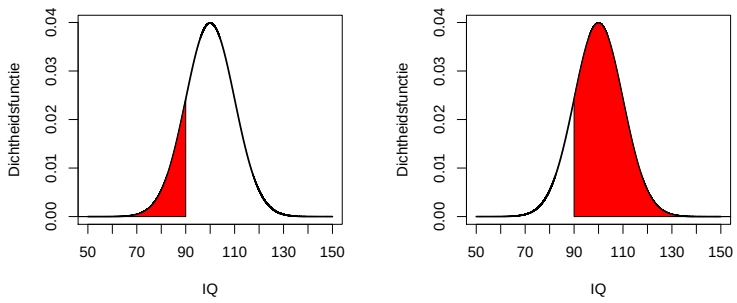
Kansen van de vorm $P(X \leq x)$ en $P(X > x)$ kunnen we op analoge wijze bekomen:

$$P(X \leq x) = \int_{-\infty}^{x} f_X(x)dx, \tag{5.5}$$

en

$$P(X > x) = \int_{x}^{+\infty} f_X(x)dx, \tag{5.6}$$

waarbij ∞ staat voor 'oneindig'. Integralen (5.5) en (5.6) kunnen we ook visueel voorstellen, zie Figuur 5.8.



Figuur 5.8: De dichtheidsfunctie van de IQ-scores. De gekleurde zone komt overeen met de kansen $P(IQ \leq 90)$ (figuur links) en $P(IQ > 90)$ (figuur rechts).

We kunnen de kans $P(x_1 \leq X \leq x_2)$ nu visueel voorstellen en dan rest de vraag: wat is de numerieke waarde van deze kans? Om hier een antwoord op te kunnen geven, moeten we de integraal in formule (5.4) berekenen en zoals eerder aangegeven zullen we deze berekening niet zelf uitvoeren omdat ze vaak complex is. Indien we beschikken over de verdelingsfunctie $F_X(x)$, is het echter eenvoudig om de kans te bekomen via de volgende eigenschap (zonder bewijs):

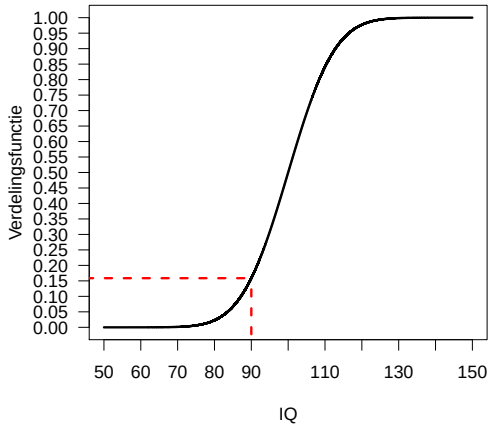
! Er geldt dat:

$$P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F_X(x_2) - F_X(x_1). \tag{5.7}$$

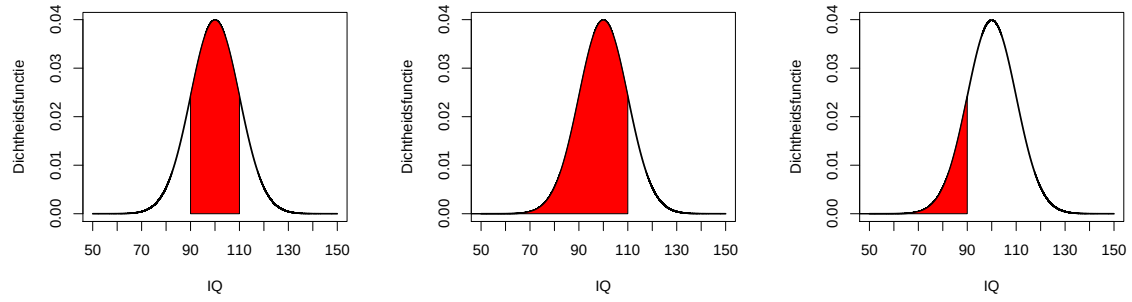
Om de kans $P(90 \leq IQ \leq 110)$ te berekenen, moeten we $P(IQ \leq 90)$ en $P(IQ \leq 110)$ bepalen. Uit Figuur 5.4 kunnen we aflezen dat $P(IQ \leq 110) \approx 0.85$. Uit Figuur 5.9 halen we dat $P(IQ \leq 90) \approx 0.15$. Bijgevolg is

$$P(90 \leq IQ \leq 110) = P(IQ \leq 110) - P(IQ \leq 90) \approx 0.85 - 0.15 = 0.70.$$

De kans om een IQ te hebben tussen 90 en 110 is ongeveer gelijk aan 0.70. Dit kunnen we ook nog formuleren als volgt: ongeveer 70% van de populatie heeft een IQ tussen 90 en 110. Formule (5.7) kunnen we ook visualiseren; zie Figuur 5.10.



Figuur 5.9: De verdelingsfunctie van de IQ-scores voor een populatie. De rode stippellijn geeft de waarde van de verdelingsfunctie bij een IQ van 90: $P(IQ \leq 90) = 0.15$.



Figuur 5.10: Illustratie van formule (5.7) met $x_1 = 90$ en $x_2 = 110$. De gekleurde oppervlakte links geeft de kans $P(90 \leq IQ \leq 110)$ weer. Deze kans is gelijk aan het verschil tussen de kans $P(IQ \leq 110)$ (gekleurde oppervlakte middelste figuur) en de kans $P(IQ \leq 90)$ (gekleurde oppervlakte rechter figuur).

We sluiten deze paragraaf af met drie interessante eigenschappen.

! De dichtheidsfunctie is een positieve functie:

$$f_X(x) \geq 0,$$

voor alle waarden x. Figuur 5.11 geeft een voorbeeld van een functie die onmogelijk een dichtheidsfunctie kan zijn omdat ze negatieve waarden aanneemt. Het is niet verwonderlijk dat een dichtheidsfunctie niet negatief kan zijn: ze is immers gebaseerd op een kans en kansen kunnen niet negatief zijn.

! De volledige oppervlakte onder de dichtheidsfunctie is gelijk aan 1:

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1. \tag{5.8}$$

Figuur 5.11 geeft een voorbeeld van een functie die onmogelijk een dichtheidsfunctie kan zijn omdat de volledige oppervlakte onder de dichtheidsfunctie verschillend is van 1 (de oppervlakte voor deze eenvoudige functie is gelijk aan de basis maal de hoogte, hier $50 \times 0.20 = 10$). De volledige oppervlakte onder de dichtheidsfunctie komt overeen met de kans $P(-\infty < X < +\infty)$ en deze is altijd gelijk aan 1 (de waarden van een continue variabele moeten altijd tussen min en plus oneindig liggen).

! Er geldt dat:

$$P(X > x) = 1 - P(X \leq x). \tag{5.9}$$

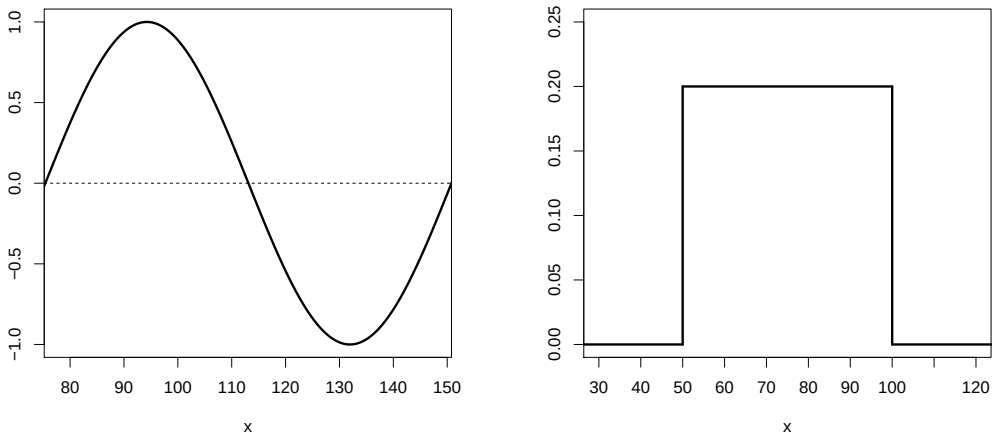
Figuur 5.12 illustreert deze formule.

5.3 Populatieparameters

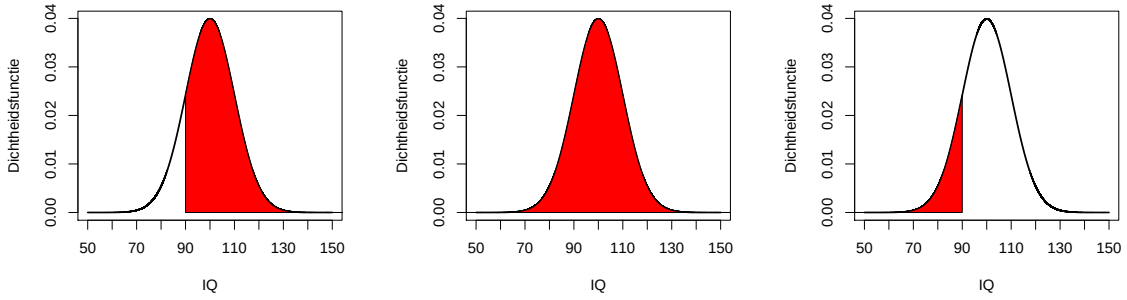
Aan de hand van de kansverdeling $P(X = x_i)$ (voor discrete variabelen) of de kansdichtheid $f_X(x)$ (voor continue variabelen) kunnen we op formele wijze verschillende *populatieparameters* definiëren. We beperken ons tot het gemiddelde en de variantie.

5.3.1 Populatiegemiddelde

Afhankelijk van het type variabele (discreet of continu) hebben we verschillende definities voor het populatiegemiddelde.



Figuur 5.11: Links: functie die geen dichtheidsfunctie kan zijn omdat ze negatieve waarden aanneemt. Rechts: functie die geen dichtheidsfunctie kan zijn omdat de volledige oppervlakte onder de dichtheidsfunctie verschillend is van 1.



Figuur 5.12: Illustratie van formule (5.9) met $x = 90$. De gekleurde oppervlakte van de figuur links geeft de kans $P(IQ > 90)$ weer. De gekleurde oppervlakte van de figuur in het midden is gelijk aan 1, zie formule (5.8). De gekleurde oppervlakte voor de figuur rechts geef de kans $P(IQ \leq 90)$ weer. Het verschil in oppervlakte tussen de figuur in het midden en de figuur rechts is gelijk aan de oppervlakte van de figuur links.

Discrete variabelen

! Het **gemiddelde** van een discrete variabele X in een populatie (symbool $E(X)$), wordt gegeven door:

$$E(X) = \sum_{i=1}^{p} P(X = x_i)x_i. \tag{5.10}$$

Om het onderscheid te kunnen maken tussen het gemiddelde van een steekproef en een populatie, zullen we expliciet refereren naar het *steekproefgemiddelde* en het *populatiegemiddelde*. Soms zullen we ook refereren naar *het gemiddelde* en uit de context moet het dan duidelijk zijn of dit betrekking heeft op een steekproef of op een populatie.

Het populatiegemiddelde wordt ook de *verwachtingswaarde* genoemd, vandaar de letter E wat afkomstig is van *expectatio* (verwachting in het Latijn). In plaats van het symbool $E(X)$, gebruikt men soms ook μ_X of kortweg μ (uitspraak *mu*).

De definitie van het populatiegemiddelde is dus verschillend van die in de steekproef zoals gegeven in formule (3.1) op pagina 68. Dit komt omdat de populatie oneindig groot kan zijn. Er is echter een sterke gelijkenis met de definitie op basis van een frequentieverdeling, zie vergelijking (3.2) op pagina 72. Het verschil tussen formule (3.2) en formule (5.10) is dat de relatieve frequenties $\frac{f_i}{n}$ vervangen zijn door de kansen $P(X = x_i)$. Deze kansen vormen net de tegenhanger van de relatieve frequenties op populatieniveau. Verder is x_i^u vervangen door x_i.

We illustreren dit aan de hand van de kansverdeling in Tabel 5.1 op pagina 142. De verwachtingswaarde (of populatiegemiddelde) van de variabele 'Score' is gelijk aan:

$$E(X) = (0.00013 \times 0) + (0.00158 \times 1) + (0.01094 \times 2) + \ldots (0.00612 \times 10) = 5.99388.$$

Tabel 5.3 toont alle stappen om dit populatiegemiddelde te berekenen. Het populatiegemiddelde van de score op de Benton Visual Retention Test is afgerond gelijk aan 6.

Continue variabelen

Voor continue variabelen is $P(X = x_i) = 0$ en kunnen we bijgevolg definitie (5.10) niet gebruiken. Analoog als bij het berekenen van kansen door middel van het integreren van de dichtheidsfunctie (dus het berekenen van oppervlaktes onder de curve), moeten we ook integreren om het populatiegemiddelde (of de verwachtingswaarde) te definiëren.

i	x_i	$P(X = x_i)$	$P(X = x_i)x_i$
1	0	0.00013	0.00000
2	1	0.00158	0.00158
3	2	0.01094	0.02188
4	3	0.04328	0.12984
5	4	0.11205	0.44820
6	5	0.20140	1.00700
7	6	0.24922	1.49532
8	7	0.21341	1.49387
9	8	0.12184	0.97472
10	9	0.04003	0.36027
11	10	0.00612	0.06120
$E(X) = \sum_{i=1}^{p} P(X = x_i)x_i$			5.99388

Tabel 5.3: De verschillende stappen om de verwachtingswaarde van de score op de Benton Visual Retention Test te berekenen waarbij x_i de i^{de} unieke waarde weergeeft van de variabele X in de populatie.

! Het **gemiddelde** (of verwachtingswaarde) van een continue variabele X in een populatie wordt gegeven door:

$$E(X) = \int_{-\infty}^{+\infty} f_X(x)xdx. \tag{5.11}$$

Niettegenstaande vergelijking (5.11) complexer is dan vergelijking (5.10), zijn er toch gelijkenissen. De som in (5.10) wordt vervangen door een integraal $\int_{-\infty}^{+\infty} dx$ en de kansverdeling door de dichtheidsfunctie $f_X(x)$. Het werken met integralen (i.p.v. sommen) is noodzakelijk omdat een continue variabele oneindig veel waarden kan aannemen. Zoals eerder aangegeven zullen we in deze cursus dergelijke integralen niet zelf uitrekenen.

5.3.2 Populatievariantie

Discrete variabelen

! De **variantie** van een discrete variabele X in een populatie (symbool $V(X)$), wordt gegeven door:

$$V(X) = \sum_{i=1}^{p} P(X = x_i)\Big(x_i - E(X)\Big)^2. \tag{5.12}$$

In plaats van het symbool $V(X)$, gebruikt men soms ook σ_X^2 of kortweg σ^2 (uitspraak *sigma-kwadraat*).

De formule van de populatievariantie vertoont gelijkenissen met de formule van de steekproefvariantie sn_X^2 op basis van de frequentieverdeling (formule (3.6) op pagina 90). We kunnen deze formule herschrijven als:

$$sn_X^2 = \sum_{i=1}^{p} \frac{f_i}{n}(x_i^u - \bar{x})^2. \tag{5.13}$$

Uit formule (5.13) kunnen we formule (5.12) afleiden door:

- de relatieve steekproeffrequenties $\frac{f_i}{n}$ te vervangen door de kansen $P(X = x_i)$.
- het steekproefgemiddelde $\bar{x}$ te vervangen door het populatiegemiddelde $E(X)$.
- x_i^u te vervangen door x_i.

Dit toont aan dat er gelijkenissen zijn tussen de formule voor de steekproefvariantie sn_X^2 en de populatievariantie σ_X^2.

Analoog als bij een steekproef kunnen we de standaarddeviatie bekomen door de vierkantswortel te nemen van de variantie.

! De **standaarddeviatie** van een variabele X in een populatie (symbool σ_X) wordt gegeven door:

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

We passen dit toe op het voorbeeld uit Tabel 5.1:

$$\begin{aligned}
\sigma_X^2 &= \sum_{i=1}^{11} P(X = x_i)(x_i - 5.99388)^2 \\
&= 0.00013 \times (0 - 5.99388)^2 + 0.00158 \times (1 - 5.99388)^2 + \ldots + 0.00612 \times (10 - 5.99388)^2 \\
&= 2.417263.
\end{aligned}$$

Tabel 5.4 toont nogmaals deze stappen. De populatievariantie van de score op de Benton Visual Retention Test is dus gelijk aan $V(X) = 2.42$ en de populatiestandaarddeviatie is $\sigma_X = \sqrt{2.42} = 1.55$.

i	x_i	$P(X = x_i)$	$(x_i - E(X))$	$(x_i - E(X))^2$	$P(X = x_i)(x_i - E(X))^2$
1	0	0.00013	-5.99388	35.92660	0.00467
2	1	0.00158	-4.99388	24.93884	0.03940
3	2	0.01094	-3.99388	15.95108	0.17450
4	3	0.04328	-2.99388	8.96332	0.38793
5	4	0.11205	-1.99388	3.97556	0.44546
6	5	0.20140	-0.99388	0.98780	0.19894
7	6	0.24922	0.00612	0.00004	0.00001
8	7	0.21341	1.00612	1.01228	0.21603
9	8	0.12184	2.00612	4.02452	0.49035
10	9	0.04003	3.00612	9.03676	0.36174
11	10	0.00612	4.00612	16.04900	0.09822
$V(X) = \sum_{i=1}^{p} P(X = x_i)(x_i - E(X))^2$					2.417263

Tabel 5.4: De verschillende stappen om de populatievariantie van de score op de Benton Visual Retention Test te berekenen.

Continue variabelen

Analoog als bij de verwachtingswaarde moeten we bij continue variabelen de som vervangen door een integraal en de kansverdeling door de dichtheidsfunctie.

! De **variantie** van een continue variabele X in een populatie wordt gegeven door:

$$V(X) = \int_{-\infty}^{+\infty} f_X(x)\Big(x - E(X)\Big)^2 dx. \tag{5.14}$$

De standaarddeviatie wordt opnieuw bekomen door de vierkantswortel te nemen.

5.4 Bivariate kansverdelingen

Gelijkaardig als in Hoofdstuk 4 kunnen we ook twee variabelen gezamenlijk bekijken. Op populatieniveau zal dit aanleiding geven tot *bivariate kansverdelingen*. Opnieuw moeten we een onderscheid maken tussen discrete en continue variabelen.

5.4.1 Discrete variabelen

Tabel 5.5 toont de bivariate kansverdeling voor een populatie kinderen die de Benton Visual Retention Test hebben afgelegd met 5 opdrachten. De eerste variabele X is de score op 5 en de tweede variabele Y is de leeftijd van het kind. In de populatie hebben we enkel kinderen van 10 of 11 jaar.

De tabel kunnen we als volgt lezen: de kans dat een kind een score 0 heeft *en* 10 jaar oud is, is 0.00341. Symbolisch schrijven we dit als

$$P(X = 0 \text{ en } Y = 10) = 0.00341.$$

Meer algemeen schrijven we de kans dat X de waarde x_i aanneemt en Y de waarde y_j als:

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

	Leeftijd Y	
Score X	10	11
0	0.00341	0.00021
1	0.02730	0.00404
2	0.08275	0.03291
3	0.12110	0.13337
4	0.09119	0.26342
5	0.02711	0.21319

Tabel 5.5: Bivariate kansverdeling voor Score en Leeftijd.

Op basis van deze bivariate verdeling kunnen we de marginale (univariate) verdelingen afleiden door kansen op te tellen (dit hebben we ook gedaan voor de bivariate frequentieverdeling, zie paragraaf 4.2 op pagina 109). We schrijven het aantal mogelijke waarden dat X kan aannemen als p. Toegepast op Tabel 5.5 is $p = 6$. Het aantal mogelijke waarden dat Y kan aannemen, noteren we als q. Voor Tabel 5.5 is $q = 2$. De univariate verdeling van X wordt bekomen via:

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

We nemen de som van de kansen waar X wordt vastgehouden bij de waarde x_i en Y varieert over alle mogelijke waarden. Dit komt overeen met het optellen van de kansen per rij in Tabel 5.5.

Als voorbeeld nemen we de kans dat een kind een score van $x_i = 4$ behaalt:

$$\begin{aligned} P(X = 4) &= \sum_{j=1}^{2} P(X = 4 \text{ en } Y = y_j) \\ &= P(X = 4 \text{ en } Y = y_1) + P(X = 4 \text{ en } Y = y_2) \\ &= P(X = 4 \text{ en } Y = 10) + P(X = 4 \text{ en } Y = 11) \\ &= 0.09119 + 0.26342 \\ &= 0.35461. \end{aligned}$$

Op een gelijkaardige wijze kunnen we ook de univariate kansverdeling van Y afleiden uit de bivariate kansverdeling via:

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

Dit komt overeen met de kansen per kolom in Tabel 5.5 op te tellen. Bijvoorbeeld, de kans dat een kind in de populatie 10 jaar is, wordt gegeven door:

$$\begin{aligned} P(Y = 10) &= \sum_{i=1}^{6} P(X = x_i \text{ en } Y = 10) \\ &= 0.00341 + 0.02730 + 0.08275 + 0.12110 + 0.09119 + 0.02711 \\ &= 0.35286. \end{aligned}$$

Statistische onafhankelijkheid is een belangrijk begrip binnen bivariate kansverdelingen.

! Twee discrete variabelen X en Y zijn **onafhankelijk** als de gelijkheid

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

geldt voor alle mogelijke combinaties van i en j.

Uit Tabel 5.5 kunnen we afleiden dat Score en Leeftijd niet onafhankelijk zijn. Inderdaad:

$$P(X = 4 \text{ en } Y = 10) = 0.09119,$$

terwijl

$$P(X = 4)P(Y = 10) = 0.35461 \times 0.35286 = 0.1251277,$$

zodat

$$P(X = 4 \text{ en } Y = 10) \neq P(X = 4)P(Y = 10).$$

De Score en Leeftijd zijn bijgevolg *afhankelijke variabelen*.

Tot slot kunnen we ook de *covariantie* en de *correlatiecoëfficiënt* definiëren voor de populatie.

! De **covariantie** voor twee discrete variabelen X en Y in een populatie (symbool $COV(X,Y)$), wordt gegeven door:

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

! De **correlatiecoëfficiënt** (symbool ρ_{XY}) wordt gegeven door:

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

met σ_X de standaarddeviatie van X en σ_Y de standaarddeviatie van Y.

5.4.2 Continue variabelen

De bivariate verdeling van continue variabelen vereist een grondige kennis van integralen. Dit maakt echter geen deel uit van deze syllabus en daarom zullen we deze bivariate verdeling slechts beknopt bespreken.

Voor continue variabelen zijn de kansen $P(X = x_i \text{ en } Y = y_i) = 0$. Gelijkaardig als in het univariate geval, kunnen we wel de *cumulatieve bivariate verdelingsfunctie* definiëren:

$$F_{X,Y}(x,y) = P(X \leq x \text{ en } Y \leq y).$$

De *bivariate dichtheidsfunctie* bekomen we door $F_{X,Y}(x,y)$ af te leiden en noteren we symbolisch als $f_{X,Y}(x,y)$.

Twee continue variabelen X en Y zijn *onafhankelijk* als geldt dat:

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

voor alle mogelijke waarden x en y. Vermits er oneindig veel dergelijke waarden bestaan, kunnen we dit niet nagaan zoals voor discrete variabelen. We zullen dit in deze syllabus niet verder bespreken. Tot slot definiëren we de covariantie en correlatie voor twee continue variabelen.

! De **covariantie** voor twee continue variabelen X en Y in een populatie (symbool $COV(X,Y)$), wordt gegeven door:

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

! De **correlatiecoëfficiënt** (symbool ρ_{XY}) wordt gegeven door:

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

met σ_X de standaarddeviatie van X en σ_Y de standaarddeviatie van Y.

5.5 Nuttige stellingen

In deze paragraaf geven we enkele nuttige stellingen die later van pas zullen komen. We zullen ze echter niet bewijzen. De stellingen gelden voor zowel discrete als continue variabelen.

Stelling 1. *Als X en Y onafhankelijke variabelen zijn, dan geldt dat:*

$$COV(X,Y) = 0.$$

De populatiecovariantie van twee variabelen is altijd nul indien de variabelen onafhankelijk zijn. Het omgekeerde geldt echter niet: een covariantie van nul impliceert niet dat de variabelen onafhankelijk zijn. Analoog als bij de steekproefcovariantie is de populatiecovariante een maat voor *lineaire* samenhang. Een covariantie van 0 kan bekomen worden indien er een samenhang is die niet lineair is.

De volgende stelling maakt gebruik van de eigenschap dat de verwachtingswaarde (dus het gemiddelde) van een constante gelijk is aan de constante: $E(a) = a$ voor een constante a. Dit is ook logisch: stel dat iedereen in een populatie hetzelfde inkomen heeft, bijvoorbeeld €2000, dan zal het gemiddelde inkomen ook €2000 bedragen.

Stelling 2. *Voor een variabele $Y = X + a$ geldt dat:*

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

waarbij a een constante is.

Stel dat iedereen in een populatie €100 opslag krijgt en we wensen het gemiddelde inkomen na deze loonopslag te berekenen. Stelling 2 impliceert dat we dit op twee manieren kunnen bekomen:

- we berekenen eerst het gemiddelde van de oorspronkelijke lonen ($E(X)$) en tellen er dan 100 bij op ($E(X) + 100$).
- we bepalen de lonen na de loonopslag ($Y = X + 100$) en we berekenen dan het gemiddelde $E(Y)$.

Stelling 3. *Voor een variabele $Y = aX$ geldt dat:*

$$E(Y) = aE(X),$$

waarbij a een constante is.

Stel dat we gegevens hebben over het inkomen van een populatie, uitgedrukt in Euro (variabele X) en we wensen het gemiddelde inkomen te weten uitgedrukt in Dollar (variabele Y). Stelling 3 impliceert dat we dit op twee manieren kunnen bekomen:

- we zetten eerst alle inkomens om naar Dollar (als wisselkoers gebruiken we €1 = $1.13, dus $Y = 1.13 \times X$) en berekenen dan het gemiddelde $E(Y)$.
- we berekenen eerst het gemiddelde van de inkomens uitgedrukt in Euro ($E(X)$) en zetten dan dit gemiddelde om naar Dollar ($1.13 \times E(X)$).

Stelling 4. *Voor twee variabelen X en Y (die onafhankelijk of afhankelijk kunnen zijn) geldt dat:*

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

en

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

Stel dat personen in een populatie twee verschillende testen moeten afleggen: X stelt de score voor op de eerste test en Y de score op de tweede test. Stelling 4 stelt dat we op twee manieren het gemiddelde van de totale score kunnen bekomen:

- we tellen eerst per persoon de scores op ($X+Y$) en berekenen dan het gemiddelde ($E(X + Y)$).
- we berekenen eerst per test het gemiddelde ($E(X)$ en $E(Y)$) en tellen dan deze gemiddelden op ($E(X) + E(Y)$).

Stelling 5. *Voor twee onafhankelijke variabelen X en Y geldt dat:*

$$E(XY) = E(X)E(Y).$$

Stellingen 2, 3 en 4 gaan ook op voor het steekproefgemiddelde, terwijl dit niet zo is voor Stellingen 1 en 5. Dit zullen we niet bewijzen in deze cursus.

De volgende stelling maakt gebruik van de eigenschap dat de variantie van een constante gelijk is aan nul: $V(a) = 0$ voor een constante a. Dit is niet verwonderlijk: als iedereen in een populatie hetzelfde inkomen heeft, is er geen spreiding en is de variantie nul.

Stelling 6. *Voor een variabele $Y = X + a$ geldt dat:*

$$V(Y) = V(X),$$

waarbij a een constante is.

Het optellen van een constante bij een variabele heeft dus geen invloed op de variantie. Als iedereen in een populatie €100 opslag krijgt, dan zal dit niets wijzigen aan de spreiding: het verschil tussen bijvoorbeeld de armste en de rijkste zal nog steeds dezelfde zijn.

Stelling 7. *Voor een variabele $Y = aX$ geldt dat:*

$$V(Y) = a^2V(X),$$

waarbij a een constante is.

Het vermenigvuldigen van een variabele met een constante heeft wel een invloed op de variantie. Als we het inkomen omzetten van Euro naar Dollar zal dit een invloed hebben op de spreiding. Beschouw bijvoorbeeld de lonen €2000 en €2100: het verschil is 100. Omgezet naar Dollar wordt dit $2260 en $2373 met een verschil van 113. De spreiding is bijgevolg groter geworden wat zal resulteren in een toename van de variantie.

Stelling 8. *Voor twee variabelen X en Y geldt dat*

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

De variantie van de som is gelijk aan de som van de varianties *plus* twee keer de covariantie.

Stel dat personen in een populatie twee verschillende testen moeten afleggen: X stelt de score voor op de eerste test en Y de score op de tweede test. Stelling 8 stelt dat we op twee manieren de variantie van de totale score kunnen bekomen:

- we tellen eerst per persoon de scores op $(X + Y)$ en berekenen dan de variantie $(V(X + Y))$.
- we berekenen eerst per test de variantie ($V(X)$ en $V(Y)$) samen met de covariantie $(COV(X, Y))$ en tellen dit dan op $(V(X) + V(Y) + 2COV(X, Y))$.

Als de covariantie positief is, impliceert dit dat de variantie van de totale score groter is dan de som van de varianties van de afzonderlijke scores. Dit komt doordat personen die een hoge score behalen op de eerste test, vaak ook een hoge score behalen op de tweede test (omwille van de positieve covariantie). Omgekeerd zullen personen die een lage score behalen op de eerste test, vaak ook een lage score behalen op de tweede. Bijgevolg zal de totale score meer spreiding vertonen en is de variantie dus groter.

Indien X en Y onafhankelijke variabelen zijn, dan volgt uit Stelling 1 en Stelling 8 dat

$$V(X + Y) = V(X) + V(Y). \tag{5.15}$$

Stelling 9. *Voor twee variabelen X en Y geldt dat:*

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

De variantie van het verschil is gelijk aan de som van de varianties *min* twee keer de covariantie.

Indien X en Y onafhankelijke variabelen zijn, dan volgt uit Stelling 1 en Stelling 9 dat:

$$V(X - Y) = V(X) + V(Y).$$

De variantie van het verschil is gelijk aan de *som* van de varianties. De variantie van het verschil is dus *niet* gelijk aan het verschil van de varianties. Dit is niet verwonderlijk, want stel dat de variantie van Y groter is dan de variantie van X dan zou dit impliceren dat de variantie van het verschil negatief is (als $V(X) < V(Y)$ dan is $V(X) - V(Y) < 0$) wat niet mogelijk is: de variantie is altijd groter dan of gelijk aan nul.

Stellingen 6, 7, 8 en 9 gaan ook op voor de steekproefvariantie. Dit zullen we niet bewijzen in deze cursus.

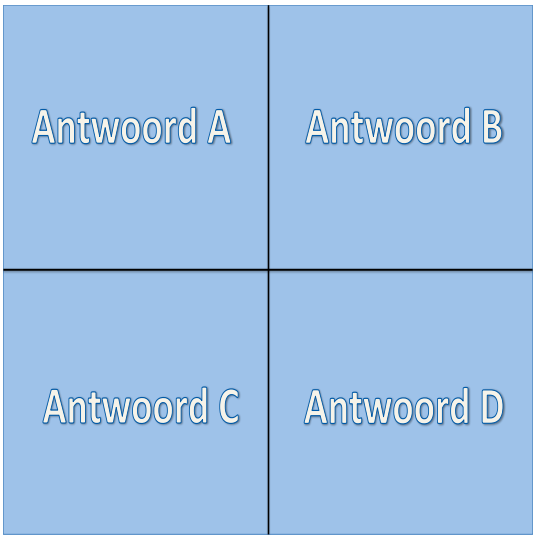
5.6 Bijzondere verdelingen

In deze paragraaf bespreken we enkele belangrijke kansverdelingen.

5.6.1 De binomiale verdeling

We illustreren de binomiale verdeling aan de hand van een meerkeuze-examen. Stel dat een grote populatie studenten (theoretisch gezien een oneindig grote populatie) een meerkeuze-examen moet afleggen. Elke vraag heeft vier antwoordmogelijkheden (A, B, C of D) en telkens is één antwoord correct. De studenten krijgen de opdracht te gokken op alle vragen: ze moeten op *willekeurige wijze* een antwoord aanduiden.

De *binomiale verdeling* zal de kansverdeling weergeven om k correcte antwoorden te hebben op een examen met N vragen. Laten we starten met een eenvoudige setting waar de studenten slechts één vraag krijgen. Omdat alle studenten op willekeurige wijze een antwoord moeten aanduiden en omdat er vier antwoordmogelijkheden zijn, zal een kwart van de populatie antwoord A aanduiden, een kwart antwoord B, een kwart antwoord C en een kwart antwoord D. Voor de eenvoud veronderstellen we dat antwoord A het correcte antwoord is. De kans om het antwoord correct te hebben is dus 0.25. Inderdaad, 25% van de populatie heeft antwoord A gegeven. Figuur 5.13 geeft dit schematisch weer. Symbolisch schrijven we deze kans als p, dus $p = 0.25$. De kans p wordt ook de kans op 'succes' genoemd. Hier komt een succes overeen met een vraag correct beantwoorden.



Figuur 5.13: Als men bij één meerkeuzevraag met 4 antwoorden volledig willekeurig een antwoord aanduidt, zal een kwart van de populatie antwoord A geven, een kwart antwoord B, een kwart antwoord C en een kwart antwoord D.

Stel dat de studenten vervolgens een tweede vraag moeten beantwoorden (dus in totaal is $N = 2$), opnieuw op volledig willekeurige wijze (zonder rekening te houden met het antwoord op de vorige vraag). Van alle studenten die antwoord A gegeven hebben op de

eerste vraag, zal een kwart antwoord A geven op de tweede vraag, een kwart antwoord B, een kwart antwoord C en een kwart antwoord D. Idem voor de andere antwoorden. Figuur 5.14 illustreert dit. De populatie wordt onderverdeeld in 16 gelijke groepen op basis van de antwoorden op beide vragen.

Stel dat bij de tweede vraag ook antwoord A correct is en beschouw de variabele X: *de totale score op het examen met N vragen*. Wat is de kansverdeling van deze variabele?

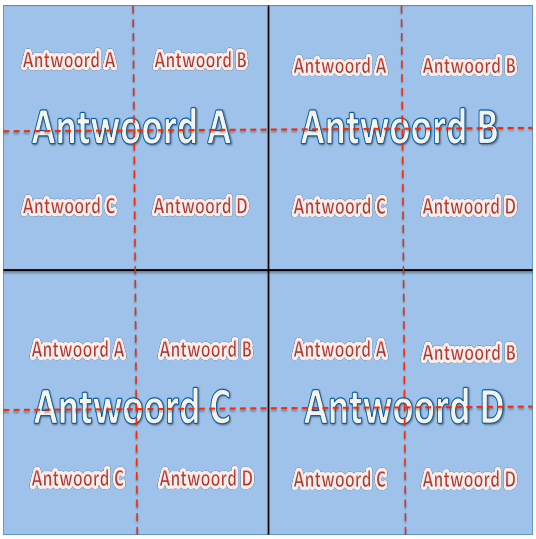
Omdat $N = 2$ kan de variabele X drie mogelijke waarden aannemen:

$$X = \begin{cases} 0 & \text{indien beide antwoorden foutief zijn.} \\ 1 & \text{indien één antwoord correct en één antwoord foutief is.} \\ 2 & \text{indien beide antwoorden correct zijn.} \end{cases}$$

De variabele X kunnen we ook interpreteren als het aantal 'successen' bij N vragen. Vervolgens bekijken we de bijhorende kansen.

- De kans $P(X = 0)$ komt overeen met de proportie studenten die een foutief antwoord hebben gegeven op beide vragen. Dit zijn de studenten die antwoorden B, C of D hebben aangeduid op de eerste vraag en antwoorden B, C of D op de tweede vraag. Dit komt overeen met 9 van de 16 groepen in Figuur 5.14. Bijgevolg is:
$$P(X = 0) = \frac{9}{16}.$$
- De kans $P(X = 1)$ komt overeen met de proportie studenten die één correct en één foutief antwoord hebben gegeven. Dit zijn de studenten die antwoord A hebben gegeven op de eerste vraag en antwoord B, C of D op de tweede vraag én de studenten die antwoord A hebben gegeven op de tweede vraag en antwoord B, C of D op de eerste vraag. Dit komt overeen met 6 van de 16 groepen in Figuur 5.14. Bijgevolg is:
$$P(X = 1) = \frac{6}{16}.$$
- De kans $P(X = 2)$ komt overeen met de proportie studenten in de populatie die zowel op de eerste als de tweede vraag antwoord A hebben gegeven. Dit komt overeen met 1 van de 16 groepen in Figuur 5.14. Bijgevolg is:
$$P(X = 2) = \frac{1}{16}.$$

Indien we de kansverdeling van X voor een examen met 3 meerkeuzevragen willen opstellen ($N = 3$), kunnen we analoog te werk gaan: we verdelen elk van de 16 groepen



Figuur 5.14: Als men bij twee meerkeuzevragen met 4 antwoorden volledig willekeurig een antwoord aanduidt, zal een kwart van de populatie op de eerste vraag antwoord A geven, een kwart antwoord B, een kwart antwoord C en een kwart antwoord D. Van de personen die antwoord A gegeven hebben op de eerste vraag, zal een kwart antwoord A geven op de tweede vraag, een kwart antwoord B, een kwart antwoord C en een kwart antwoord D. Idem voor de andere antwoorden.

in 4 kleinere groepen en we tellen het aantal correcte antwoorden op. Dit kunnen we dan herhalen voor $N = 4$, $N = 5$, etc. Omdat dit een omslachtige werkwijze is, zullen we anders te werk gaan. We zullen gebruik maken van een wiskundige formule om de kansverdeling van X te berekenen: *de binomiale kansverdeling*. Deze wordt gegeven door:

$$P(X = k) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}, \tag{5.16}$$

waar het symbool $N!$ staat voor N *faculteit*[b], p voor de kans op succes, k is een gegeven geheel getal en N voor het maximaal aantal successen.

Deze formule laat ons toe om de kansverdeling van X te berekenen, zonder dat we groepen moeten tellen zoals in Figuur 5.14.

De kansen die we eerder bekomen hebben voor een meerkeuze-examen met 2 vragen, kunnen we bekomen door formule (5.16) toe te passen $N = 2$ en $p = 0.25$ en k gelijk aan 0, 1 of 2:

$$P(X = 0) = \frac{2!}{0!(2-0)!} 0.25^0 (1-0.25)^{2-0} = 0.75^2 = 0.5625 = \frac{9}{16}.$$

In voorgaande berekening hebben we gebruik gemaakt van de eigenschap dat $x^0 = 1$

[b] $N! = N \times (N-1) \times (N-2) \times \ldots \times 2 \times 1$ en $0! = 1$. Bijvoorbeeld $4! = 4 \times 3 \times 2 \times 1 = 24$.

voor alle reële getallen x, vb. $0.25^0 = 1$. Analoog voor de andere kansen:

$$P(X = 1) = \frac{2!}{1!(2-1)!}0.25^1(1-0.25)^{2-1} = 0.375 = \frac{6}{16},$$

en

$$P(X = 2) = \frac{2!}{2!(2-2)!}0.25^2(1-0.25)^{2-2} = 0.0625 = \frac{1}{16}.$$

Een variabele die een binomiale verdeling heeft, noemen we ook een *binomiale variabele* en we noteren dit symbolisch als $X \sim Binom(N, p)$. Figuur 5.15 toont de staafdiagrammen van de binomiale kansverdeling voor verschillende keuzes van N en p. De figuur linksboven geeft bijvoorbeeld de score op een examen met 5 vragen waarbij er bij elke vraag 10 antwoordmogelijkheden zijn, zodat de kans op succes gelijk is aan $p = 0.10$. De kans om bijvoorbeeld alle vragen foutief te beantwoorden is veel groter dan de kans om alle vragen correct te beantwoorden. Merk op dat voor $p = 0.5$ de verdeling perfect symmetrisch is, terwijl voor $p = 0.1$ en $p = 0.9$ de verdeling scheef is (scheef naar rechts voor $p = 0.1$ en scheef naar links voor $p = 0.9$).

De verwachtingswaarde van een binomiale variabele $X \sim Binom(N, p)$ wordt gegeven door:

$$E(X) = Np, \tag{5.17}$$

en de variantie door:

$$V(X) = Np(1-p). \tag{5.18}$$

We zullen deze eigenschappen niet bewijzen, maar we kunnen ze wel illustreren aan de hand van het meerkeuze-examen met 2 vragen ($N = 2$ en $p = 0.25$). Uit formule (5.17) volgt:

$$E(X) = 2 \times 0.25 = 0.5.$$

Als we vervolgens de algemene formule van de verwachtingswaarde toepassen, formule (5.10) op pagina 154, bekomen we inderdaad dezelfde uitkomst:

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2) = 0 \times \frac{9}{16} + 1 \times \frac{6}{16} + 2 \times \frac{1}{16} = 0.5.$$

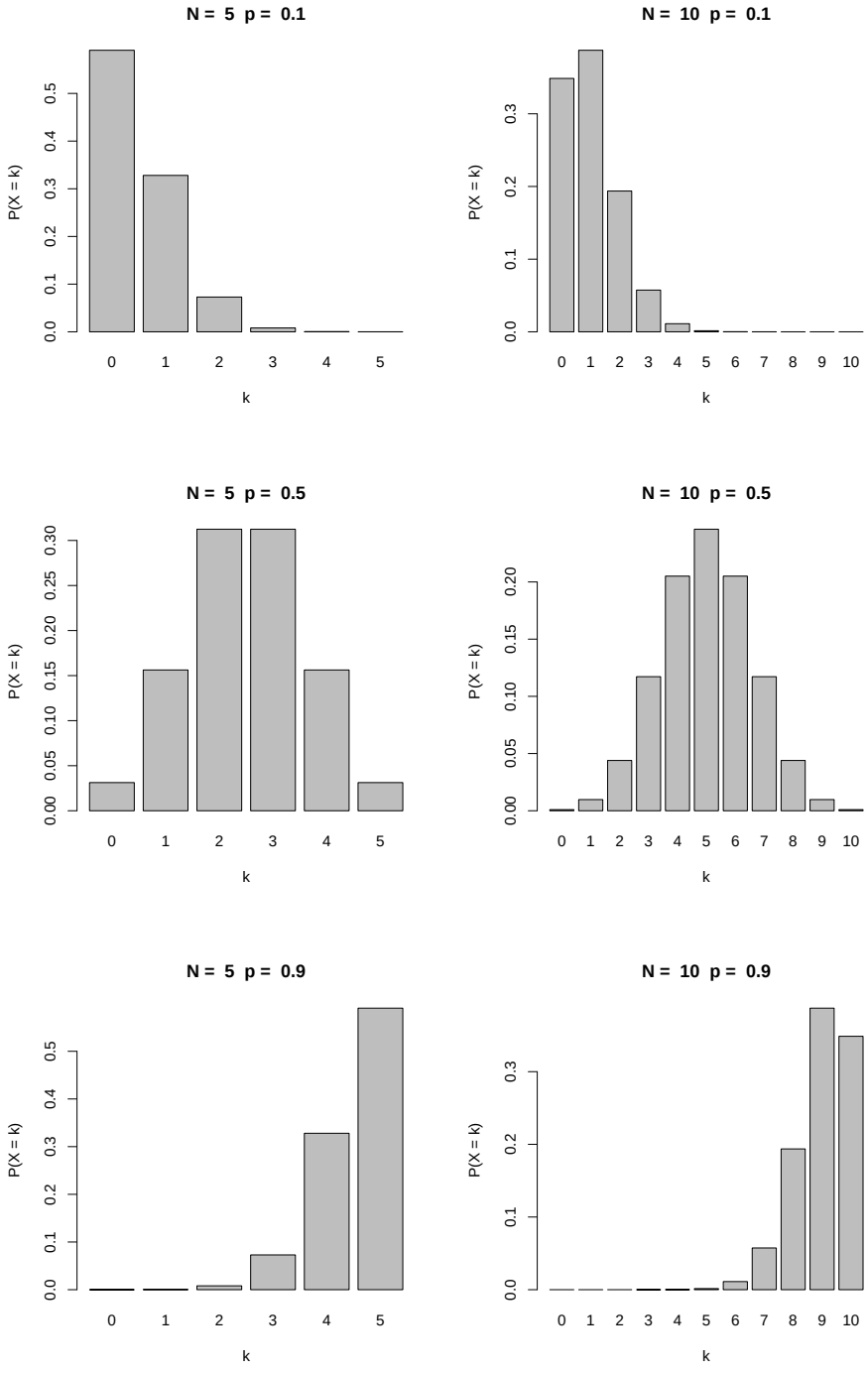
De gemiddelde score van de studenten in de populatie is dus 0.5 op 2.

Op gelijkaardige wijze kan je de variantie berekenen via formule (5.18):

$$V(X) = 2 \times 0.25 \times (1 - 0.25) = 0.375.$$

Op basis van formule (5.12) op pagina 155 bekomen we dezelfde uitkomst:

$$V(X) = P(X = 0) \times (0-0.5)^2 + P(X = 1) \times (1-0.5)^2 + P(X = 2) \times (2-0.5)^2 = 0.375.$$



Figuur 5.15: De kansverdeling van een binomiale variabele $X \sim Binom(N, p)$ voor verschillende keuzes van N en p.

De binomiale verdeling kan enkel gebruikt worden als N vast is en indien de kans op succes p ongewijzigd blijft. Voor het meerkeuze-examen is dit inderdaad het geval: N is het aantal vragen en ligt op voorhand vast en voor elke vraag is de kans op succes gelijk aan $p = 0.25$ (omdat de studenten moeten gokken), ze blijft dus ongewijzigd.

Illustratie in R

De kansverdeling $P(X = k)$ kunnen we bekomen met het commando `dbinom(k, N, p)`. We illustreren dit voor $N = 2$, $p = 0.25$ en $k = 0$, $k = 1$, $k = 2$:

```
> dbinom(0, 2, 0.25)

[1] 0.5625

> dbinom(1, 2, 0.25)

[1] 0.375

> dbinom(2, 2, 0.25)

[1] 0.0625
```

De cumulatieve verdelingsfunctie $P(X \leq k)$ wordt bekomen via het commando `pbinom(k, N, p)`:

```
> pbinom(0, 2, 0.25)

[1] 0.5625

> pbinom(1, 2, 0.25)

[1] 0.9375

> pbinom(2, 2, 0.25)
```

```
[1] 1
```

Merk op dat bijvoorbeeld $P(X \leq 1) = P(X = 0) + P(X = 1)$:

```
> pbinom(1, 2, 0.25)
```

```
[1] 0.9375
```

```
> dbinom(0, 2, 0.25) + dbinom(1, 2, 0.25)
```

```
[1] 0.9375
```

5.6.2 De normale verdeling

Binnen de statistiek zijn *normaal verdeelde variabelen* zeer belangrijk. Enerzijds blijkt dat de normale verdeling een goede benadering is voor verschillende verdelingen in de praktijk en anderzijds is ze zeer nuttig omwille van de *centrale limietstelling*. We komen later terug op deze stelling.

Een normaal verdeelde variabele is continu en de dichtheidsfunctie wordt gegeven door:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \tag{5.19}$$

met $\pi = 3.14\ldots$ en $e = 2.71\ldots$. Een variabele die normaal verdeeld is, noteren we als $X \sim N(\mu, \sigma^2)$. De dichtheidsfunctie hangt af van *twee parameters*: μ en σ^2, waarvoor geldt dat (zonder bewijs):

$$E(X) = \mu,$$

en

$$V(X) = \sigma^2.$$

Deze gelijkheden laten ons toe een interpretatie te geven aan de parameters: μ is het populatiegemiddelde en σ^2 is de populatievariantie.

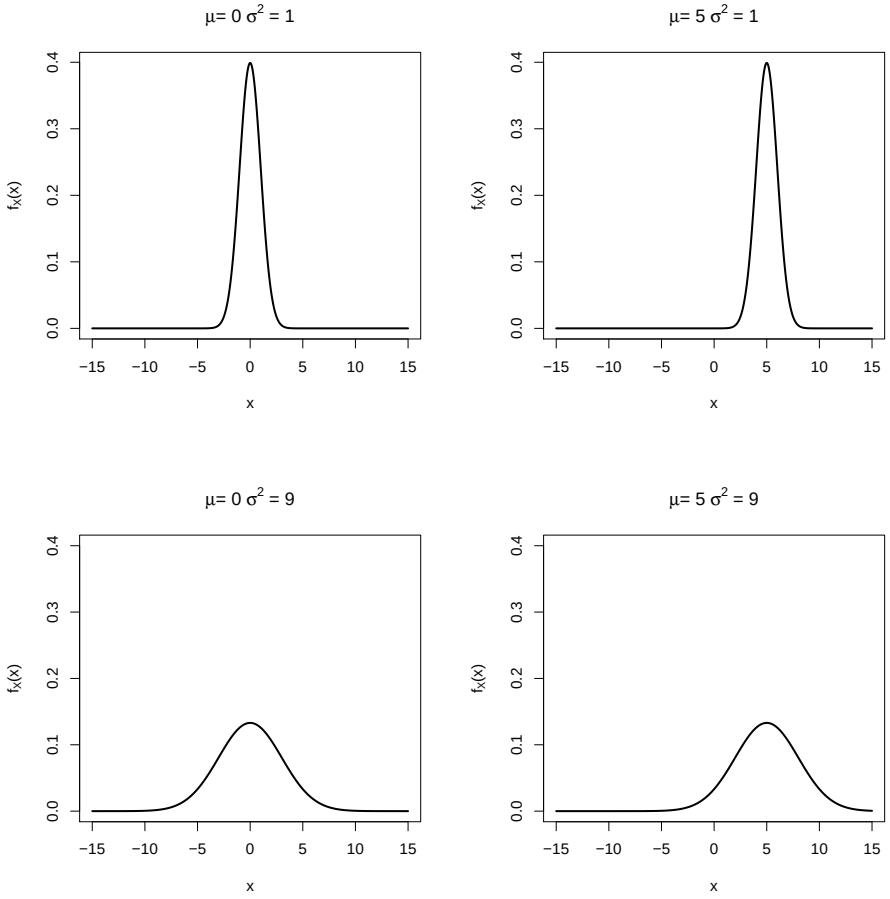
Voor elke keuze van μ en σ^2 bekomen we een andere dichtheidsfunctie. We illustreren dit voor $x = 0$ en $x = 2$. Indien $\mu = 0$ en $\sigma^2 = 1$ dan volgt uit formule (5.19) dat

$$f(0) = \frac{1}{1 \times \sqrt{2\pi}} e^{\frac{-(0-0)^2}{2\times 1^2}} = 0.3989 \quad \text{en} \quad f(2) = \frac{1}{1 \times \sqrt{2\pi}} e^{\frac{-(2-0)^2}{2\times 1^2}} = 0.0540.$$

Indien we bijvoorbeeld $\mu = 1.5$ en $\sigma^2 = 1$ kiezen, dan krijgen we de waarden

$$f(0) = \frac{1}{1 \times \sqrt{2\pi}} e^{\frac{-(0-1.5)^2}{2 \times 1^2}} = 0.1295 \quad \text{en} \quad f(2) = \frac{1}{1 \times \sqrt{2\pi}} e^{\frac{-(2-1.5)^2}{2 \times 1^2}} = 0.3521.$$

Figuur 5.16 illustreert dit grafisch en toont de dichtheidsfunctie voor verschillende waarden van μ en σ^2. Merk op dat de dichtheidsfunctie haar hoogste punt bereikt in het gemiddelde. Bij een grotere variantie σ^2 (dus bij meer spreiding rond het gemiddelde) wordt de dichtheidsfunctie breder en minder hoog. Hoewel dit moeilijk af te leiden is uit de figuren, wordt de dichtheidsfunctie nergens nul omdat $f_X(x) > 0$ voor alle waarden x.



Figuur 5.16: De dichtheidsfunctie van een normaal verdeelde variabele $X \sim N(\mu, \sigma^2)$ voor verschillende keuzes van μ en σ^2.

Zoals aangegeven in formule (5.4) gebruiken we de dichtheidsfunctie om kansen van de vorm $P(x_1 \leq X \leq x_2)$ te berekenen. Voor de normale verdeling moeten we de volgende integraal berekenen:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx. \tag{5.20}$$

Deze integraal kunnen we echter niet analytisch oplossen. Om deze kansen te berekenen, zullen we beroep doen op tabellen waarvoor deze integraal met behulp van een computer is uitgerekend. Voor elke keuze van μ en σ^2 is de integraal (5.20) echter verschillend en hebben we bijgevolg andere tabellen nodig. Omdat μ en σ^2 oneindig veel verschillende waarden kunnen aannemen, hebben we in principe een oneindig aantal tabellen nodig. Dit is uiteraard niet mogelijk. Het blijkt echter dat een tabel voor $\mu = 0$ en $\sigma^2 = 1$ voldoende is om de kansen te berekenen voor *elke* normale verdeling (dus ook als $\mu \neq 0$ en/of $\sigma^2 \neq 1$). We illustreren dit in de rest van deze paragraaf.

De normale verdeling met $\mu = 0$ en $\sigma^2 = 1$ wordt ook de *standaardnormale verdeling* genoemd. Tabel 5.6 geeft enkele waarden weer voor de kansen:

$$P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx.$$

Uit deze tabel lezen we bijvoorbeeld af dat

$$P(X \leq -0.5) = 0.3085,$$

en

$$P(X \leq 1) = 0.8413.$$

Door gebruik te maken van formule (5.7) op pagina 150 kunnen we de kans $P(-0.5 \leq X \leq 1)$ afleiden:

$$P(-0.5 \leq X \leq 1) = P(X \leq 1) - P(X \leq -0.5) = 0.8413 - 0.3085 = 0.5328.$$

Zoals eerder aangegeven in paragraaf 5.2.2 kunnen we deze kansen ook visueel weergeven aan de hand van oppervlaktes onder de dichtheidskromme; zie Figuur 5.17.

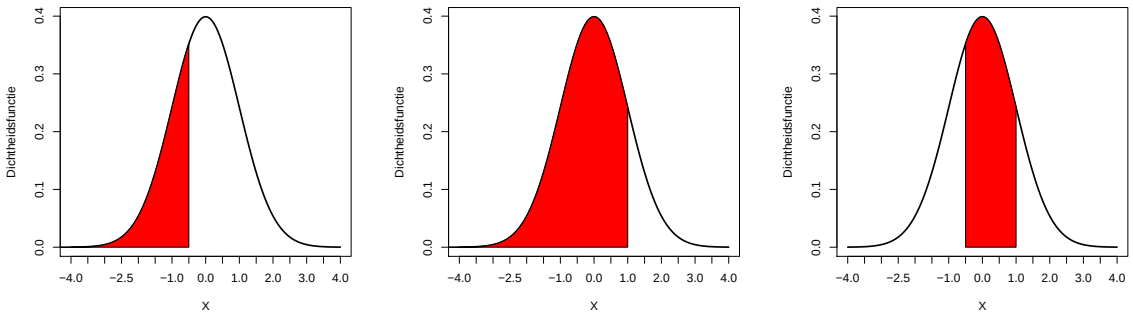
Uit Tabel 5.6 halen we dat $P(X \leq 0) = 0.5$. Dit komt omdat de standaardnormale verdeling *symmetrisch* is rond 0; zie Figuur 5.18. We zien dat het deel links van de rode volle lijn het spiegelbeeld is van het deel rechts van de rode volle lijn. Omdat de totale oppervlakte onder de dichtheidsfunctie altijd gelijk is aan 1 en de rode volle lijn de oppervlakte in 2 gelijke delen opdeelt, kunnen we afleiden dat de oppervlakte links (dus de kans $P(X \leq 0)$) gelijk is aan 0.5. Ook de oppervlakte rechts (dus de kans $P(X > 0)$) is gelijk aan 0.5.

Meer algemeen geldt voor de standaardnormale verdeling dat

$$P(X > x) = P(X \leq -x).$$

x	$P(X \leq x)$
-3.0	0.0013
-2.5	0.0062
-2.0	0.0228
-1.5	0.0668
-1.0	0.1587
-0.5	0.3085
0.0	0.5000
0.5	0.6915
1.0	0.8413
1.5	0.9332
2.0	0.9772
2.5	0.9938
3.0	0.9987

Tabel 5.6: De cumulatieve verdelingsfunctie van een standaardnormale variabele voor bepaalde waarden van x.

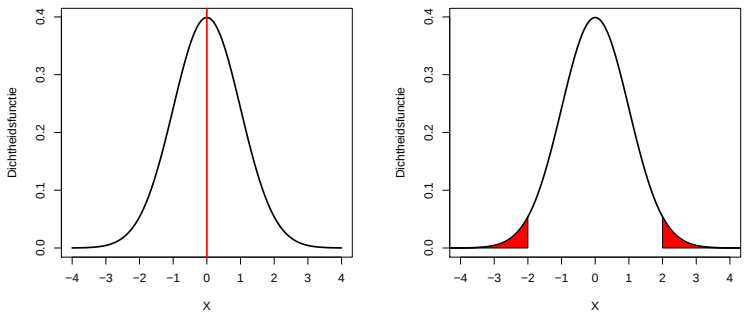


Figuur 5.17: De grafische weergave van de kansen $P(X \leq -0.5)$ (gekleurde oppervlakte figuur links), $P(X \leq 1)$ (gekleurde oppervlakte figuur midden) en $P(-0.5 \leq X \leq 1)$ (gekleurde oppervlakte figuur rechts). Merk op dat de gekleurde oppervlakte in de figuur rechts gelijk is aan het verschil in gekleurde oppervlakte van de figuur in het midden en de figuur links. Dit komt overeen met de eigenschap dat $P(-0.5 \leq X \leq 1) = P(X \leq 1) - P(X \leq -0.5)$.

Uit Tabel 5.6 kunnen we ook een verband tussen de kansen $P(X \leq -x)$ en $P(X \leq x)$ afleiden, namelijk:

$$P(X \leq -x) = 1 - P(X \leq x), \tag{5.21}$$

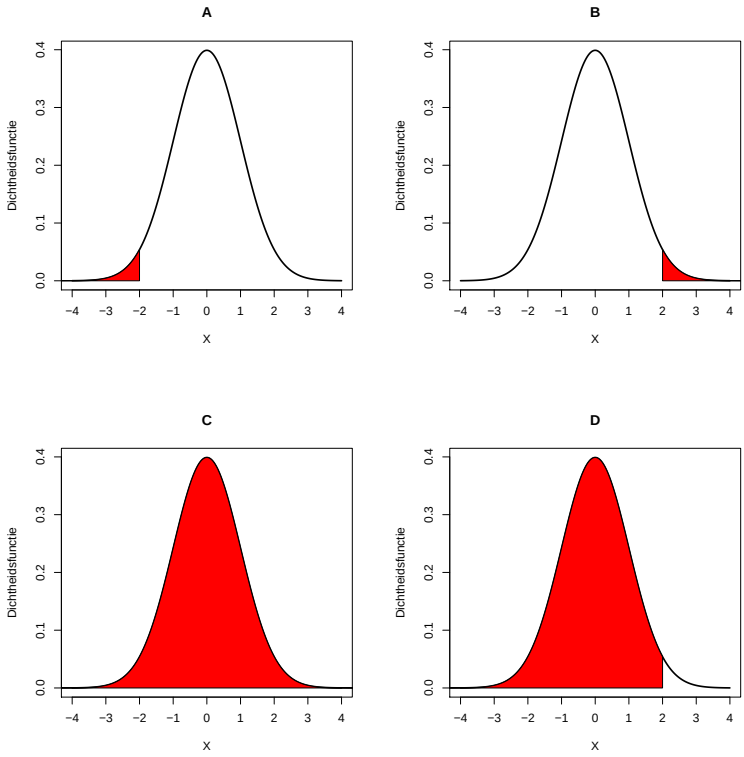
voor alle x. Bijvoorbeeld voor $x = 2$ zien we dat $P(X \leq -2) = 0.0228$ en $P(X \leq 2) = 0.9772$, zodat inderdaad $P(X \leq -2) = 1 - P(X \leq 2)$. Figuur 5.19 illustreert formule (5.21) door grafisch de oppervlaktes te bepalen.



Figuur 5.18: Links: de standaardnormale verdeling is symmetrisch rond 0: het deel links van de rode volle lijn is het spiegelbeeld van het deel rechts. Rechts: omwille van de symmetrie geldt dat $P(X \leq -x) = P(X > x)$. Hier illustreren we dit met de waarde $x = 2$ waar $P(X \leq -2)$ de oppervlakte links is en $P(X > 2)$ de oppervlakte rechts is. Beide gekleurde oppervlaktes zijn gelijk aan elkaar.

Formule (5.21) is nuttig omdat ze impliceert dat we Tabel 5.6 meer beknopt kunnen schrijven door enkel de kansen $P(X \leq x)$ te geven waarvoor x positief is. Indien we een kans wensen te berekenen voor een negatieve x, dan kunnen we beroep doen op formule (5.21).

Tabel 5.6 heeft echter een nadeel: ze is niet nauwkeurig. Als we bijvoorbeeld de kans $P(X \leq 1.55)$ wensen te weten, kunnen we ze niet afleiden. Daarom zullen we beroep doen op een meer nauwkeurige tabel, namelijk Tabel 5.7. Omdat de tabel zeer veel getallen bevat en op één pagina moet worden kunnen weergegeven, is ze speciaal opgesteld. We kunnen ze als volgt gebruiken: vb. de kans $P(X \leq 1.55)$ lezen we af op de plaats waar de rij 1.5 de kolom 0.05 snijdt (merk op dat $1.5 + 0.05 = 1.55$). Hier komt dit overeen met 0.9394 en bijgevolg is $P(X \leq 1.55) = 0.9394$.



Figuur 5.19: Figuur A: grafische weergave van de kans $P(X \leq -2)$. Figuur B: grafische weergave van de kans $P(X > 2)$ (uit Figuur 5.18 weten we dat deze kans gelijk is aan $P(X \leq -2)$). Figuur C: de volledige oppervlakte is altijd gelijk aan 1. Figuur D: grafische weergave van de kans $P(X \leq 2)$. Merk op dat de gekleurde oppervlakte in figuur B gelijk is aan het verschil in gekleurde oppervlakte van figuur C en D (dus $P(X > 2) = 1 - P(X \leq 2)$) en dat de gekleurde oppervlakte van figuur B gelijk is aan de gekleurde oppervlakte van figuur A (dit is $P(X > 2) = P(X \leq -2)$). Dit kunnen we samenbrengen tot de eigenschap $P(X \leq -2) = 1 - P(X \leq 2)$.

x	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974

Tabel 5.7: De cumulatieve verdelingsfunctie van een standaardnormale variabele $X \sim N(0,1)$ voor bepaalde waarden van x.

Tabel 5.7 geeft enkel de cumulatieve verdelingsfunctie voor een *standaardnormale* variabele (waarvoor $\mu = 0$ en $\sigma^2 = 1$). De volgende stelling is cruciaal om deze tabel ook te kunnen gebruiken voor normale variabelen waarbij $\mu \neq 0$ of $\sigma^2 \neq 1$.

Stelling 10. *Als X een normale verdeling heeft met gemiddelde μ en variantie σ^2, dus $X \sim N(\mu, \sigma^2)$, dan heeft de variabele*

$$Z = \frac{X - \mu}{\sigma},$$

een standaardnormale verdeling, dus $Z \sim N(0, 1)$.

Deze stelling impliceert de volgende vergelijking: als $X \sim N(\mu, \sigma^2)$ dan geldt dat:

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right), \tag{5.22}$$

waarbij $Z \sim N(0, 1)$. Dit noemen we ook het *standaardiseren* van X.

Als voorbeeld nemen we een variabele $X \sim N(1, 4)$ waarvoor we de kans $P(X \leq 3)$ wensen te berekenen. We passen eerst formule (5.22) toe:

$$P(X \leq 3) = P\left(Z \leq \frac{3 - 1}{\sqrt{4}}\right) = P(Z \leq 1).$$

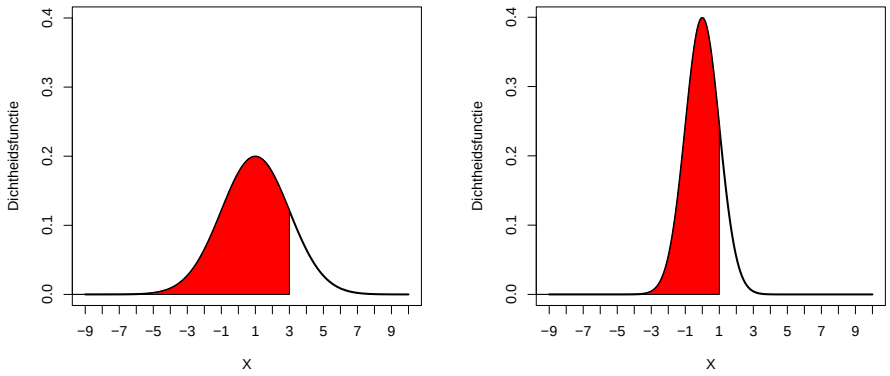
Omdat $Z \sim N(0, 1)$ kunnen we Tabel 5.7 gebruiken om deze kans af te lezen: $P(Z \leq 1) = 0.8413$. Bijgevolg is $P(X \leq 3) = 0.8413$. Figuur 5.20 illustreert dit grafisch.

Door gebruik te maken van formule (5.22) en Tabel 5.7, kunnen we voor elke normaal verdeelde variabele $X \sim N(\mu, \sigma^2)$ de kansen $P(X \leq x)$ berekenen.

Illustratie in R

Indien we beschikken over een computer, hebben we Tabel 5.7 niet nodig om de kansen te berekenen. Voor een standaardnormale variabele X, kunnen we rechtstreeks de kansen $P(X \leq x)$ vinden via het commando `pnorm(x)`:

```
> pnorm(1.55)

[1] 0.9394292
```



Figuur 5.20: De grafische weergave van de kans $P(X \leq 3)$ waar $X \sim N(1,4)$ (gekleurde oppervlakte figuur links) en de kans $P(Z \leq 1)$ met $Z \sim N(0,1)$ (gekleurde oppervlakte figuur rechts). Beide oppervlaktes zijn aan elkaar gelijk (dit is echter moeilijk op het zicht vast te stellen).

Via `dnorm(x)` kunnen we de kansdichtheid $f_X(x)$ berekenen:

```
> dnorm(1.55)

[1] 0.120009
```

Voor een variabele $X \sim N(1,4)$ kunnen we de kans $P(X \leq 3)$ in R ook direct berekenen:

```
> pnorm(3, mean = 1, sd = sqrt(4))

[1] 0.8413447
```

Via het argument `mean = 1` duiden we aan dat het gemiddelde van de normale verdeling $\mu = 1$ is en via `sd = sqrt(4)` duiden we aan dat de standaarddeviatie $\sigma = \sqrt{4} = 2$ is. Merk op dat we de *standaarddeviatie* moeten ingeven en dat de notatie $N(1,4)$ staat voor een normale verdeling met gemiddelde $\mu = 1$ en *variantie* $\sigma^2 = 4$.

De kans $P(X \leq 3)$ kunnen we ook bekomen via formule (5.22):

```
> z <- (3-1)/sqrt(4)
> pnorm(z)

[1] 0.8413447
```

5.6.3 De χ^2-verdeling

Laat $X_1, X_2, \ldots, X_k$ onafhankelijke standaardnormale variabelen zijn (dus $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, ..., $X_k \sim N(0,1)$). De χ^2_k-verdeling (uitspraak *chi-kwadraat*) is de verdeling van de variabele:

$$Y = X_1^2 + X_2^2 + \ldots + X_k^2.$$

De χ^2-verdeling is bijgevolg de verdeling van de som van k gekwadrateerde standaardnormale variabelen. De parameter k wordt *het aantal vrijheidsgraden* genoemd. Men kan aantonen dat:

$$E(Y) = k,$$

en

$$V(Y) = 2k.$$

Dit laat ons toe k te interpreteren als het populatiegemiddelde. Merk op dat voor een χ^2_k-verdeling de variantie steeds gelijk is aan twee maal het populatiegemiddelde.

Figuur 5.21 toont de dichtheidsfunctie van de χ^2_k-verdeling voor verschillende waarden van k. Een variabele Y die een χ^2_k-verdeling heeft, noteren we als $Y \sim \chi^2_k$.

Analoog als bij een normale verdeling, gebruiken we tabellen om kansen van de vorm $P(Y \leq y)$ te bepalen; zie Tabel 5.8. Het gebruik van deze tabel is echter anders dan die van de standaardnormale.

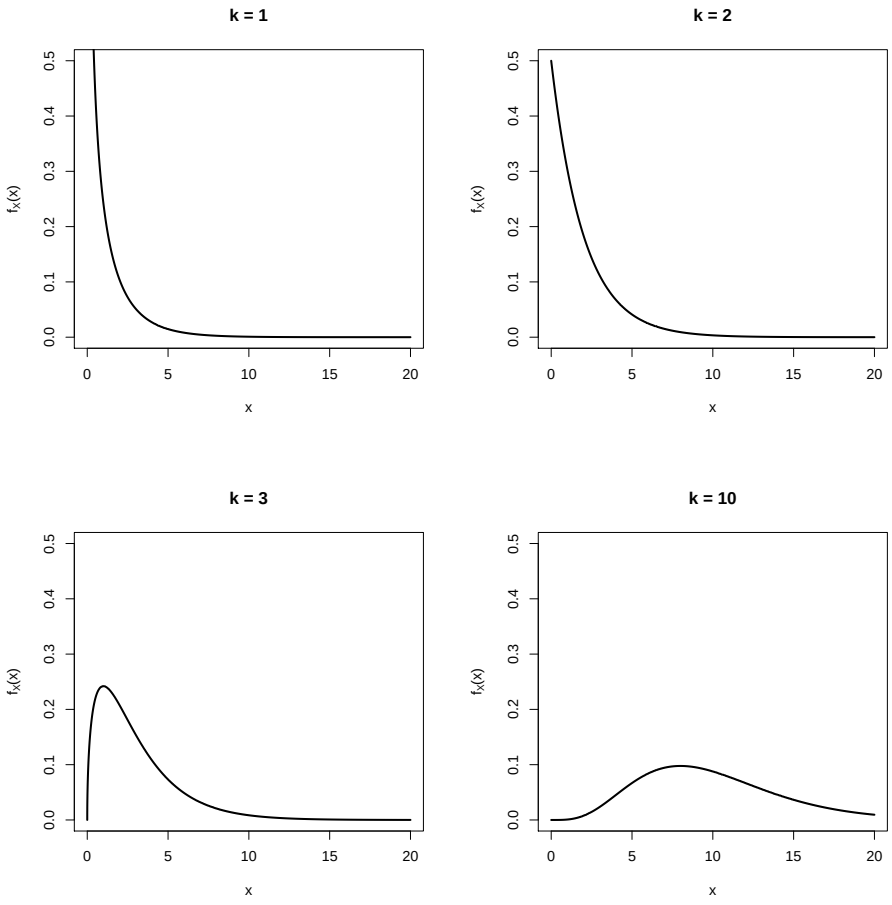
Ter illustratie: voor de variabele $Y \sim \chi^2_{28}$ geldt dat $F_Y(16.93) = 0.050$ omdat de waarde in rij 28 en kolom 0.050 gelijk is aan 16.93. Tabel 5.8 geeft maar een zeer beperkt aantal waarden van de verdelingsfunctie, maar voor deze cursus zal de tabel echter volstaan.

In tegenstelling tot de normale verdeling, komt de χ^2_k-verdeling minder vaak voor in de natuur. Ze zal echter wel belangrijk zijn voor Deel III van de syllabus en voor de cursus 'Statistiek II'.

Illustratie in R

Via het commando `pchisq(y, k)` kunnen we de kansen $P(Y \leq y)$ bekomen voor een variabele $Y \sim \chi^2_k$:

```
> pchisq(16.93, 28)

[1] 0.05004119
```



Figuur 5.21: De dichtheidsfunctie van een χ^2_k-verdeling voor verschillende waarden van k.

	$P(Y \leq y) = F_Y(y)$											
k	0.001	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995	0.999
1	0.00	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88	10.83
2	0.00	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60	13.82
3	0.02	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75	20.52
6	0.38	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55	22.46
7	0.60	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95	26.12
9	1.15	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59	27.88
10	1.48	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19	29.59
11	1.83	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76	31.26
12	2.21	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30	32.91
13	2.62	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82	34.53
14	3.04	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32	36.12
15	3.48	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80	37.70
16	3.94	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27	39.25
17	4.42	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72	40.79
18	4.90	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16	42.31
19	5.41	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58	43.82
20	5.92	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00	45.31
21	6.45	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40	46.80
22	6.98	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80	48.27
23	7.53	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18	49.73
24	8.08	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56	51.18
25	8.65	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93	52.62
26	9.22	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29	54.05
27	9.80	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64	55.48
28	10.39	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99	56.89
29	10.99	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34	58.30
30	11.59	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67	59.70
40	17.92	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77	73.40
50	24.67	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49	86.66
60	31.74	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95	99.61
70	39.04	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	104.21	112.32
80	46.52	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32	124.84
90	54.16	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12	128.30	137.21
100	61.92	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17	149.45

Tabel 5.8: De verdelingsfunctie $F_Y(y)$ van de χ^2_k-verdeling. De linkse kolom geeft het aantal vrijheidsgraden k. De bovenste rij geeft de waarden van de verdelingsfunctie $F_Y(y)$. In de tabel bevinden zich de waarden y van de variabele.

5.6.4 De t-verdeling

Laat $X \sim N(0,1)$ en $Y \sim \chi^2_k$ onafhankelijke variabelen zijn. De t_k-verdeling is de verdeling van de variabele

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

Net als bij de χ^2_k-verdeling wordt ook hier k het aantal vrijheidsgraden genoemd, maar hier zullen we geen interpretatie geven aan deze parameter. Figuur 5.22 toon de dichtheidsfunctie van de t_k-verdeling voor verschillende keuzes van k. De dichtheidsfunctie gelijkt op die van een normale verdeling, maar is niet volledig gelijk. Echter, naarmate k toeneemt, gelijkt de t_k-verdeling meer en meer op de dichtheid van een standaardnormale (als k naar oneindig convergeert, valt ze exact samen met de standaardnormale). Tabel 5.9 geeft de cumulatieve kansverdeling weer voor bepaalde waarden en wordt op dezelfde manier gelezen als Tabel 5.8. Met t_∞ duiden we in deze tabel de t_k-verdeling aan met 'oneindig' veel vrijheidsgraden - deze is exact gelijk aan de standaardnormale verdeling.

Als $T \sim t_k$ dan geldt dat

$$E(T) = 0,$$

en

$$V(T) = \frac{k}{k-2}, \quad \text{voor } k > 2.$$

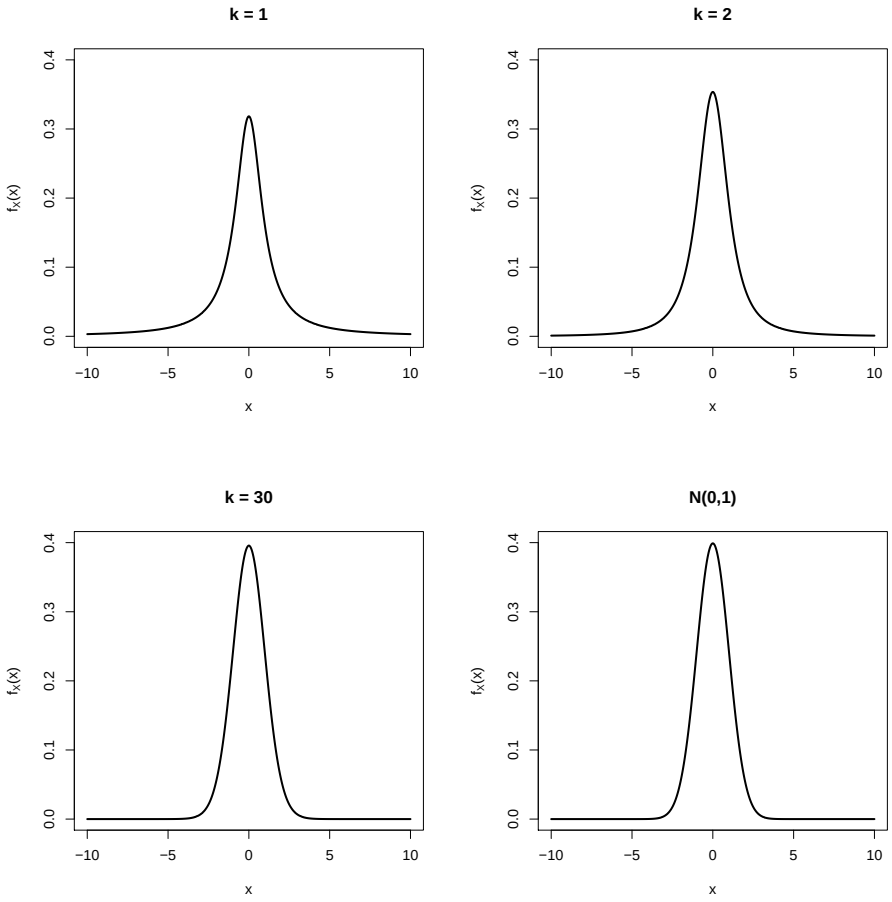
Analoog als de χ^2_k-verdeling komt de t_k-verdeling niet vaak voor in de natuur, maar is ze vooral van belang voor Deel III van de syllabus en voor de cursus 'Statistiek II'.

Illustratie in R

Via het commando `pt(t, k)` kunnen we de kansen $P(T \leq t)$ bekomen voor een variabele $T \sim t_k$:

```
> pt(1, 2)
```

```
[1] 0.7886751
```



Figuur 5.22: De dichtheidsfunctie van een t-verdeling met k vrijheidsgraden voor verschillende waarden van k. Naarmate k toeneemt, zal de dichtheidsfunctie meer gelijken op die van een standaard normale (figuur rechtsonder).

	$P(T \leq t) = F_T(t)$											
k	0.001	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995	0.999
1	−318.309	−63.657	−31.821	−12.706	−6.314	−3.078	3.078	6.314	12.706	31.821	63.657	318.309
2	−22.327	−9.925	−6.965	−4.303	−2.920	−1.886	1.886	2.920	4.303	6.965	9.925	22.327
3	−10.215	−5.841	−4.541	−3.182	−2.353	−1.638	1.638	2.353	3.182	4.541	5.841	10.215
4	−7.173	−4.604	−3.747	−2.776	−2.132	−1.533	1.533	2.132	2.776	3.747	4.604	7.173
5	−5.893	−4.032	−3.365	−2.571	−2.015	−1.476	1.476	2.015	2.571	3.365	4.032	5.893
6	−5.208	−3.707	−3.143	−2.447	−1.943	−1.440	1.440	1.943	2.447	3.143	3.707	5.208
7	−4.785	−3.499	−2.998	−2.365	−1.895	−1.415	1.415	1.895	2.365	2.998	3.499	4.785
8	−4.501	−3.355	−2.896	−2.306	−1.860	−1.397	1.397	1.860	2.306	2.896	3.355	4.501
9	−4.297	−3.250	−2.821	−2.262	−1.833	−1.383	1.383	1.833	2.262	2.821	3.250	4.297
10	−4.144	−3.169	−2.764	−2.228	−1.812	−1.372	1.372	1.812	2.228	2.764	3.169	4.144
11	−4.025	−3.106	−2.718	−2.201	−1.796	−1.363	1.363	1.796	2.201	2.718	3.106	4.025
12	−3.930	−3.055	−2.681	−2.179	−1.782	−1.356	1.356	1.782	2.179	2.681	3.055	3.930
13	−3.852	−3.012	−2.650	−2.160	−1.771	−1.350	1.350	1.771	2.160	2.650	3.012	3.852
14	−3.787	−2.977	−2.624	−2.145	−1.761	−1.345	1.345	1.761	2.145	2.624	2.977	3.787
15	−3.733	−2.947	−2.602	−2.131	−1.753	−1.341	1.341	1.753	2.131	2.602	2.947	3.733
16	−3.686	−2.921	−2.583	−2.120	−1.746	−1.337	1.337	1.746	2.120	2.583	2.921	3.686
17	−3.646	−2.898	−2.567	−2.110	−1.740	−1.333	1.333	1.740	2.110	2.567	2.898	3.646
18	−3.610	−2.878	−2.552	−2.101	−1.734	−1.330	1.330	1.734	2.101	2.552	2.878	3.610
19	−3.579	−2.861	−2.539	−2.093	−1.729	−1.328	1.328	1.729	2.093	2.539	2.861	3.579
20	−3.552	−2.845	−2.528	−2.086	−1.725	−1.325	1.325	1.725	2.086	2.528	2.845	3.552
21	−3.527	−2.831	−2.518	−2.080	−1.721	−1.323	1.323	1.721	2.080	2.518	2.831	3.527
22	−3.505	−2.819	−2.508	−2.074	−1.717	−1.321	1.321	1.717	2.074	2.508	2.819	3.505
23	−3.485	−2.807	−2.500	−2.069	−1.714	−1.319	1.319	1.714	2.069	2.500	2.807	3.485
24	−3.467	−2.797	−2.492	−2.064	−1.711	−1.318	1.318	1.711	2.064	2.492	2.797	3.467
25	−3.450	−2.787	−2.485	−2.060	−1.708	−1.316	1.316	1.708	2.060	2.485	2.787	3.450
26	−3.435	−2.779	−2.479	−2.056	−1.706	−1.315	1.315	1.706	2.056	2.479	2.779	3.435
27	−3.421	−2.771	−2.473	−2.052	−1.703	−1.314	1.314	1.703	2.052	2.473	2.771	3.421
28	−3.408	−2.763	−2.467	−2.048	−1.701	−1.313	1.313	1.701	2.048	2.467	2.763	3.408
29	−3.396	−2.756	−2.462	−2.045	−1.699	−1.311	1.311	1.699	2.045	2.462	2.756	3.396
30	−3.385	−2.750	−2.457	−2.042	−1.697	−1.310	1.310	1.697	2.042	2.457	2.750	3.385
40	−3.307	−2.704	−2.423	−2.021	−1.684	−1.303	1.303	1.684	2.021	2.423	2.704	3.307
50	−3.261	−2.678	−2.403	−2.009	−1.676	−1.299	1.299	1.676	2.009	2.403	2.678	3.261
60	−3.232	−2.660	−2.390	−2.000	−1.671	−1.296	1.296	1.671	2.000	2.390	2.660	3.232
70	−3.211	−2.648	−2.381	−1.994	−1.667	−1.294	1.294	1.667	1.994	2.381	2.648	3.211
80	−3.195	−2.639	−2.374	−1.990	−1.664	−1.292	1.292	1.664	1.990	2.374	2.639	3.195
90	−3.183	−2.632	−2.368	−1.987	−1.662	−1.291	1.291	1.662	1.987	2.368	2.632	3.183
100	−3.174	−2.626	−2.364	−1.984	−1.660	−1.290	1.290	1.660	1.984	2.364	2.626	3.174
200	−3.131	−2.601	−2.345	−1.972	−1.653	−1.286	1.286	1.653	1.972	2.345	2.601	3.131
∞	−3.090	−2.576	−2.326	−1.960	−1.645	−1.282	1.282	1.645	1.960	2.326	2.576	3.090

Tabel 5.9: De verdelingsfunctie $F_T(t)$ van de t_k-verdeling. De linkse kolom geeft het aantal vrijheidsgraden k. De bovenste rij geeft de waarden van de verdelingsfunctie $F_T(t)$. In de tabel bevinden zich de waarden t van de variabele.

5.7 Samenvatting

In dit hoofdstuk hebben we verschillende eigenschappen van de populatie besproken: de verdelingsfunctie, het gemiddelde, de variantie, de covariantie, etc. De wiskundige formulering van deze eigenschappen hangt af van het type van variabele: discreet of continu. Voor het gemiddelde en de variantie hebben we enkele nuttige stellingen gezien die later van pas zullen komen.

We hebben twee verdelingen in detail besproken: de binomiale verdeling en de normale verdeling. De binomiale verdeling is een voorbeeld van een verdeling van een discrete variabele. De kansverdeling kan berekend worden door gebruik te maken van een wiskundige formule. De normale verdeling is een voorbeeld van een verdeling van een continue variabele. De cumulatieve verdelingsfunctie wordt bekomen door de dichtheidsfunctie te integreren, wat overeenkomt met het berekenen van oppervlaktes. Omdat we deze oppervlaktes niet zelf zullen berekenen, maken we gebruik van een tabel. Door variabelen te standaardiseren hebben we slechts één tabel nodig voor alle normale verdelingen.

De χ^2_k-verdeling en de t_k-verdeling zijn andere voorbeelden van verdelingen van continue variabelen en zullen vooral later in de cursus een rol spelen.

Deel III

Inductieve statistiek

Hoofdstuk 6

De steekproevenverdeling

In Deel I van de syllabus hebben we statistische methodes besproken om data afkomstig uit een steekproef te analyseren. In Deel II hebben we formeel de populatie omschreven. Vaak wenst de onderzoeker een uitspraak te doen over deze populatie. In dit hoofdstuk bestuderen we eigenschappen van variabelen die we bekomen door op willekeurige wijze een steekproef te trekken uit de populatie.

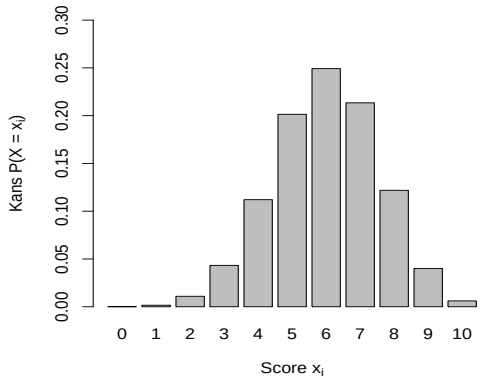
Centraal hierbij staat de *reproduceerbaarheid* van de onderzoeksresultaten: indien we op basis van een steekproef een bepaald besluit formuleren, moeten deze resultaten reproduceerbaar zijn. Hiermee bedoelen we dat we gelijkaardige conclusies verwachten wanneer we het experiment opnieuw uitvoeren op basis van een nieuwe steekproef. De steekproevenverdeling zal ons toelaten hierover besluiten te formuleren, zonder dat we het experiment opnieuw moeten uitvoeren. Dit is een zeer krachtige eigenschap: niettegenstaande reproduceerbaarheid zeer belangrijk is, hebben we vaak in de praktijk geld en tijd om slechts één experiment uit te voeren (op basis van één steekproef). Via de steekproevenverdeling kunnen we dan de reproduceerbaarheid inschatten zonder dat we het experiment opnieuw moeten uitvoeren.

6.1 Steekproeftrekking

Zoals kort besproken in Hoofdstuk 2 is de wijze waarop de steekproef uit de populatie wordt getrokken zeer belangrijk. Wij beperken ons tot de *aselecte steekproeftrekking*: op volledig willekeurige wijze (lukraak) worden n elementen geselecteerd uit de populatie. We veronderstellen verder dat deze n elementen onafhankelijk zijn van elkaar. Zoals voorheen noteren we de variabele met een hoofdletter: X. Omdat we deze va-

riabele kunnen meten voor elk element in de steekproef, zullen we ze ook noteren als $X_1, X_2, \ldots, X_n$. De waarden van de variabelen voor één specifieke steekproef schrijven we met kleine letters (zoals in Deel I van de syllabus): $x_1, x_2, \ldots, x_n$. Bijgevolg stelt X_i de variabele X voor van element i in een steekproef zonder dat we deze steekproef effectief getrokken hebben, terwijl x_i de waarde voorstelt van de variabele X bij element i voor een specifiek getrokken steekproef.

In dit hoofdstuk hernemen we het voorbeeld van de score op de Benton Visual Retention Test uit paragraaf 5.1. Figuur 6.1 geeft de kansverdeling weer van de score (X) voor alle kinderen in de populatie (merk op dat we deze figuur reeds besproken hebben, zie Figuur 5.1).



Figuur 6.1: Staafdiagram van de kansverdeling voor de score op de Benton Visual Retention Test.

Zoals besproken in paragraaf 5.1, drukt $P(X = x_i)$ de relatieve frequentie uit van x_i in de populatie. Bijvoorbeeld, $P(X = 5) = 0.2014$ zodat 20.14% van de kinderen in de populatie een score van 5 hebben. De notatie P komt van *probabiliteit* en dit komt omdat we een *kansinterpretatie* kunnen geven aan $P(X = x_i)$. Alvorens we dit kunnen doen, moeten we eerst weten wat we precies bedoelen met *een kans*.

Intermezzo: de betekenis van een kans

De betekenis van een kans kunnen we eenvoudig illustreren aan de hand van het opwerpen van een geldstuk. Bij een worp kan de uitkomst ofwel 'munt' ofwel 'kop' zijn en deze wordt volledig door het toeval bepaald. Als we een geldstuk eenmaal opwerpen, is de kans dat we munt gooien gelijk aan 0.5 (of 50%). Wat bedoelen we precies met de zin "de kans om munt te werpen is 50%"? Er bestaan verschillende interpretaties van een kans. Diegene die we in deze cursus bespreken is de *frequentistische*:

Als we een geldstuk een *oneindig* aantal keer opwerpen, zullen we in 50% van de

gevallen munt geworpen hebben.

De kans op een gebeurtenis (hier het werpen van munt) is dus gelijk aan de relatieve frequentie van de gebeurtenis indien we het experiment (hier het opwerpen van een geldstuk) een oneindig aantal keer herhalen.

Deze theoretische omschrijving van een kans kunnen we in de praktijk niet toepassen: we kunnen een geldstuk geen oneindig aantal keer opwerpen. We kunnen dit echter wel *benaderen*: als we een geldstuk een groot aantal keer opwerpen, zullen we in *ongeveer* 50% van de gevallen munt gegooid hebben. Hoe meer we dit herhalen, hoe dichter deze relatieve frequentie bij 50% zal liggen.

Deze interpretatie kunnen we koppelen aan de formule van de kans (formule (5.1) op pagina 140). We schrijven de uitkomst van het opwerpen van een geldstuk symbolisch als Y (dus Y kan de waarden 'munt' of 'kop' aannemen). Door formule (5.1) toe te passen, bekomen we:

$$P(Y = munt) = \lim_{n\to\infty} \frac{f_{munt}}{n},$$

waarbij f_{munt} staat voor de absolute frequentie van 'munt' (het aantal keer dat we munt geworpen hebben bij n opwerpingen), zodat $\frac{f_{munt}}{n}$ staat voor de relatieve frequentie. Aangezien n naar oneindig gaat, geeft $P(Y = munt)$ inderdaad de relatieve frequentie weer van het aantal keer dat we 'munt' gooien indien we het geldstuk een oneindig aantal keer opwerpen.

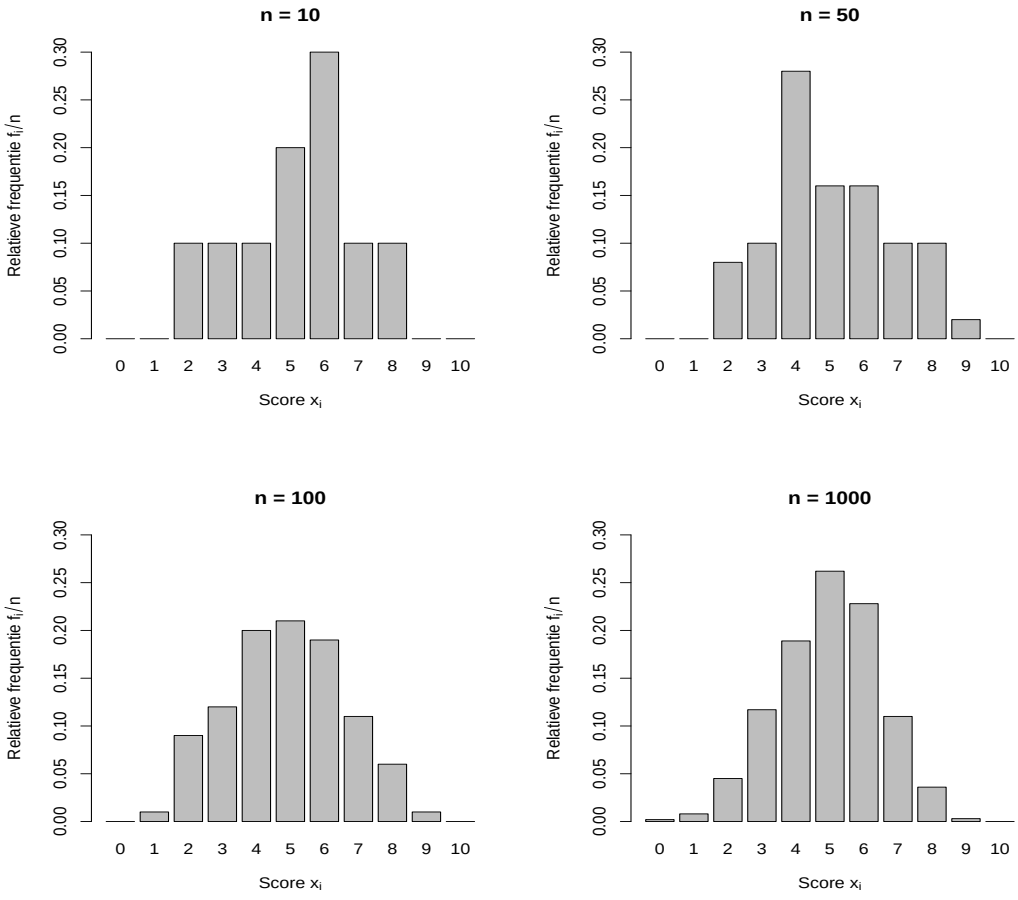
Terugkeer naar de Benton Visual Retention Test

Gelijkaardig aan het voorbeeld met het geldstuk, kunnen we bij de Benton Visual Retention Test de waarde $P(X = x_i)$ ook interpreteren als een kans via de *herhaalde steekproeftrekking*.

We illustreren dit voor $x_i = 5$ waarvoor $P(X = 5) = 0.2014$. Op willekeurige wijze selecteren we een kind uit de populatie, we nemen de Benton Visual Retention Test af en we noteren de score. Vervolgens selecteren we willekeurig een tweede kind uit de populatie, we nemen de test af en we noteren de score. Dit proces herhalen we vele malen (theoretisch gezien een oneindig aantal keer). Van alle scores die we genoteerd hebben, zullen er 20.14% gelijk zijn aan 5. We zeggen dat de kans om een score van 5 te bekomen gelijk is aan 0.2014 of 20.14%.

Als we dit vergelijken met het opwerpen van een geldstuk, komt het trekken van een steekproef van een persoon gevolgd door de afname van de Benton Visual Retention Test (het experiment) overeen met het herhaaldelijk opwerpen van een geldstuk, terwijl de gebeurtenis $X = 5$ overeenkomt met het werpen van 'munt'.

Merk op dat de kansen $P(X = x_i)$ zoals gegeven in Figuur 6.1 enkel opgaan indien we de steekproeftrekking een oneindig aantal keer herhalen. Figuur 6.2 illustreert dit door staafdiagrammen te tonen op basis van een steekproef met n elementen (hier kinderen) voor verschillende keuzes van n. Naarmate n groter wordt, zal het staafdiagram van de relatieve frequentieverdeling beter gelijken op het staafdiagram van de kansverdeling zoals gegeven in Figuur 6.1.



Figuur 6.2: Staafdiagram op basis van een steekproef met n elementen voor de score op de Benton Visual Retention Test.

Een variabele X die bekomen wordt door op toevallige wijze een element uit de populatie te trekken, wordt ook een *toevalsvariabele* genoemd omdat:

- ze het resultaat aanduidt van een toevallige trekking van een element uit de populatie.

- ze veranderlijk (variabel) is omdat niet alle elementen in de populatie dezelfde waarde hebben.

Doorheen de syllabus zullen we echter kortweg blijven spreken over een *variabele* in plaats van een toevalsvariabele.

6.2 Steekproevenverdeling van het gemiddelde

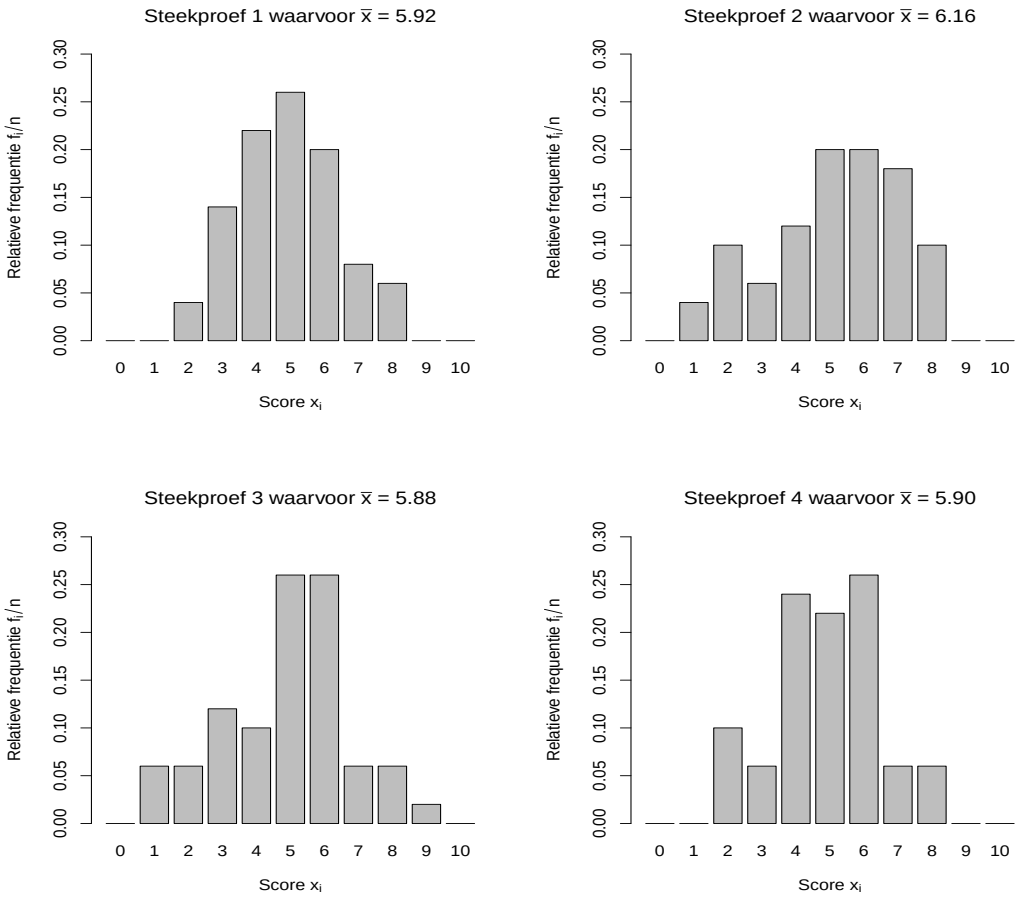
In Hoofdstuk 3 hebben we het steekproefgemiddelde genoteerd als $\bar{x}$. Dit komt overeen met het gemiddelde van de waarden van een variabele gemeten in één steekproef. Figuur 6.3 toont een staafdiagram van de frequentieverdeling voor een steekproef bestaande uit 50 kinderen (steekproef 1) die de Benton Visual Retention Test hebben afgelegd. Voor deze steekproef berekenen we de gemiddelde score, hier $\bar{x} = 5.92$. Vervolgens kunnen we een tweede steeproef trekken (steekproef 2, Figuur 6.3). Voor deze steekproef berekenen we ook het gemiddelde $\bar{x} = 6.16$. Het is niet verwonderlijk dat het gemiddelde van beide steekproeven verschillend is, het zijn immers twee verschillende steekproeven (verschillende groepen van kinderen). Indien we een derde steekproef trekken (steekproef 3) dan is de gemiddelde score $\bar{x} = 5.88$, terwijl dit voor een vierde steekproef (steekproef 4) gelijk is aan $\bar{x} = 5.90$.

Op basis van deze vier steekproeven zien we dat het steekproefgemiddelde *variabel* is: de waarde hangt af van de frequentieverdeling van de scores in de steekproef en verschillende steekproeven hebben verschillende frequentieverdelingen. *Het steekproefgemiddelde is bijgevolg een variabele.*

Omdat we de conventie hebben om variabelen met een hoofdletter te schrijven, zullen we ook het steekproefgemiddelde met een hoofdletter schrijven:

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

Hier stelt $\bar{X}$ het steekproefgemiddelde voor van een steekproef in het algemeen. Eenmaal we een steekproef hebben getrokken en waarden $x_1, x_2, \ldots, x_n$ van de variabele observeren, schrijven we het steekproefgemiddelde met een kleine letter $\bar{x}$. Bijvoorbeeld $\bar{x} = 5.92$ in Figuur 6.3 stelt de waarde voor van het steekproefgemiddelde in steekproef 1. Samengevat duiden we met $\bar{X}$ het steekproefgemiddelde aan voor een steekproef in het algemeen, en duiden we met $\bar{x}$ de waarde aan van het steekproefgemiddelde berekend op basis van één specifieke steekproef.



Figuur 6.3: Staafdiagram van de score op de Benton Visual Retention Test en de gemiddelde score voor 4 verschillende steekproeven met $n = 50$.

Het steekproefgemiddelde is een voorbeeld van een *steekproefgrootheid*: het is een bewerking toegepast op de variabelen $X_1, \ldots, X_n$. Voor het gemiddelde bestaat deze bewerking uit het sommeren en te delen door n. Andere voorbeelden van steekproefgrootheden zijn de mediaan, de modus, de variantie, etc. Een steekproefgrootheid wordt soms ook een *statistiek* genoemd.

Figuur 6.3 geeft de gemiddelden voor 4 steekproeven, maar we kunnen dit blijven herhalen voor meerdere steekproeven. Tabel 6.1 geeft enkele steekproefgemiddelden weer voor 1000 steekproeven. Figuur 6.4 geeft het histogram weer van deze 1000 steekproefgemiddelden (figuur links). Indien we een oneindig aantal steekproeven trekken en we een histogram opstellen van het steekproefgemiddelde met oneindig veel klassen, bekomen we de dichtheidsfunctie van het gemiddelde (Figuur 6.4 rechts). Deze dichtheidsfunctie wordt ook de *steekproevenverdeling* van het gemiddelde genoemd: ze geeft de verdeling (hier in termen van een dichtheidsfunctie) weer van het steekproefgemiddelde voor zeer veel steekproeven (theoretisch gezien oneindig veel steekproeven).

Meer algemeen kunnen we de steekproevenverdeling bekomen voor elke steekproefgrootheid.

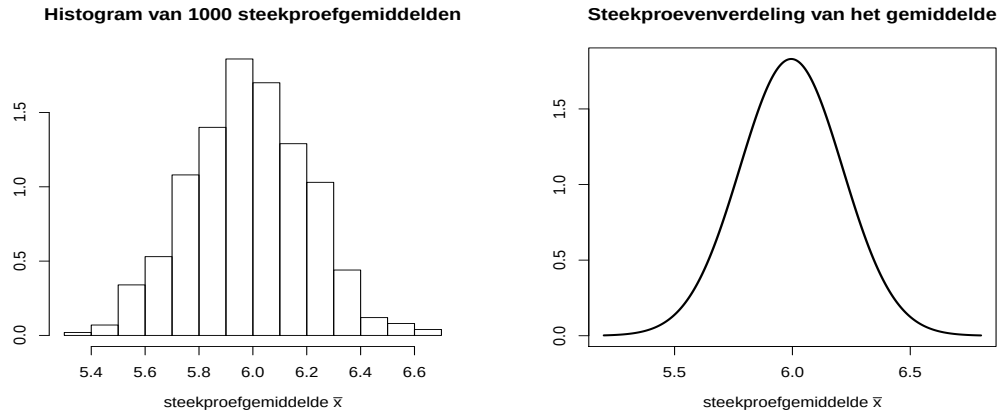
! De verdeling van een steekproefgrootheid wordt de **steekproevenverdeling** van die steekproefgrootheid genoemd.

Opgelet: De frequentieverdeling geeft de verdeling van een variabele weer, terwijl de steekproevenverdeling de verdeling van een steekproefgrootheid weergeeft.

Steekproef	Steekproefgemiddelde $\bar{x}$
1	5.92
2	6.16
3	5.88
4	5.90
5	6.02
$\vdots$	$\vdots$
999	6.00
1000	6.12

Tabel 6.1: De steekproefgemiddelden horende bij 1000 *steekproeven.*

Analoog als bij de binomiale en normale verdeling, kunnen we ook voor de steekproevenverdeling de verwachtingswaarde bepalen. Volgende stelling geeft deze verwachtingswaarde.



Figuur 6.4: Links: histogram van 1000 steekproefgemiddelden met $n = 50$. Rechts: dichtheidsfunctie die we bekomen door oneindig veel steekproeven met $n = 50$ te nemen en het gemiddelde te berekenen. Dit wordt ook de steekproevenverdeling *van het gemiddelde genoemd.*

Stelling 11. *De verwachtingswaarde van het steekproefgemiddelde $\bar{X}$ is gelijk aan het populatiegemiddelde van de variabele X:*

$$E(\bar{X}) = \mu_X.$$

Bewijs. Door gebruik te maken van Stelling 3 volgt dat:

$$E(\bar{X}) = E\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n}E\left(X_1 + \ldots + X_n\right).$$

Stelling 4 impliceert dat:

$$E\left(X_1 + \ldots + X_n\right) = E(X_1) + \ldots + E(X_n).$$

Omdat $E(X_i) = \mu_X$ volgt dat:

$$E(X_1) + \ldots + E(X_n) = \mu_X + \ldots + \mu_X = n\mu_X,$$

zodat:

$$E(\bar{X}) = \frac{1}{n}E\left(X_1 + \ldots + X_n\right) = \frac{n\mu_X}{n} = \mu_X.$$

□

Voor één steekproef is het steekproefgemiddelde over het algemeen *niet* gelijk aan het populatiegemiddelde. Voor de Benton Visual Retention Test is het populatiegemiddelde

gelijk aan $\mu_X = 5.9939$ (zie pagina 155), terwijl de steekproefgemiddelden in Tabel 6.1 verschillend zijn. Het gemiddelde van de 1000 steekproefgemiddelden is gelijk aan 5.9949 en ligt wel zeer dicht bij het populatiegemiddelde. Indien we nu theoretisch gezien een oneindig aantal steekproeven zouden trekken (i.p.v. 1000) en het gemiddelde berekenen, garandeert Stelling 11 dat het gemiddelde van deze steekproefgemiddelden exact gelijk zal zijn aan het populatiegemiddelde.

De volgende stelling geeft de variantie van het steekproefgemiddelde.

Stelling 12. *De variantie van het steekproefgemiddelde is gelijk aan de populatievariantie van de variabele gedeeld door de steekproefgrootte:*

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

Bewijs. Door gebruik te maken van Stelling 7 volgt dat:

$$V(\bar{X}) = V\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n^2}V(X_1 + \ldots + X_n).$$

Omdat $X_1, \ldots, X_n$ onafhankelijk zijn, volgt uit formule (5.15) dat:

$$V(X_1 + \ldots + X_n) = V(X_1) + \ldots + V(X_n).$$

Omdat $V(X_i) = \sigma_X^2$ volgt dat:

$$V(X_1) + \ldots + V(X_n) = \sigma_X^2 + \ldots + \sigma_X^2 = n\sigma_X^2,$$

zodat:

$$V(\bar{X}) = \frac{1}{n^2}V(X_1 + \ldots + X_n) = \frac{n\sigma_X^2}{n^2} = \frac{\sigma_X^2}{n}.$$

□

De variantie van het steekproefgemiddelde is dus *niet* gelijk aan de populatievariantie van de variabele. De variantie van het steekproefgemiddelde zal altijd kleiner dan of gelijk zijn aan de populatievariantie van de variabele omdat $n \geq 1$.

Wat is de betekenis van Stelling 12? Laten we terug het voorbeeld rond de Benton Visual Retention Test nemen. De populatievariantie is $\sigma_X^2 = 2.417$ (zie pagina 157). De variantie van de steekproefgemiddelden in Tabel 6.1 is gelijk aan 0.047 en is veel kleiner dan σ_X^2. Volgens de stelling is de variantie van het steekproefgemiddelde gelijk aan $\sigma_X^2/n = 2.417/50 = 0.048$ wat inderdaad dicht bij 0.047 ligt. Indien we meer dan 1000 steekproeven nemen (theoretisch gezien oneindig veel), zullen deze waarden dichter bij elkaar komen te liggen.

Naarmate we grotere steekproeven nemen om het gemiddelde te berekenen, zal de variatie tussen de steekproefgemiddelden afnemen (de variantie wordt dus kleiner). Tabel 6.2 illustreert dit: ze geeft de gemiddelden voor 'kleine' steekproeven ($n = 10$) en 'grote' steekproeven ($n = 500$). De steekproefgemiddelden verschillen (variëren) meer bij de kleine steekproeven en bijgevolg is hun variantie groter. Bij grotere steekproeven hebben we meer informatie over de populatie (omdat we meer elementen in de steekproef hebben) en zal het steekproefgemiddelde 'dichter' bij het populatiegemiddelde liggen en minder variëren. De variantie is dus kleiner. Dit wordt bevestigd als we de varianties van de steekproefgemiddelden berekenen: voor $n = 10$ is dit 0.223 terwijl dit voor $n = 500$ gelijk is aan 0.005.

Als we Stelling 11 en Stelling 12 combineren, krijgen we de *wet van de grote aantallen* die stelt dat het steekproefgemiddelde met hoge waarschijnlijkheid weinig zal verschillen van het populatiegemiddelde indien de steekproef 'groot' is (i.e. als ze oneindig groot is). Dit kan je als volgt interpreteren: als de steekproef groter en groter wordt, zal het steekproefgemiddelde beter en beter het populatiegemiddelde benaderen.

Steekproef $n = 10$	Steekproefgemiddelde $\bar{x}$	Steekproef $n = 500$	Steekproefgemiddelde $\bar{x}$
1	6.20	1	5.95
2	6.30	2	5.92
3	6.00	3	5.98
4	5.80	4	5.93
5	6.10	5	5.98
⋮	⋮	⋮	
999	6.00	999	5.94
1000	6.40	1000	5.92

Tabel 6.2: De steekproefgemiddelden horende bij 1000 *steekproeven. Links: steekproeven met* 10 *elementen. Rechts: steekproeven met* 500 *elementen.*

Nu we de verwachtingswaarde en de variantie van het steekproefgemiddelde weten, wensen we ook de verdelingsfunctie te weten te komen. De volgende twee stellingen zullen ons deze informatie geven.

Stelling 13. *Stel dat $X_1, \ldots, X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met een normale verdeling $N(\mu_X, \sigma_X^2)$, dan zal $\bar{X}$ ook normaal verdeeld zijn:*

$$\bar{X} \sim N(\mu_X, \sigma_X^2/n). \tag{6.1}$$

Als de populatie een normale verdeling heeft, dan volgt automatisch dat het steekproefgemiddelde ook een normale verdeling heeft. Dit geldt voor elke keuze van de

steekproefgrootte n. Merk op dat we in Stelling 13 gebruik hebben gemaakt van Stellingen 11 en 12 door $E(\bar{X})$ en $V(\bar{X})$ onmiddellijk in te vullen in formule (6.1).

Het histogram en de dichtheidsfunctie in Figuur 6.4 visualiseren de verdelingsfunctie van het steekproefgemiddelde en deze heeft inderdaad de vorm van een normale verdeling. Stelling 13 stelt dat dit een normale verdeling moet zijn indien X (de score op de Benton Visual Retention Test) uit een normale verdeling komt. De verdeling van X wordt weergegeven in Figuur 6.1. Hoewel deze op een normale verdeling lijkt, is ze het niet. Dit komt omdat X een discrete variabele is en de normale verdeling enkel opgaat voor continue variabelen.

Niettegenstaande Stelling 13 niet opgaat voor de score op de Benton Visual Retention Test, lijkt de verdelingsfunctie van het steekproefgemiddelde (Figuur 6.4) toch sterk op die van een normale verdeling. Dit kan verklaard worden door een andere zeer belangrijke stelling: *de centrale limietstelling*.

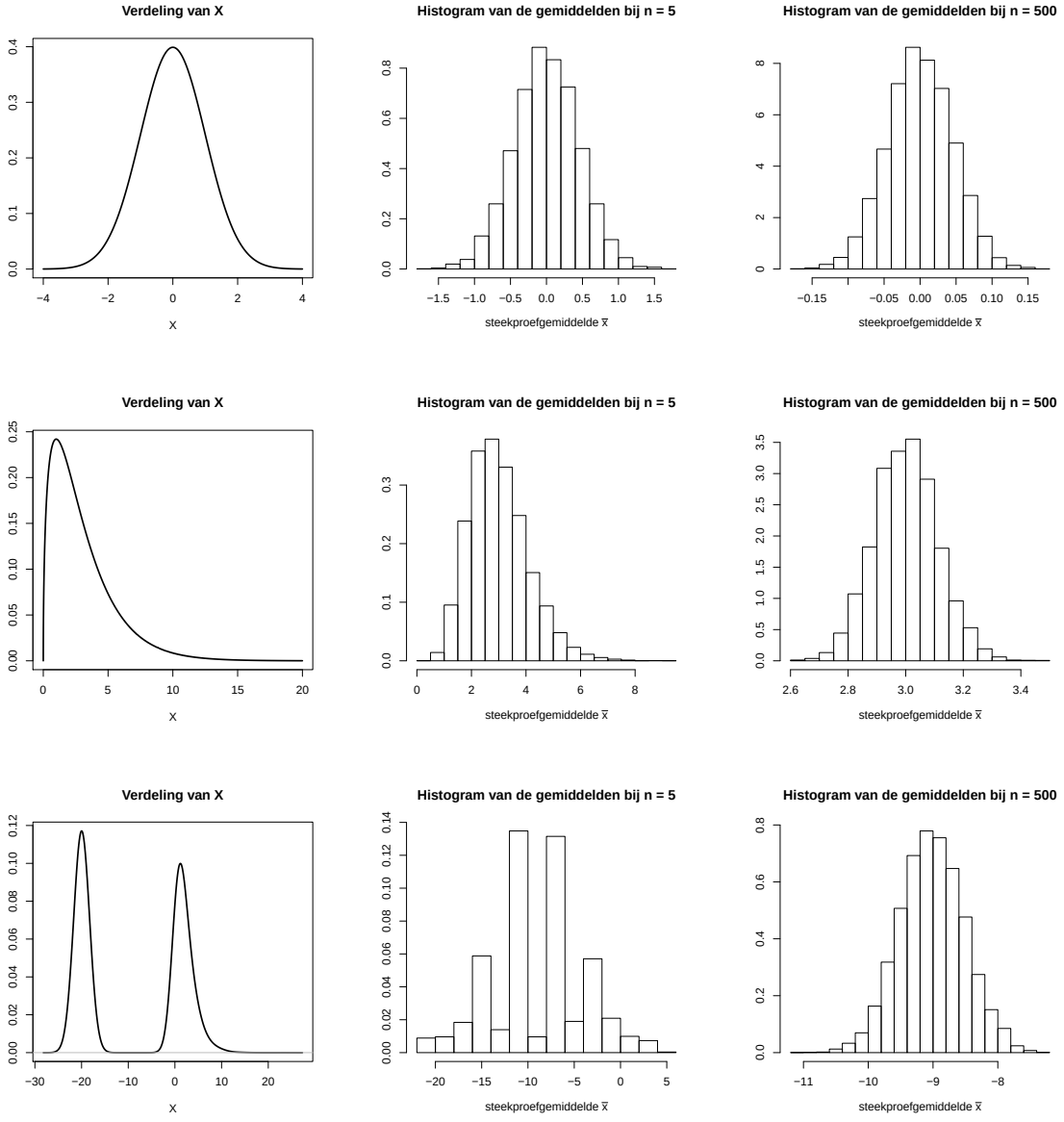
Stelling 14 (Centrale limietstelling)**.** *Stel dat $X_1, \ldots, X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met gemiddelde μ_X en variantie σ_X^2, dan wordt de verdeling van het steekproefgemiddelde $\bar{X}$ naarmate n groter wordt, steeds beter benaderd door de normale verdeling met gemiddelde μ_X en variantie σ_X^2/n.*

Stelling 14 stelt dat het steekproefgemiddelde, bij benadering, *altijd* normaal verdeeld zal zijn, zolang *de steekproef maar groot genoeg is*. Hoe groter de steekproef, hoe beter de verdeling van het steekproefgemiddelde zal lijken op een normaalverdeling. Stelling 14 is dus veel breder toepasbaar dan Stelling 13 omdat ze opgaat voor elke verdeling van X (niet noodzakelijk de normaalverdeling). Anderzijds gaat stelling 13 op voor elke keuze van n, terwijl Stelling 14 maar opgaat voor 'grote' n. In de praktijk gebruiken we vaak de regel $n \geq 30$ om aan te geven dat een steekproef 'groot' is. Merk op dat dit slechts een vuistregel is.

Figuur 6.5 toont de verdeling van het steekproefgemiddelde voor verschillende keuzes van de verdeling van X. De histogrammen in de bovenste rij stellen steekproefgemiddelden voor van steekproeven afkomstig uit een normale verdeling. Voor zowel $n = 5$ als $n = 500$ lijkt het histogram op dat van een normale verdeling. Dit is een gevolg van Stelling 13.

Bij de middelste rij figuren zijn de steekproeven afkomstig uit een χ_3^2-verdeling. Voor $n = 5$ is de verdeling van het steekproefgemiddelde scheef naar rechts en dus verschillend van een normale verdeling. Dit komt omdat we Stelling 13 niet kunnen toepassen. Voor $n = 500$ wordt de verdeling symmetrisch en kan ze benaderd worden door de normale verdeling. Dit is een gevolg van Stelling 14.

Voor de onderste rij figuren zijn de steekproeven afkomstig uit een verdeling met twee pieken. Voor $n = 5$ is het histogram sterk verschillend van de normale verdeling. Dit komt doordat we Stelling 13 niet kunnen toepassen. Voor $n = 500$ is de verdeling bij benadering gelijk is aan die van de normale verdeling. Dit is een gevolg van Stelling 14.



Figuur 6.5: Steekproefverdeling van het gemiddelde voor een variabele met een normale verdeling (boven), een χ^2_3-verdeling (midden) en een verdeling met twee pieken (onder). De histogrammen zijn gebaseerd op 10000 steekproeven.

In paragraaf 5.6.2 hebben we besproken hoe we normaal verdeelde variabelen kunnen

standaardiseren. We kunnen dit ook toepassen op het steekproefgemiddelde:

$$P(\bar{X} \leq x) = P\left(Z \leq \frac{x - \mu_X}{\sqrt{\sigma_X^2/n}}\right), \quad Z \sim N(0,1). \tag{6.2}$$

Indien X uit een normale verdeling komt, geldt eigenschap (6.2) voor alle keuzes van n. Indien X niet uit een normale verdeling komt, geldt deze eigenschap enkel maar voor grote n.

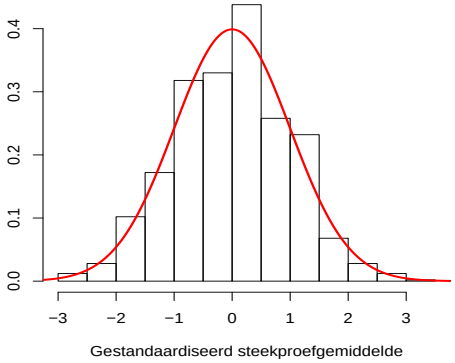
Eigenschap (6.2) kunnen we visualiseren op basis van de scores op de Benton Visual Retention Test. Zoals eerder aangegeven impliceert de centrale limietstelling dat 1000 steekproefgemiddelden in Tabel 6.3 bij benadering normaal verdeeld zijn (omdat $n = 50$ 'groot' is). Als we deze nu standaardiseren, stelt eigenschap (6.2) dat deze gestandaardiseerde gemiddelden bij benadering standaardnormaal verdeeld zijn.

Tabel 6.3 illustreert het standaardiseren (herinner je dat voor de scores op de Benton Visual Retention test geldt dat $n = 50$, $\mu_X = 5.9939$ en $\sigma_X^2 = 2.417$). Ter illustratie: het gemiddelde van de eerste steekproef is 5.92 zodat het gestandaardiseerde gemiddelde gelijk is aan $(5.92 - 5.9939)/(\sqrt{2.417/50}) = -0.34$. Figuur 6.6 toont het histogram van deze gestandaardiseerde gemiddelden samen met de dichtheidsfunctie van de standaardnormale verdeling. De dichtheidsfunctie is een goede benadering van het histogram.

Steekproef	Steekproefgemiddelde $\bar{x}$	Gestandaardiseerd $\frac{\bar{x}-\mu_X}{\sqrt{\sigma_X^2/n}}$
1	5.92	−0.34
2	6.16	0.76
3	5.88	−0.52
4	5.90	−0.43
5	6.02	0.12
⋮	⋮	
999	6.00	0.03
1000	6.12	0.57

Tabel 6.3: De steekproefgemiddelden horende bij 1000 *steekproeven en de gestandaardiseerde gemiddelden.*

Stellingen 13 en 14 zijn samen met eigenschap (6.2) van fundamenteel belang voor de statistiek. Ze laten toe kansen te berekenen die weergeven wat er zou gebeuren indien we een experiment blijven herhalen. In de praktijk zullen we vaak slechts één experiment uitvoeren, maar om anderen te overtuigen van de onderzoeksresultaten moeten de conclusies *reproduceerbaar* zijn: indien we het experiment opnieuw uitvoeren (op basis van een nieuwe steekproef), moeten we gelijkaardige resultaten bekomen.



Figuur 6.6: Histogram van 1000 gestandaardiseerde steekproefgemiddelden $\frac{\bar{x}-\mu_X}{\sqrt{\sigma_X^2/n}}$ *met* $n = 50$*. De rode volle lijn stelt de dichtheidsfunctie voor van de standaardnormale verdeling.*

We nemen terug het voorbeeld van de Benton Visual Retention Test en Tabel 6.3 om dit te illustreren. Het gemiddelde in steekproef 1 werd bekomen door het experiment één keer uit te voeren: de onderzoeker heeft 50 kinderen geselecteerd uit de populatie en bij elk kind de test afgenomen. Het gemiddelde van de scores op deze 50 testen is gelijk aan $\bar{x} = 5.92$.

De onderzoeker wenst te weten of dit een ongewoon lage of ongewoon hoge score is: indien het experiment vele malen opnieuw wordt uitgevoerd, zal de gemiddelde score dan sterk verschillen van 5.92? Eén manier om hierop een antwoord te geven is door de kans $P(\bar{X} \leq 5.92)$ te berekenen. Indien deze kans zeer klein is (rond 0), weten we dat de gemiddelde score 5.92 zeer laag is (relatief gezien t.o.v. de scores in de populatie). Als de kans anderzijds zeer groot is (rond 1), impliceert dit dat de gemiddelde score 5.92 zeer groot is. Er zijn twee mogelijkheden om de kans $P(\bar{X} \leq 5.92)$ te berekenen:

A. De onderzoeker zal het experiment vele malen herhalen, bijvoorbeeld 1000 keer zoals weergeven in Tabel 6.3. Vervolgens berekenen we de proportie van gemiddelden dat kleiner dan of gelijk is aan 5.92. Het histogram in Figuur 6.7 geeft dit grafisch weer en de oppervlakte links van de rode volle lijn is een benadering van de kans $P(\bar{X} \leq 5.92)$ (om de exacte kans te weten te komen, moeten we het experiment een oneindig aantal keer herhalen). Hier is deze oppervlakte 0.374 en bijgevolg is de kans om een gemiddelde score te bekomen van maximaal 5.92 ongeveer gelijk aan 37.4%. Dit geeft aan dat 5.92 geen ongewone gemiddelde score is.

B. De onderzoeker voert het experiment maar 1 keer uit en maakt gebruik van Stel-

ling 14 en eigenschap 6.2. We standaardiseren eerst het gemiddelde:

$$P(\bar{X} \leq 5.92) = P\left(Z \leq \frac{5.92 - \mu_X}{\sqrt{\sigma_X^2/n}}\right) = P\left(Z \leq \frac{5.92 - 5.9939}{\sqrt{2.417/50}}\right) = P(Z \leq -0.34),$$

en vervolgens lezen we de kans af $P(Z \leq -0.34)$ met $Z \sim N(0,1)$ uit Tabel 5.7, waarbij we gebruik maken van formule (5.21):

$$P(Z \leq -0.34) = 1 - P(Z \leq 0.34) = 1 - 0.6331 = 0.3669.$$

Bijgevolg is de kans om een gemiddelde score te bekomen van maximaal 5.92 ongeveer gelijk aan 36.7%. Merk op dat deze kans ook slechts een benadering is omdat we beroep doen op de centrale limietstelling. Ze ligt echter dicht in de buurt van de kans bekomen door het experiment 1000 keer te herhalen.

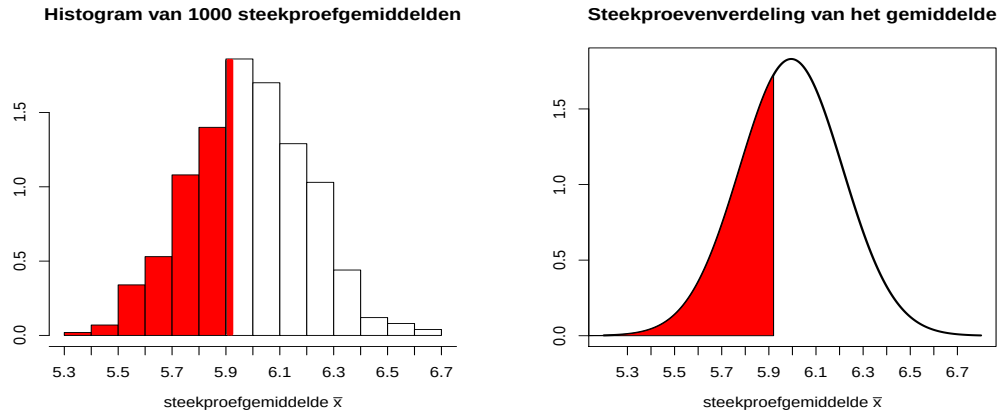
Aanpak A waarbij we het experiment 1000 keer herhalen is conceptueel eenvoudig, maar in de praktijk niet uitvoerbaar: vaak is er maar geld en tijd om het experiment *één keer* uit te voeren. Aanpak B is wiskundig meer complex en doet beroep op verschillende resultaten uit de kansrekening, maar kan uitgevoerd worden op basis van één experiment en is dus praktisch gezien meer bruikbaar dan de eerste mogelijkheid. Aanpak B laat ons toe een uitspraak te doen over wat er zou gebeuren indien we het experiment een oneindig aantal keer zouden herhalen, *zonder dat we dit effectief moeten uitvoeren.* Dit is een zeer krachtige eigenschap.

Aanpak B heeft echter nog een nadeel: om de kans te kunnen berekenen, moeten we μ_X en σ_X^2 invullen, terwijl deze populatieparameters voor vele studies typisch ongekend zijn. In Hoofdstuk 7 zullen we methodes zien die deze tekortkoming omzeilen.

6.3 Steekproevenverdeling van de variantie

De steekproefvariantie (zie paragraaf 3.2.3) is een ander voorbeeld van een steekproefgrootheid. Voor deze grootheid kunnen we ook de steekproevenverdeling bepalen. Analoog als het gemiddelde, zullen we hoofdletters gebruiken om te benadrukken dat de steekproefvariantie een variabele is: als we voor verschillende steekproeven de variantie berekenen, zal deze variëren. We gebruiken opnieuw twee formules voor de steekproefvariantie:

$$SN_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$



Figuur 6.7: Links: histogram van 1000 steekproefgemiddelden met $n = 50$. Rechts: dichtheidsfunctie die we theoretisch bekomen door oneindig veel steekproeven met $n = 50$ te nemen en het gemiddelde te berekenen. De gekleurde oppervlakte komt bij benadering overeen met de kans $P(\bar{X} \leq 5.92)$.

en

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

Voor deze steekproefgrootheden kunnen we ook de verwachtingswaarde bepalen. Er geldt dat:

$$E(SN_X^2) = \frac{n-1}{n}\sigma_X^2.$$

De verwachtingswaarde van de steekproefvariantie SN_X^2 is dus *niet* gelijk aan de populatievariantie. Voor S_X^2 is dit echter wel zo:

$$E(S_X^2) = \sigma_X^2.$$

De verwachtingswaarde van de steekproefvariantie S_X^2 is gelijk aan de populatievariantie. Als we zeer veel (theoretisch gezien oneindig veel) steekproeven trekken en telkens de steekproefvariantie berekenen via formule S_X^2, dan zal het gemiddelde van de varianties gelijk zijn aan de populatievariantie. Dit is een gunstige eigenschap en daarom zal men in de praktijk vaak de variantie berekenen via S_X^2 in plaats van SN_X^2.

De volgende stelling geeft de verdeling van de steekproefgrootheid $(n-1)S_X^2/\sigma_X^2$ wanneer de populatie waaruit de steekproef wordt getrokken een normale verdeling heeft. Deze verdeling heeft vooral een theoretisch nut voor het vervolg van de cursus.

Stelling 15. *Stel dat $X_1, \ldots, X_n$ n onafhankelijke lukrake trekkingen zijn uit een populatie met normale verdeling $N(\mu_X, \sigma_X^2)$, dan geldt:*

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2.$$

6.4 Samenvatting

In dit hoofdstuk hebben we de steekproevenverdeling van het gemiddelde uitgewerkt. Deze steekproevenverdeling laat ons toe uitspraken te formuleren over het gemiddelde indien we het experiment (en dus ook de steekproeftrekking) een oneindig aantal keer zouden herhalen. Dit is een belangrijke eigenschap en hangt nauw samen met de reproduceerbaarheid van de conclusies bekomen op basis van één steekproef.

Tot slot hebben we kort de steekproevenverdeling van de variantie uitgewerkt. Deze heeft vooral een theoretisch nut voor het vervolg van de cursus.

Hoofdstuk 7

Betrouwbaarheidsintervallen en statistische toetsen voor het populatiegemiddelde

In dit hoofdstuk zullen we beroep doen op de kennis van de voorgaande hoofdstukken om op basis van een steekproef een uitspraak te formuleren over de populatie. Een dergelijke uitspraak heeft typisch betrekking op één of meerdere populatieparameters. In dit hoofdstuk staat het populatiegemiddelde centraal. Andere populatieparameters zullen behandeld worden in vervolgcursussen.

Eerst zullen we de populatieparameter *schatten* op basis van de data in de steekproef. Vervolgens kunnen we op basis van deze schatting een uitspraak doen over de populatieparameter door een betrouwbaarheidsinterval op te stellen en/of door gebruik te maken van een statistische toets. Zoals aangegeven in paragraaf 1.2 zullen we nooit 100% zeker zijn of onze uitspraak over de populatieparameter correct is, maar we kunnen deze onzekerheid wel meten, controleren en communiceren.

7.1 Schatters

We hernemen het voorbeeld van de masterstudent die onderzoek doet naar het gemiddeld IQ van studenten eerste bachelor Psychologie en eerste bachelor Pedagogische Wetenschappen; zie paragraaf 1.2.

Het gemiddeld IQ van alle studenten is de populatieparameter waarover men een uit-

spraak wenst te doen. Zoals eerder aangegeven, kunnen we geen IQ test afnemen bij de volledige populatie, maar enkel bij een steekproef van n studenten. In het voorbeeld in paragraaf 1.2 was $n = 50$, maar in dit hoofdstuk zullen we ook andere keuzes van n toelaten. Voor de eenvoud zullen we ook veronderstellen dat alle steekproeven via aselecte steekproeftrekking bekomen zijn: de elementen in de steekproef zijn lukraak geselecteerd uit de populatie en zijn onafhankelijk van elkaar.

Indien we een uitspraak wensen te doen over het populatiegemiddelde op basis van een steekproef, zullen we dit gemiddelde moeten *schatten* op basis van de informatie in de steekproef. Intuïtief lijkt het steekproefgemiddelde een logische keuze om het populatiegemiddelde te schatten. We moeten dit echter meer formeel onderbouwen.

Een schatter voor een populatieparameter θ (uitspraak *thèta*) noteren we als $\hat{\theta}$, waarbij $\hat{\theta}$ een steekproefgrootheid is. We wensen uiteraard goede schatters voor de populatieparameter te bekomen. Hiervoor moeten we eerst formeel omschrijven wat we precies bedoelen met 'goed'.

! $\hat{\theta}$ is een **goede schatter** van θ indien:

- ze *zuiver* is, wat wil zeggen dat de verwachtingswaarde van de schatter gelijk is aan de populatieparameter:

  $$E(\hat{\theta}) = \theta.$$

  Dit houdt in dat de populatieparameters niet systematisch te klein of te groot wordt geschat.
- de variantie van de schatter, $V(\hat{\theta})$, kleiner wordt naarmate de steekproefgrootte toeneemt. Dit drukt uit dat de schatter *meer nauwkeurig* zal zijn wanneer de steekproef groter wordt.

De standaarddeviatie van de schatter, $\sqrt{V(\hat{\theta})}$, wordt ook de *standaardfout* genoemd. Als we verschillende schatters hebben voor een bepaalde populatieparameter, dan zeggen we dat de schatter met de kleinste standaardfout het *efficiëntst* is.

7.1.1 Het gemiddelde

Zoals eerder aangegeven lijkt het steekproefgemiddelde een logische keuze om het populatiegemiddelde te schatten, dus $\hat{\theta} = \bar{X}$ als $\theta = \mu$ (voor de eenvoud laten we het subscript X weg bij de populatieparameters en schrijven we μ i.p.v. μ_X). Uit Stelling

11 volgt dat het steekproefgemiddelde een zuivere schatter is voor het populatiegemiddelde omdat:

$$E(\bar{X}) = \mu.$$

Herinner je dat we deze gelijkheid als volgt kunnen interpreteren: indien we één steekproef trekken uit de populatie zal het steekproefgemiddelde zeer waarschijnlijk niet gelijk zijn aan het populatiegemiddelde (dus $\bar{x} \neq \mu$). Indien we vele steekproeven trekken (theoretisch gezien oneindig veel) en telkens per steekproef het steekproefgemiddelde berekenen, dan zal het gemiddelde van deze steekproefgemiddelden gelijk zijn aan het populatiegemiddelde.

Uit Stelling 12 volgt dat de variantie van het steekproefgemiddelde gelijk is aan:

$$V(\bar{X}) = \frac{\sigma^2}{n}.$$

Als n toeneemt, dan wordt σ^2/n kleiner: hoe groter de steekproef, hoe nauwkeuriger we het populatiegemiddelde kunnen schatten via het steekproefgemiddelde. Uit deze uitdrukking leiden we ook af dat de standaardfout van het steekproefgemiddelde gelijk is aan $\sigma/\sqrt{n}$.

We kunnen dus besluiten dat het steekproefgemiddelde een goede schatter van het populatiegemiddelde is, omdat aan beide voorwaarden van een goede schatter voldaan is.

De waarde van een schatter op basis van één steekproef noemen we een *schatting*. Toegepast op het gemiddelde is $\bar{X}$ de schatter en $\bar{x}$ de schatting die we bekomen op basis van één steekproef.

7.1.2 De variantie

Om de populatievariantie ($\theta = \sigma^2$) te schatten lijkt het logisch om beroep te doen op de steekproefvariantie. We hebben echter twee formules voor de steekproefvariantie: SN_X^2 en S_X^2, zie paragraaf 6.3. Welke van deze twee schatters is nu de beste?

Uit paragraaf 6.3 weten we dat:

$$E(SN_X^2) = \frac{n-1}{n}\sigma^2.$$

Omdat

$$\frac{n-1}{n}\sigma^2 \neq \sigma^2,$$

volgt dat SN_X^2 *geen* zuivere schatter is van de populatievariantie. Als we oneindig veel steekproeven nemen en telkens de steekproefvariantie via SN_X^2 berekenen, dan zal het gemiddelde van deze varianties *niet* gelijk zijn aan de populatievariantie. Omdat

$$\frac{n-1}{n}\sigma^2 < \sigma^2,$$

zal de populatieparameter systematisch te klein geschat worden. Bijgevolg is SN_X^2 geen goede schatter van de populatievariantie.

De populatievariantie kan echter wel zuiver worden geschat door:

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

omdat

$$E(S_X^2) = \sigma^2.$$

Door te delen door $n-1$ in plaats van door n bekomen we een zuivere schatter. Indien we een oneindig aantal steekproeven nemen en telkens de steekproefvariantie berekenen via S_X^2, dan zal het gemiddelde van deze varianties *wel* gelijk zijn aan de populatievariantie. We geven daarom in de praktijk de voorkeur aan de schatter S_X^2 om de populatievariantie te schatten. Men kan verder ook aantonen dat de variantie van zowel SN_X^2 als S_X^2 afneemt naarmate de steekproef groter wordt.

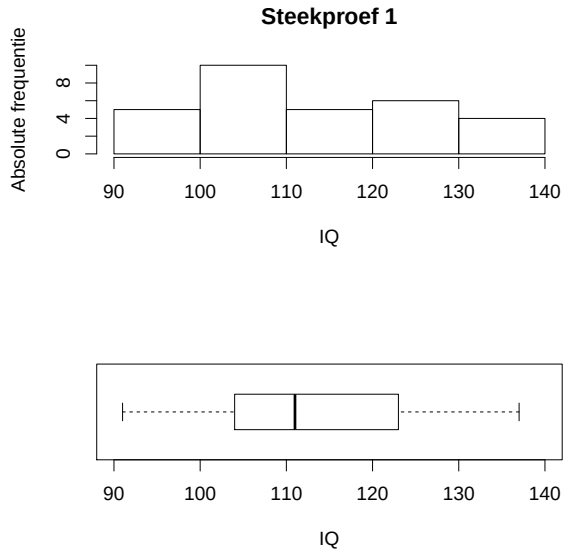
7.2 Betrouwbaarheidsintervallen

In het onderzoek naar het IQ wensen we een uitspraak te doen over het populatiegemiddelde. We weten nu dat het steekproefgemiddelde een goede schatter is voor het populatiegemiddelde. Figuur 7.1 toont het histogram en boxplot van een steekproef van 50 studenten. Voor deze 50 studenten is het gemiddelde $\bar{x} = 112.7$. Het is zeer onwaarschijnlijk dat dit steekproefgemiddelde exact gelijk is aan het populatiegemiddelde (het populatiegemiddelde is het gemiddelde van alle studenten en niet enkel van de 50 in de steekproef). Hoe kunnen we nu een uitspraak doen over het populatiegemiddelde? De steekproevenverdeling zal hierin een cruciale rol spelen en zal ons toelaten *betrouwbaarheidsintervallen* te construeren. Een betrouwbaarheidsinterval zal ons in staat stellen om met een bepaalde zekerheid een uitspraak te doen over het populatiegemiddelde.

We zullen eerst veronderstellen dat de populatievariantie gekend is en dat het IQ van alle studenten normaal verdeeld is. Dit maakt de wiskundige afleidingen wat eenvoudiger. In de praktijk is dit echter onrealistisch omdat de populatievariantie vaak niet

gekend is. We zullen daarom in het tweede deel van deze paragraaf ook betrouwbaarheidsintervallen construeren wanneer de populatievariantie ongekend is.

Merk op dat het populatiegemiddelde uiteraard ongekend is: indien ze gekend zou zijn, hoeven we ze natuurlijk niet te schatten en is er geen nood aan statistische methodes.



Figuur 7.1: Histogram en boxplot van één steekproef met $n = 50$ studenten. Het steekproefgemiddelde is $\bar{x} = 112.7$.

7.2.1 X normaal verdeeld en gekende populatievariantie

Om een betrouwbaarheidsinterval te construeren, hebben we nog wat extra notatie nodig. We duiden met z_α de waarde van de standaardnormale verdeling aan zodat de oppervlakte onder de curve *rechts* van de waarde gelijk is aan α (uitspraak *alfa*). De waarde z_α is dus het getal waarvoor geldt dat:

$$P(Z > z_\alpha) = \alpha,$$

met $Z \sim N(0, 1)$, zie Figuur 7.2. Ter illustratie: als $\alpha = 0.025$ dan is $z_{0.025}$ de waarde zodat:

$$P(Z > z_{0.025}) = 0.025. \tag{7.1}$$

Tabel 5.7 kunnen we echter niet onmiddellijk gebruiken om $z_{0.025}$ af te lezen, want deze tabel bevat de oppervlaktes *links* onder de curve (dus kansen van de vorm $P(Z \leq z)$).

Door gebruik te maken van formule (5.9) kunnen we uitdrukking (7.1) echter herschrijven als:

$$0.025 = 1 - P(Z \leq z_{0.025}),$$

wat gelijk is aan:

$$P(Z \leq z_{0.025}) = 1 - 0.025 = 0.975.$$

We kunnen nu vervolgens Tabel 5.7 gebruiken om $z_{0.025}$ af te lezen. We zien dat:

$$P(Z \leq 1.96) = 0.975,$$

zodat $z_{0.025} = 1.96$. Merk op dat we uit Tabel 5.7 niet alle waarden kunnen aflezen. Bijvoorbeeld $z_{0.05}$ (dus de waarde zodat $P(Z \leq z_{0.05}) = 0.95$) kan niet direct afgelezen worden. Omdat $P(Z \leq 1.64) = 0.9495$ en $P(Z \leq 1.65) = 0.9505$, weten we dat $z_{0.05}$ tussen de waarden 1.64 en 1.65 zal liggen. Voor de eenvoud zullen wij het gemiddelde nemen van deze twee waarden, dus $z_{0.05} = (1.64 + 1.65)/2 = 1.645$. Anderzijds kan je dit ook afleiden uit Tabel 5.9 op pagina 185 met $k = \infty$ (herinner je dat de t-verdeling met oneindig veel vrijheidsgraden gelijk is aan de standaardnormale verdeling).

Omdat de standaardnormale verdeling symmetrisch is rond 0, kan men aantonen dat:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha. \tag{7.2}$$

Figuur 7.3 illustreert dit voor $\alpha = 0.05$, waarvoor $z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$. De oppervlakte onder de kromme tussen de grenzen -1.96 ($= -z_{\alpha/2}$) en 1.96 ($= z_{\alpha/2}$) is gelijk aan $1 - \alpha = 1 - 0.05 = 0.95$. Herinner je dat deze waarde een kans is. We kunnen dit ook als volgt interpreteren:

De kans dat een standaardnormale variabele een waarde aanneemt tussen -1.96 en 1.96 is 95%.

Formule (7.2) is geldig voor elke standaardnormaal verdeelde variabele Z en zal de basis vormen om een betrouwbaarheidsinterval te construeren. Uit paragraaf 6.2 weten we dat het steekproefgemiddelde een normale verdeling volgt indien X een normale verdeling heeft:

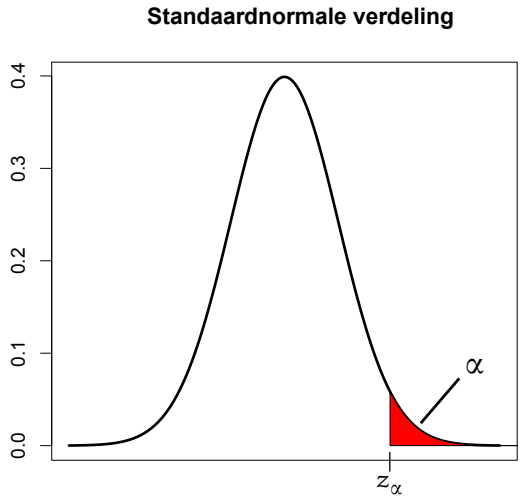
$$\text{als} \quad X \sim N(\mu, \sigma^2) \quad \text{dan} \quad \bar{X} \sim N(\mu, \sigma^2/n).$$

Indien we het steekproefgemiddelde standaardiseren (zie paragraaf 6.2) volgt dat:

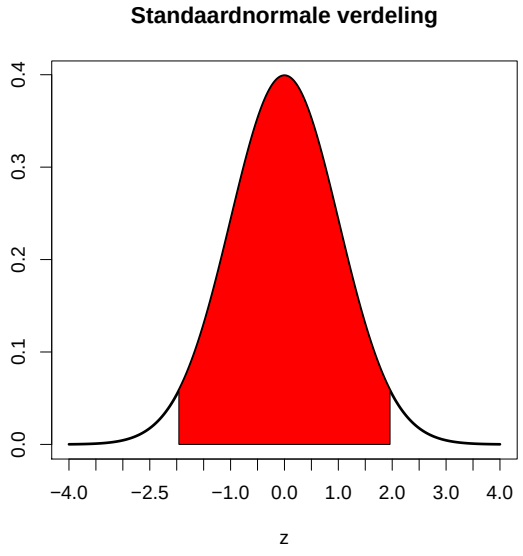
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

We kunnen bijgevolg Z in formule (7.2) vervangen door $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$:

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha. \tag{7.3}$$



Figuur 7.2: z_α is de waarde zodat de gekleurde oppervlakte rechts gelijk is aan α.



Figuur 7.3: De gekleurde oppervlakte komt overeen met de kans $P(-1.96 \leq Z \leq 1.96)$, $Z \sim N(0,1)$, en is gelijk aan 0.95.

Uit uitdrukking (7.3) kunnen we een betrouwbaarheidsinterval afleiden. Om dit te bekomen, moeten we deze uitdrukking eerst herschrijven zodat enkel μ overblijft in de middelste term van de ongelijkheden. We kunnen dit doen in vier stappen.

Stap 1. In een eerste stap vermenigvuldigen we de drie termen van de ongelijkheid in uitdrukking (7.3) met $\sigma/\sqrt{n}$. Omdat $\sigma/\sqrt{n}$ nooit negatief kan zijn, zullen de ongelijkheidstekens niet wijzigen. We bekomen:

$$P(-z_{\alpha/2}\times\sigma/\sqrt{n} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\times\sigma/\sqrt{n} \leq z_{\alpha/2}\times\sigma/\sqrt{n}) = 1-\alpha.$$

Omdat $\sigma/\sqrt{n}$ bij de middelste term zowel in de teller als in de noemer voorkomt, kunnen we ze schrappen:

$$P(-z_{\alpha/2}\sigma/\sqrt{n} \leq \bar{X}-\mu \leq z_{\alpha/2}\sigma/\sqrt{n}) = 1-\alpha.$$

Stap 2. In een volgende stap willen we $\bar{X}$ wegwerken uit de middelste term van de ongelijkheid. We kunnen dit bekomen door bij iedere term $\bar{X}$ af te trekken. Het aftrekken heeft geen invloed op de ongelijkheidstekens. We bekomen:

$$P(-z_{\alpha/2}\sigma/\sqrt{n}-\bar{X} \leq \bar{X}-\mu-\bar{X} \leq z_{\alpha/2}\sigma/\sqrt{n}-\bar{X}) = 1-\alpha.$$

Dit kunnen we herschikken:

$$P(-\bar{X}-z_{\alpha/2}\sigma/\sqrt{n} \leq -\mu \leq -\bar{X}+z_{\alpha/2}\sigma/\sqrt{n}) = 1-\alpha.$$

Stap 3. In een derde stap willen we μ bekomen in de middelste term in plaats van $-\mu$. Dit kunnen we bekomen door de drie termen met -1 te vermenigvuldigen. Merk op dat het vermenigvuldigen met een negatief getal de ongelijkheidstekens zal wijzigen. We bekomen:

$$P(\bar{X}+z_{\alpha/2}\sigma/\sqrt{n} \geq \mu \geq \bar{X}-z_{\alpha/2}\sigma/\sqrt{n}) = 1-\alpha.$$

Stap 4. In een laatste stap herschrijven we de ongelijkheid zodat de kleinste waarde links staat en de grootste rechts:

$$P(\bar{X}-z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X}+z_{\alpha/2}\sigma/\sqrt{n}) = 1-\alpha. \tag{7.4}$$

Formule (7.4) is zeer belangrijk en we kunnen ze als volgt interpreteren: de kans dat het populatiegemiddelde in het interval

$$[\bar{X}-z_{\alpha/2}\sigma/\sqrt{n}, \bar{X}+z_{\alpha/2}\sigma/\sqrt{n}] \tag{7.5}$$

ligt, is gelijk aan $1-\alpha$. Interval (7.5) wordt het $(1-\alpha)100\%$ *betrouwbaarheidsinterval (BI) genoemd.*

Kiezen we bijvoorbeeld $\alpha = 0.05$ dan is $z_{\alpha/2} = 1.96$ en wordt dit interval:

$$[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}],$$

en de kans dat μ in dit interval ligt, is gelijk aan 95% ($1-\alpha = 1-0.05 = 0.95$). We zijn dus in staat om een uitspraak te doen over het populatiegemiddelde op basis van een steekproef. *We zijn echter niet* 100% *zeker van onze conclusie, maar* 95% *zeker.* Er bestaat dus een kans dat onze conclusie foutief is.

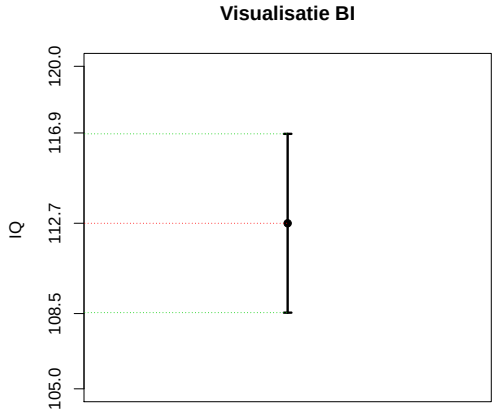
Op basis van de steekproef uit Figuur 7.1 volgt dat $\bar{x} = 112.7$, $n = 50$ en we veronderstellen dat $\sigma = 15$. Vervolgens berekenen we de grenzen van het betrouwbaarheidsinterval (wanneer we $\alpha = 0.05$ kiezen):

$$\bar{x} - 1.96\sigma/\sqrt{n} = 112.7 - 1.96 \times 15/\sqrt{50} = 108.5,$$

en

$$\bar{x} + 1.96\sigma/\sqrt{n} = 112.7 + 1.96 \times 15/\sqrt{50} = 116.9.$$

Het 95% betrouwbaarheidsinterval voor μ is bijgevolg gelijk aan [108.5, 116.9]. Figuur 7.4 visualiseert dit betrouwbaarheidsinterval. Merk op dat het steekproefgemiddelde bij constructie altijd exact in het midden van het interval ligt.



Figuur 7.4: Visualisatie van het 95% betrouwbaarheidsinterval $[108.5, 116.9]$ *(volle zwarte lijn). De stip geeft het steekproefgemiddelde aan* $\bar{x} = 112.7$.

Wat is de precieze betekenis van '95% betrouwbaarheid'? Het populatiegemiddelde μ is een ongekend getal waarvoor er twee mogelijkheden zijn: ofwel ligt μ in het interval $[108.5, 116.9]$, ofwel ligt ze er niet in. Als we concluderen dat μ in het interval

$[108.5, 116.9]$ ligt, dan is deze conclusie ofwel correct ofwel fout. Hoe staat dit in verhouding tot die 95%?

Interpretatie van het betrouwbaarheidsinterval

Om dit te begrijpen, moeten we terugkeren naar het concept van de herhaalde steekproeftrekking. Zoals aangegeven in Hoofdstuk 6 is reproduceerbaarheid een belangrijk begrip: wat gebeurt er indien we het experiment herhalen op basis van een nieuwe steekproef?

We trekken een nieuwe steekproef van 50 studenten en bepalen het IQ. Voor deze steekproef is het steekproefgemiddelde $\bar{x} = 111.5$. Merk op dat σ en $z_{\alpha/2}$ ongewijzigd blijven omdat ze niet afhangen van de steekproef. Vervolgens berekenen we het betrouwbaarheidsinterval:

$$\bar{x} - 1.96\sigma/\sqrt{n} = 111.5 - 1.96 \times 15/\sqrt{50} = 107.3,$$

en

$$\bar{x} + 1.96\sigma/\sqrt{n} = 111.5 + 1.96 \times 15/\sqrt{50} = 115.7.$$

Op basis van deze tweede steekproef is het 95% betrouwbaarheidsinterval gelijk aan $[107.3, 115.7]$. Dit is een ander interval dan het interval op basis van de eerste steekproef. Dit is niet verwonderlijk: het betrouwbaarheidsinterval hangt af van het steekproefgemiddelde $\bar{x}$ en verschillende steekproeven zullen verschillende gemiddelden hebben. Dit resulteert bijgevolg in verschillende betrouwbaarheidsintervallen.

Het betrouwbaarheidsinterval is dus *variabel*: per steekproef zullen de grenzen verschillen. Tabel 7.1 illustreert dit: ze toont de steekproefgemiddelden en de grenzen van het 95% betrouwbaarheidsinterval voor 100 verschillende steekproeven.

Om de betekenis van de 95% zekerheid te begrijpen, moeten we de echte waarde van het populatiegemiddelde kennen. Hier is dit $\mu = 110$.

Opgelet: In de praktijk kennen we μ niet. Hier veronderstellen we ze dat ze gekend is, louter om de betekenis van het betrouwbaarheidsinterval te demonstreren.

Per steekproef kunnen we kijken of μ in het bijhorende betrouwbaarheidsinterval ligt. Voor de eerste steekproef is het interval $[108.5, 116.9]$ en ligt $\mu = 110$ er in (omdat $108.5 \leq \mu \leq 116.9$). Voor de tweede steekproef is dit ook zo (110 ligt in het interval $[107.3, 115.7]$). Dit kunnen we nagaan voor alle 100 steekproeven in Tabel 7.1. Voor sommige steekproeven zal μ niet tot het interval behoren. Voor steekproef 3 bijvoorbeeld is het interval $[110.2, 118.6]$ en deze bevat 110 niet (omdat $110 < 110.2$).

Steekproef	$\bar{x}$	$\bar{x} - 1.96\sigma/\sqrt{n}$	$\bar{x} + 1.96\sigma/\sqrt{n}$	Ligt $\mu = 110$ in het interval?
1	112.7	108.5	116.9	JA
2	111.5	107.3	115.7	JA
3	114.4	110.2	118.6	NEEN
4	110.3	106.1	114.5	JA
5	113.4	109.2	117.6	JA
⋮	⋮	⋮	⋮	⋮
99	112.3	108.1	116.5	JA
100	110.3	106.1	114.5	JA

Tabel 7.1: Het steekproefgemiddelde, de grenzen van het betrouwbaarheidsinterval en de aanduiding of het populatiegemiddelde in het interval ligt voor 100 *steekproeven met* $n = 50$ *(slechts enkele resultaten worden getoond in de tabel).*

Figuur 7.5 visualiseert de 100 betrouwbaarheidsintervallen horende bij de 100 steekproeven. Betrouwbaarheidsintervallen die zwart gekleurd zijn, bevatten het populatiegemiddelde, terwijl de rode intervallen in stippellijn het populatiegemiddelde niet bevatten. We zien dat er 95 van de 100 intervallen zwart gekleurd zijn. Bijgevolg bevatten 95% van de intervallen het populatiegemiddelde. De theorie van betrouwbaarheidsintervallen garandeert dat er exact 95% van alle intervallen het populatiegemiddelde zullen bevatten indien we het experiment een oneindig aantal keer herhalen[a].

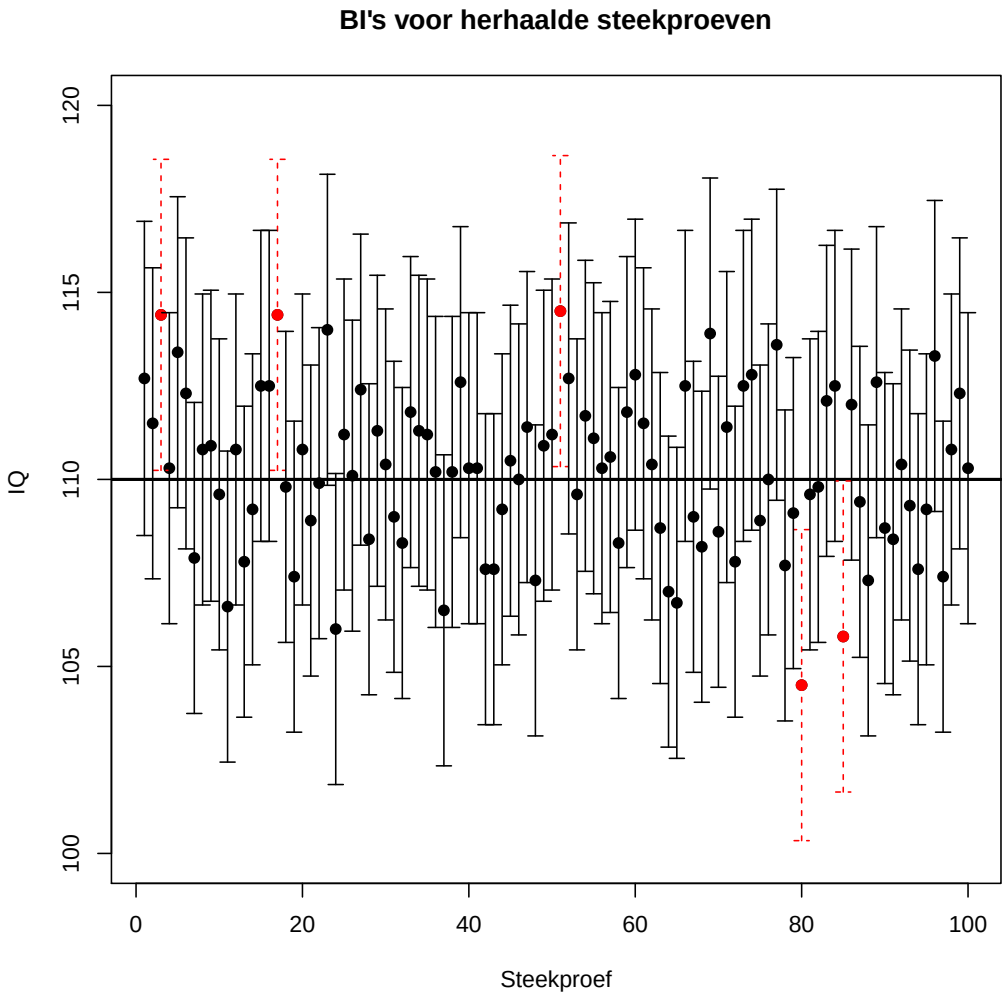
Eigenschappen van het betrouwbaarheidsinterval

Nu we de interpretatie van betrouwbaarheidsintervallen begrijpen, kunnen we de formule (7.5) in meer detail bekijken. De breedte van een interval $[a, b]$ is gelijk aan $b - a$. Als we dit toepassen op het betrouwbaarheidsinterval (7.5) krijgen we:

$$(\bar{X} + z_{\alpha/2}\sigma/\sqrt{n}) - (\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}) = 2 \times z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

De breedte hangt af van de steekproefgrootte n, de waarde $z_{\alpha/2}$ en de populatiestandaarddeviatie σ. De standaarddeviatie σ is een populatieparameter en kunnen we niet wijzigen. Anderzijds kunnen we n wel wijzigen door een kleinere of grotere steekproef te nemen. Als de steekproef groter wordt (dus n neemt toe), dan zal $1/\sqrt{n}$ kleiner worden zodat de breedte van het interval kleiner wordt. Figuur 7.6 illustreert dit voor $n = 100$ en $n = 500$: de betrouwbaarheidsintervallen voor $n = 500$ zijn smaller dan die voor $n = 100$.

[a]Bij dit voorbeeld is het exact 95% als we het experiment 100 keer herhalen, maar dit is toeval. Het is enkel als we het experiment een oneindig aantal keer herhalen dat het exact 95% moet zijn.



Figuur 7.5: Visualisatie van de 95% betrouwbaarheidsintervallen voor de 100 verschillende steekproeven met $n = 50$. De horizontale volle lijn duidt het populatiegemiddelde $\mu = 110$ aan en de stippen duiden de steekproefgemiddelden aan. De rode intervallen in stippellijn snijden de horizontale lijn niet en bevatten bijgevolg het populatiegemiddelde niet. Er zijn 95 van de 100 intervallen zwart gekleurd, wat overeenkomt met 95%.

We kunnen hier een intuïtieve verklaring aan geven: een smaller betrouwbaarheidsinterval impliceert dat we een meer nauwkeurige uitspraak kunnen formuleren over het populatiegemiddelde. Als de steekproefgrootte n toeneemt, verkrijgen we meer informatie over de populatie, wat zal resulteren in een meer nauwkeurige uitspraak over het populatiegemiddelde.

Als we α wijzigen, zal $z_{\alpha/2}$ en bijgevolg ook de breedte van het interval wijzigen. Als α toeneemt, zal $z_{\alpha/2}$ afnemen: als we in Figuur 7.2 z_α opschuiven naar links, zal de oppervlakte rechts toenemen. Figuur 7.7 toont de betrouwbaarheidsintervallen wanneer $\alpha = 0.01$ en $\alpha = 0.10$. Voor $\alpha = 0.01$ is het betrouwbaarheidsniveau 99% $(1-\alpha = 0.99)$ terwijl dit 90% is voor $\alpha = 0.1$ $(1 - \alpha = 0.90)$.

Als α toeneemt, dan zal de breedte van het interval afnemen. We kunnen hier een intuïtieve verklaring aan geven: als α toeneemt, zal $1 - \alpha$ afnemen en bijgevolg zal de kans dat het interval het populatiegemiddelde bevat, afnemen. Hoe smaller de intervallen, hoe kleiner de kans dat ze de populatieparameter zullen bevatten.

7.2.2 X normaal verdeeld en ongekende populatievariantie

Zoals eerder aangegeven is σ^2 in de praktijk vaak ongekend en kunnen we bijgevolg het interval zoals gegeven door formule (7.5) niet berekenen. We weten dat we de populatievariantie kunnen schatten door de steekproefvariantie S_X^2. We kunnen echter σ in uitdrukking (7.5) niet zomaar vervangen door S_X. Dit komt doordat S_X een variabele is terwijl σ een constante is. De afleiding van het betrouwbaarheidsinterval in paragraaf 7.2.1 gaat niet meer op indien we σ vervangen door S_X.

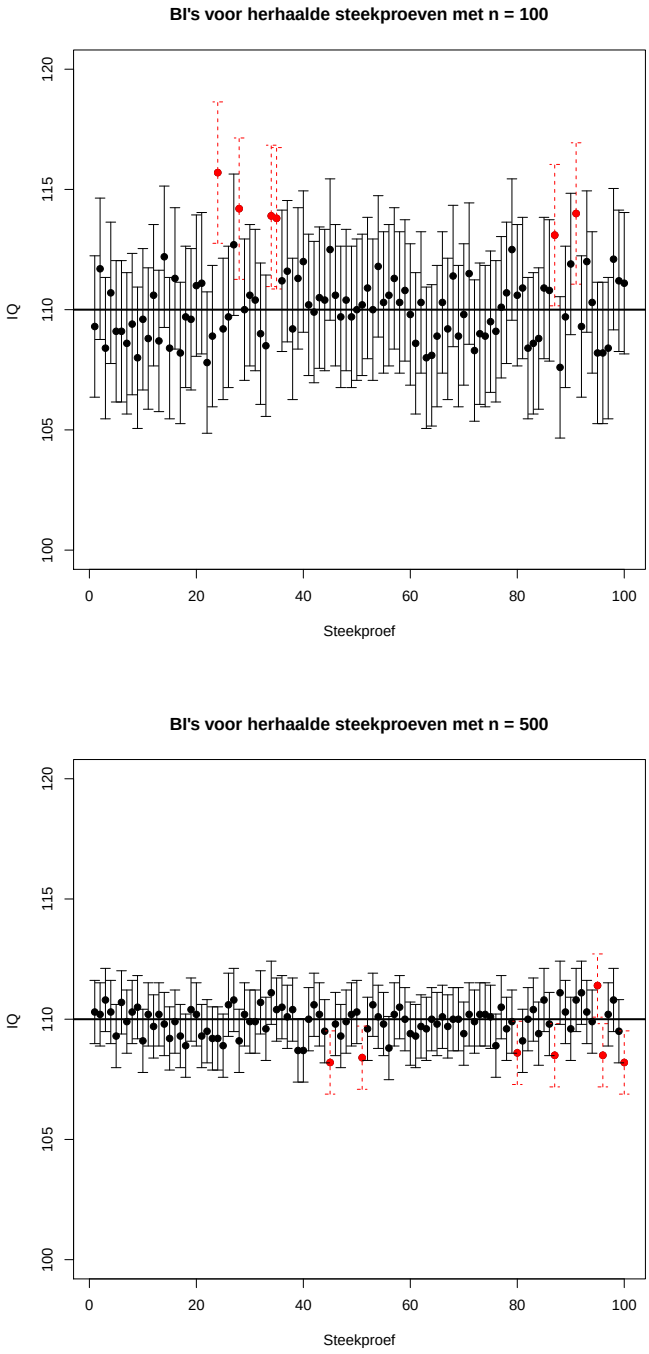
Door twee gekende eigenschappen te combineren, kunnen we echter een nieuw betrouwbaarheidsinterval opstellen die gebruik maakt van S_X in plaats van σ.

Eigenschap 1. Als X normaal verdeeld is, dan volgt uit Stelling 15 dat:

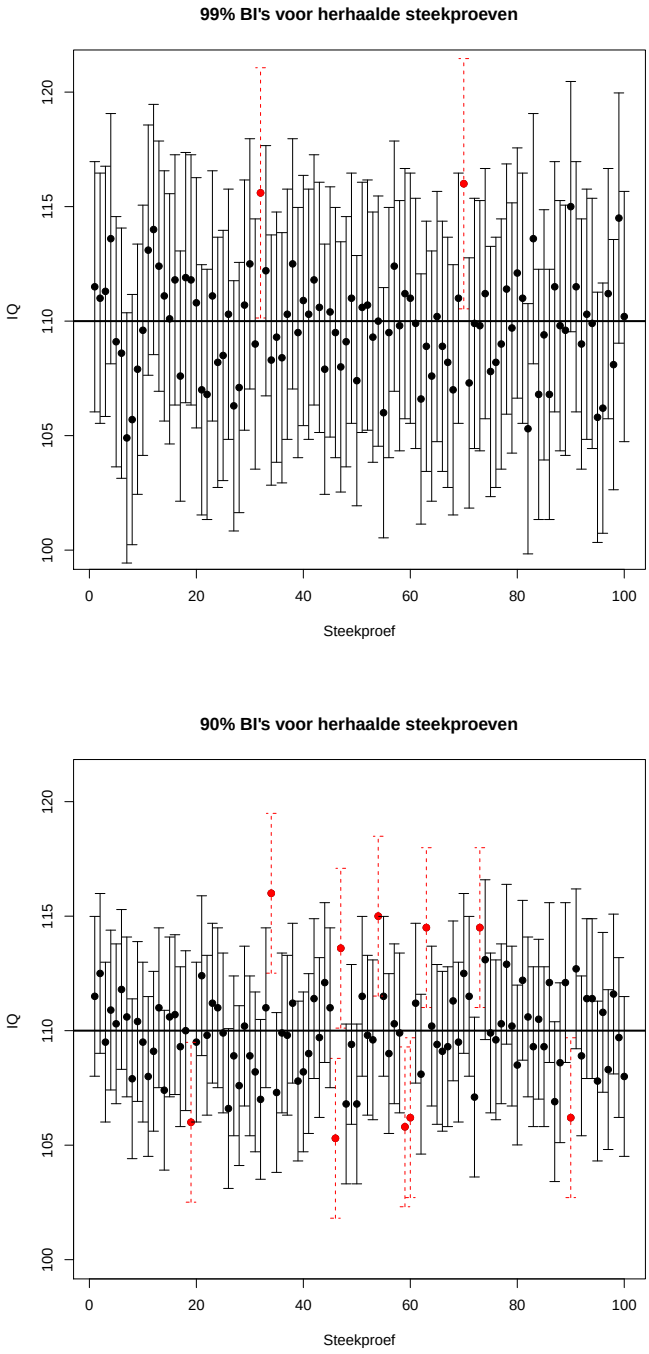
$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Eigenschap 2. Als X normaal verdeeld is, dan volgt:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$



Figuur 7.6: Visualisatie van de 95% betrouwbaarheidsintervallen voor 100 verschillende steekproeven met $n = 100$ (boven) en $n = 500$ (onder). De intervallen worden smaller als n toeneemt.



Figuur 7.7: Visualisatie van de 99% en 90% betrouwbaarheidsintervallen voor 100 verschillende steekproeven met $n = 50$. De intervallen worden smaller als α toeneemt (en dus $1 - \alpha$ afneemt).

Door eigenschappen 1 en 2 te combineren, kunnen we aantonen dat (zie paragraaf 5.6.4):

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)S_X^2}{\sigma^2}}{n-1}}} \sim t_{n-1}.$$

Deze uitdrukking kunnen we vereenvoudigen naar:

$$\frac{\bar{X}-\mu}{S_X/\sqrt{n}} \sim t_{n-1}. \tag{7.6}$$

Indien we σ in $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ vervangen door de steekproefstandaarddeviatie S_X dan wijzigt de verdeling van een standaardnormale naar een t_{n-1}-verdeling.

Gelijkaardig als bij de standaardnormale verdeling, duiden we met $t_{n-1;\alpha/2}$ de waarde aan van de t_{n-1}-verdeling zodat de oppervlakte rechts gelijk is aan $\alpha/2$, dus:

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

Omdat de t_{n-1}-verdeling symmetrisch is rond 0, kan men aantonen dat:

$$P(-t_{n-1;\alpha/2} \leq T \leq t_{n-1;\alpha/2}) = 1-\alpha, \tag{7.7}$$

waar $T \sim t_{n-1}$. Als we formules (7.6) en (7.7) samenvoegen, bekomen we:

$$P(-t_{n-1;\alpha/2} \leq \frac{\bar{X}-\mu}{S_X/\sqrt{n}} \leq t_{n-1;\alpha/2}) = 1-\alpha.$$

Analoog als in paragraaf 7.2.1 kunnen we dit herschrijven als:

$$P(\bar{X} - t_{n-1;\alpha/2}S_X/\sqrt{n} \leq \mu \leq \bar{X} + t_{n-1;\alpha/2}S_X/\sqrt{n}) = 1-\alpha,$$

zodat het $(1-\alpha)100\%$-betrouwbaarheidsinterval gelijk is aan:

$$[\bar{X} - t_{n-1;\alpha/2}S_X/\sqrt{n}, \bar{X} + t_{n-1;\alpha/2}S_X/\sqrt{n}]. \tag{7.8}$$

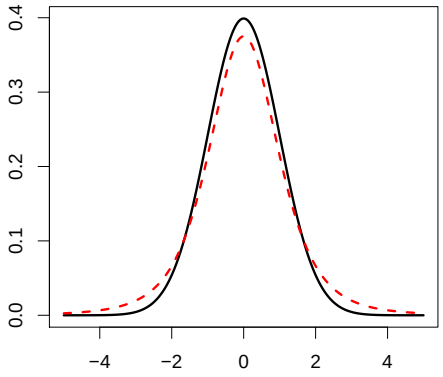
Merk op dat deze formule lijkt op formule (7.5) met $z_{\alpha/2}$ vervangen door $t_{n-1;\alpha/2}$ en σ door S_X. Indien we de populatievariantie niet kennen, moeten we het betrouwbaarheidsinterval dus berekenen op basis van formule (7.8).

Indien we de steekproefstandaarddeviatie gebruiken in plaats van de populatiestandaarddeviatie, moeten we de t_{n-1}-verdeling gebruiken. Figuur 7.8 toont de dichtheidsfunctie van een standaardnormale verdeling en een t-verdeling (met 4 vrijheidsgraden) ter illustratie. Niettegenstaande beide functies goed op elkaar lijken, zijn er toch enkele belangrijke verschillen:

- de t_{n-1}-verdeling heeft een grotere variantie dan de standaardnormale verdeling.
- de $t_{n-1;\alpha/2}$-waarde van een t_{n-1}-verdeling is groter dan de $z_{\alpha/2}$-waarde van een standaardnormale verdeling:

$$t_{n-1;\alpha/2} > z_{\alpha/2}.$$

Deze eigenschappen impliceren dat het betrouwbaarheidsinterval berekend op basis van formule (7.8) vaak breder zal zijn dan het betrouwbaarheidsinterval berekend op basis van formule (7.5). Dit komt doordat we de populatiestandaarddeviatie moeten schatten, wat zal resulteren in extra variabiliteit. Let wel: naarmate n groter wordt, zal de t_{n-1}-verdeling steeds beter de standaardnormale verdeling benaderen en zal $t_{n-1;\alpha/2} \approx z_{\alpha/2}$.



Figuur 7.8: De dichtheidsfunctie van een standaardnormale (zwarte volle lijn) en een t-verdeling (rode stippellijn) met 4 vrijheidsgraden.

We passen formule (7.8) toe op een nieuwe steekproef bestaande uit 30 studenten. Voor deze steekproef is $n = 30$, $\bar{x} = 112.7$ en $s_X = 12.8$. Voor de waarde $t_{29,0.025}$ geldt dat:

$$P(T \leq t_{29,0.025}) = 1 - 0.025 = 0.975, \quad T \sim t_{29},$$

zodat we uit Tabel 5.9 kunnen aflezen dat $t_{29,0.025} = 2.045$. Het betrouwbaarheidsinterval voor deze steekproef wordt gegeven door:

$$\bar{x} - t_{n-1;\alpha/2} s_X/\sqrt{n} = 112.7 - 2.045 \times 12.8/\sqrt{30} = 107.9,$$

en

$$\bar{x} + t_{n-1;\alpha/2} s_X/\sqrt{n} = 112.7 + 2.045 \times 12.8/\sqrt{30} = 117.5.$$

Het 95% betrouwbaarheidsinterval voor deze steekproef wordt dus gegeven door $[107.9, 117.5]$.

7.2.3 X niet normaal verdeeld en ongekende populatievariantie

Als X niet normaal verdeeld is, kunnen we voor een grote steekproef beroep doen op de centrale limietstelling. Deze garandeert dat het interval (7.8) bij benadering een $(1-\alpha)\times 100\%$ betrouwbaarheidsinterval is voor het populatiegemiddelde μ. We zullen dit niet bewijzen. Als X niet normaal verdeeld is en indien de steekproef klein is, hebben we andere methodes nodig om een betrouwbaarheidsinterval op te stellen. Deze methodes maken echter geen deel uit van de leerstof Statistiek I.

Illustratie in R

We illustreren hoe we formule (7.8) kunnen implementeren in R. De vector `IQ` bevat de IQ-scores van de 30 studenten uit paragraaf 7.2.2 en we veronderstellen dat deze uit een normale verdeling komt. We illustreren enkel het betrouwbaarheidsinterval wanneer σ^2 ongekend is omdat dit vaak in de praktijk zo is.

```
> IQ <- c(114, 137, 105, 123, 117,  91, 110, 126, 108,  94, 123,
+         105, 112,  91, 132, 110, 110, 110,  92, 102, 131,
+                 131, 104, 103, 125, 116, 121, 100, 115, 124)
> n <- length(IQ)
> n
```

```
[1] 30
```

```
> gem <- mean(IQ)
> gem
```

```
[1] 112.7333
```

```
> st.dev <- sd(IQ)
> st.dev
```

```
[1] 12.78721
```

De waarde $t_{29;0.025}$ staat voor de waarde zodat de oppervlakte *rechts* gelijk is aan 0.025:

$$P(T > t_{29;0.025}) = 0.025, \quad T \sim t_{n-1},$$

terwijl het commando `qt()` in R de oppervlakte *links* uitwerkt (dus kansen van de vorm $P(T \leq t)$). We kunnen nu de kritische waarde bekomen door gebruik te maken van de eigenschap:

$$P(T \leq t_{29;0.025}) = 1 - 0.025 = 0.975, \quad T \sim t_{n-1}.$$

In R wordt dit

```
> t.waarde <- qt(0.975, n-1)
> t.waarde
```

```
[1] 2.04523
```

Anderzijds kunnen we ook de optie `lower.tail = FALSE` gebruiken om aan te geven dat we de oppervlakte *rechts* ingeven:

```
> t.waarde <- qt(0.025, n-1, lower.tail = FALSE)
> t.waarde
```

```
[1] 2.04523
```

Tot slot berekenen we het 95% betrouwbaarheidsinterval:

```
> ondergrens <- gem - t.waarde*st.dev/sqrt(n)
> bovengrens <- gem + t.waarde*st.dev/sqrt(n)
> ondergrens
```

```
[1] 107.9585
```

```
> bovengrens
```

```
[1] 117.5082
```

7.3 Statistische toetsen

Naast het construeren van betrouwbaarheidsintervallen, zijn statistische toetsen ook een belangrijk middel om conclusies te formuleren over een populatieparameter. Wij zullen ons beperken tot de *t-toets voor één steekproef* wanneer X normaal verdeeld is of wanneer de steekproef groot is (in de praktijk vaak $n \geq 30$). We veronderstellen dat σ^2 ongekend is. We hernemen het onderzoek naar het gemiddeld IQ van studenten eerste bachelor Psychologie en eerste bachelor Pedagogische Wetenschappen.

Stel dat we de volgende vraag wensen te beantwoorden:

Is het gemiddeld IQ van de populatie verschillend van 115?

We kunnen deze vraag herformuleren in termen van *hypotheses*:

$$H_0 : \mu = 115 \quad \text{en} \quad H_a : \mu \neq 115.$$

H_0 wordt de *nulhypothese* genoemd en ze stelt dat het populatiegemiddelde gelijk is aan 115. Anderzijds wordt H_a de *alternatieve hypothese* genoemd en ze stelt dat het populatiegemiddelde verschillend is van 115.

Ofwel is H_0 correct ofwel is H_a correct en op basis van één steekproef wensen we een besluit te bekomen. Om dit te realiseren kunnen we bijvoorbeeld het steekproefgemiddelde $\bar{x}$ vergelijken met 115. Indien H_0 opgaat, verwacht je dat het steekproefgemiddelde 'in de buurt' van 115 zal liggen. Als anderzijds H_a opgaat, verwacht je dat $\bar{x}$ 'redelijk' verschillend van 115 zal zijn. Als we deze redenering omdraaien, kunnen we volgende beslissingsregels formuleren:

- als $\bar{x}$ 'ongeveer' gelijk is aan 115 zullen we H_0 niet verwerpen.
- als $\bar{x}$ 'sterk' verschilt van 115 zullen we H_0 verwerpen en H_a besluiten.

Bovenstaande redenering is intuïtief duidelijk en vormt de essentie van de statistische toets. De exacte uitwerking van de toets is echter iets complexer: we moeten immers op een objectieve manier vastleggen wat 'ongeveer' gelijk is en wat 'sterk' verschillend is. Voor onze steekproef van 30 studenten is $\bar{x} = 112.7$. Is dit verschillend genoeg van 115 om H_0 te verwerpen? Door gebruik te maken van de steekproevenverdeling van $(\bar{X} - 115)/(S_X/\sqrt{n})$ zullen we hierop een antwoord kunnen geven.

Om formeel de toets in te voeren, gebruiken we meer algemene notatie. We schrijven de hypotheses als:

$$H_0 : \mu = \mu_0 \quad \text{en} \quad H_a : \mu \neq \mu_0,$$

waar μ_0 een gegeven waarde is. In ons voorbeeld is $\mu_0 = 115$. Het is de conventie om H_0 te schrijven in termen van een gelijkheid en H_a in termen van een ongelijkheid. De keuze $H_0 : \mu \neq \mu_0$ en $H_a : \mu = \mu_0$ is bijgevolg niet correct. De alternatieve hypothese $H_a : \mu \neq \mu_0$ wordt de *tweezijdig* alternatieve hypothese genoemd. De *eenzijdige* alternatieve hypotheses zijn $H_a : \mu > \mu_0$ en $H_a : \mu < \mu_0$ en worden besproken in paragraaf 7.3.5.

Bij een statistische toets zullen we trachten H_0 te verwerpen: op basis van de data in de steekproef zullen we trachten 'bewijs' te vinden tegen H_0. Eenmaal we voldoende bewijs gevonden hebben, zullen we H_0 'verwerpen'. Het kan uiteraard ook zijn dat we geen bewijs vinden tegen H_0 en dan zullen we H_0 'niet verwerpen' (soms zullen we ook spreken over H_0 'aanvaarden'). Het bewijs tegen H_0 zullen we samenvatten door middel van een *toetsingsgrootheid*.

7.3.1 Toetsingsgrootheid

In paragraaf 7.2.2 hebben we gezien dat:

$$\frac{\bar{X} - \mu}{S_X/\sqrt{n}} \sim t_{n-1}, \quad \text{als} \quad X \sim N(\mu, \sigma^2).$$

Veronderstel nu dat de nulhypothese opgaat (zodat $\mu = \mu_0$), dan volgt dat:

$$\frac{\bar{X} - \mu_0}{S_X/\sqrt{n}} \sim t_{n-1}, \quad \text{als} \quad X \sim N(\mu, \sigma^2).$$

Deze steekproefgrootheid vormt een belangrijk onderdeel van de statistische toets en noteren we kort als G:

$$G = \frac{\bar{X} - \mu_0}{S_X/\sqrt{n}}. \tag{7.9}$$

Om te benadrukken dat G een t_{n-1}-verdeling volgt op voorwaarde dat H_0 correct is, noteren we dit nog als:

$$G \overset{H_0}{\sim} t_{n-1}. \tag{7.10}$$

De steekproefgrootheid G wordt de *toetsingsgrootheid* genoemd en t_{n-1} is de verdeling van de toetsingsgrootheid wanneer de nulhypothese waar is. De waarde van G die we bekomen op basis van één steekproef noteren we als g. Als we terugkeren naar de steekproef met 30 studenten waarvoor $\bar{x} = 112.7$ en $s_X = 12.8$, dan volgt dat

$$g = \frac{\bar{x} - \mu_0}{s_X/\sqrt{n}} = \frac{112.7 - 115}{12.8/\sqrt{30}} = -0.98.$$

Hoe kunnen we nu op basis van de waarde $g = -0.98$ besluiten of H_0 of H_a waar is? Om dit te weten te komen, moeten we eerst een beter idee krijgen welke waarden G kan aannemen wanneer H_0 waar is en welke waarden G kan aannemen wanneer H_0 niet waar is.

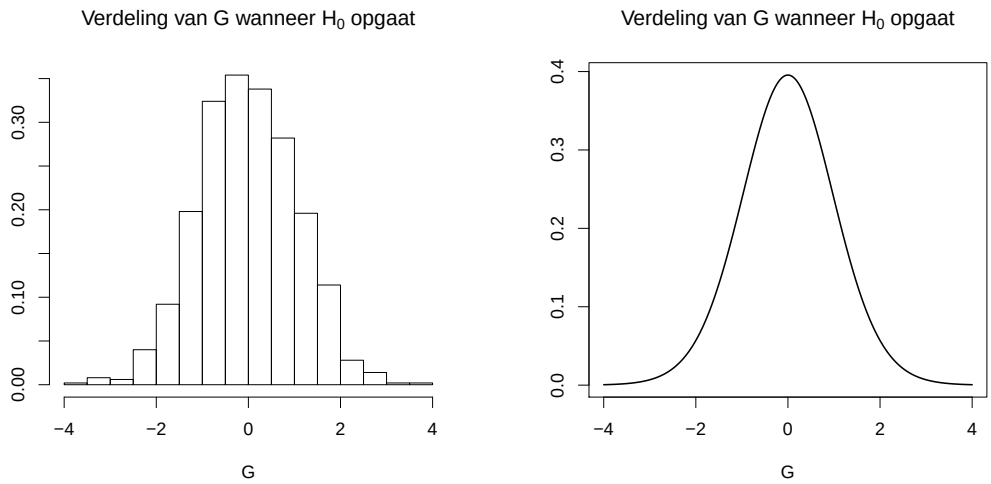
Veronderstel eerst dat H_0 waar is (dus $\mu = 115$) en veronderstel dat we het experiment 1000 keer herhalen. Tabel 7.2 illustreert dit op basis van gesimuleerde (artificiële) data. Per steekproef berekenen we het steekproefgemiddelde en de steekproefstandaarddeviatie om vervolgens g te berekenen. Figuur 7.9 toont het histogram van deze 1000 waarden van de toetsingsgrootheid G. Indien we dit een oneindig aantal keer herhalen (i.p.v. 1000 herhalingen), dan stelt eigenschap (7.10) dat de verdeling van G gelijk is aan een t_{n-1}-verdeling (Figuur 7.9 rechts). Als we kijken naar de waarden van G zien we dat de meeste waarden klein zijn en rond 0 liggen.

Steekproef	$\bar{x}$	s_X	$g = \frac{\bar{x}-\mu_0}{s_X/\sqrt{n}}$
1	117.8	12.8	1.21
2	120.6	14.0	2.18
3	112.0	14.8	-1.10
4	114.7	16.7	-0.10
5	111.5	18.2	-1.05
6	113.0	16.9	-0.64
7	114.4	14.8	-0.21
8	115.0	18.0	0.01
⋮	⋮	⋮	⋮
999	118.2	14.7	1.18
1000	116.1	17.4	0.35

Tabel 7.2: Steekproefgemiddelden, steekproefstandaarddeviaties en toetsingsgrootheden horende bij 1000 *verschillende steekproeven met* $n = 30$ *en wanneer* H_0 *opgaat.*

Veronderstel vervolgens dat H_0 niet waar is, bijvoorbeeld $\mu = 125$ zodat $\mu > \mu_0$. Tabel 7.3 toont de waarden van G (die we nog steeds berekenen via formule (7.9)) afkomstig uit 1000 verschillende (gesimuleerde) steekproeven en Figuur 7.10 toont het bijhorende histogram. We zien dat G nu enkel positieve waarden aanneemt en dat de waarden vaak groter zijn in vergelijking met de waarden uit Figuur 7.9.

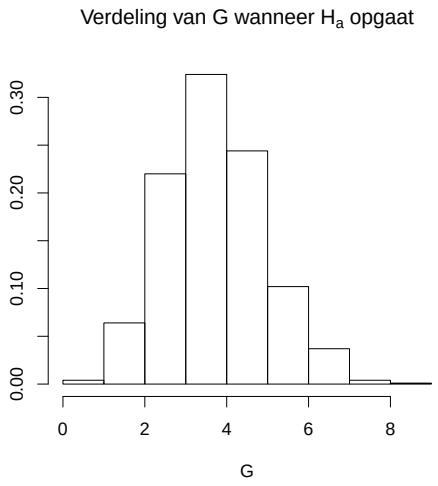
Veronderstel dat $\mu = 105$ zodat H_0 nog steeds niet waar is, maar nu is $\mu < \mu_0$. Tabel 7.4 toont de waarden van G afkomstig uit 1000 verschillende (gesimuleerde) steekproeven en Figuur 7.11 toont het bijhorende histogram. We zien dat G nu enkel negatieve waarden aanneemt en dat de waarden vaak kleiner zijn in vergelijking met de waarden uit Figuur 7.9.



Figuur 7.9: De verdeling van G wanneer de nulhypothese opgaat op basis van 1000 steekproeven (links) en de theoretische verdeling t_{n-1} op basis van oneindig veel steekproeven (rechts).

Steekproef	$\bar{x}$	s_X	$g = \frac{\bar{x}-\mu_0}{s_X/\sqrt{n}}$
1	127.8	12.8	5.50
2	130.6	14.0	6.09
3	122.0	14.8	2.60
4	124.7	16.7	3.18
5	121.5	18.2	1.95
6	123.0	16.9	2.61
7	124.4	14.8	3.50
8	125.0	18.0	3.05
⋮	⋮	⋮	⋮
999	128.2	14.7	4.91
1000	126.1	17.4	3.49

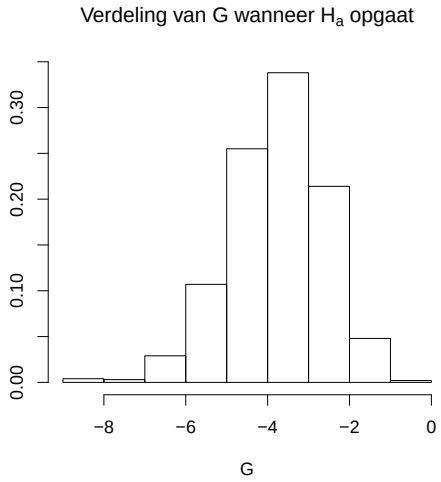
Tabel 7.3: Steekproefgemiddelden, steekproefstandaarddeviaties en toetsingsgrootheden horende bij 1000 *verschillende steekproeven met $n = 30$ en wanneer H_a opgaat met $\mu = 125$ (dus $\mu > \mu_0$).*



Figuur 7.10: De verdeling van G wanneer de alternatieve hypothese opgaat (met $\mu = 125$) op basis van 1000 steekproeven.

Steekproef	$\bar{x}$	s_X	$g = \frac{\bar{x}-\mu_0}{s_X/\sqrt{n}}$
1	107.8	12.8	-3.08
2	110.6	14.0	-1.74
3	102.0	14.8	-4.79
4	104.7	16.7	-3.37
5	101.5	18.2	-4.06
6	103.0	16.9	-3.88
7	104.4	14.8	-3.92
8	105.0	18.0	-3.02
⋮	⋮	⋮	⋮
999	108.2	14.7	-2.55
1000	106.1	17.4	-2.79

Tabel 7.4: Steekproefgemiddelden, steekproefstandaarddeviaties en toetsingsgrootheden horende bij 1000 *verschillende steekproeven met $n = 30$ en wanneer H_a opgaat met $\mu = 105$ (dus $\mu < \mu_0$).*



Figuur 7.11: De verdeling van G wanneer de alternatieve hypothese opgaat (met $\mu = 105$) op basis van 1000 steekproeven.

7.3.2 Beslissingsregels

Door de toetsingsgrootheid te bestuderen onder H_0 en H_a bekomen we volgende conclusies:

- als H_0 waar is, verwachten we dat G waarden zal aannemen 'rond' 0.
- als H_0 niet waar is, verwachten we dat G waarden zal aannemen 'sterk' verschillend van 0.

In de praktijk weten we uiteraard niet of H_0 of H_a waar is, maar op basis van één steekproef kunnen we g wel berekenen. Op basis van die g kunnen we dan volgende regels opstellen:

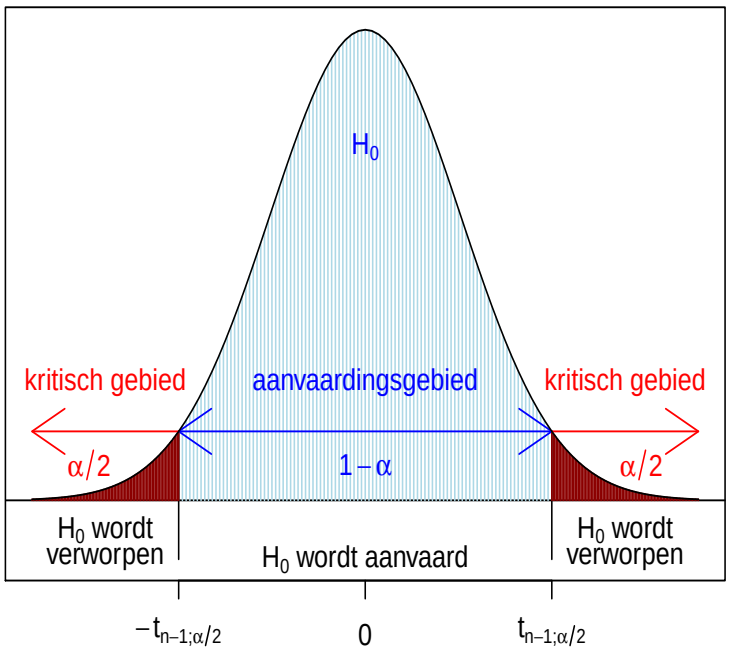
- als g 'rond' 0 ligt, verwerpen we H_0 niet.
- als g 'sterk' verschilt van 0, verwerpen we H_0 en besluiten we H_a.

Dan rest ons nog de vraag: wanneer is g 'sterk' verschillend van 0? Wij zullen volgende *beslissingsregels* gebruiken:

- als $-t_{n-1;\alpha/2} \leq g \leq t_{n-1;\alpha/2}$ verwerpen we H_0 niet.

- als $g < -t_{n-1;\alpha/2}$ of $g > t_{n-1;\alpha/2}$ verwerpen we H_0 en besluiten we H_a.

De waarden $-t_{n-1;\alpha/2}$ en $t_{n-1;\alpha/2}$ worden de *kritische waarden* van de toets genoemd. Deze beslissingsregels worden visueel weergegeven in Figuur 7.12. Het gebied tussen de waarden $-t_{n-1;\alpha/2}$ en $t_{n-1;\alpha/2}$ noemt men het *aanvaardingsgebied*. Het gebied buiten deze twee waarden wordt het *kritisch gebied* genoemd.



Figuur 7.12: Visualisatie van de beslissingsregels.

Als we bijvoorbeeld $\alpha = 0.05$ nemen, dan halen we uit Tabel 5.9 dat $t_{30-1;0.05/2} = t_{29;0.025} = 2.045$. We hadden eerder al berekend dat $g = -0.98$. Omdat $-2.045 < -0.98 < 2.045$ verwerpen we $H_0 : \mu = 115$ niet. Op basis van de data hebben we *geen* bewijs gevonden dat het populatiegemiddelde verschillend is van 115.

Door middel van een statistische toets hebben we dus op basis van de steekproef een besluit geformuleerd over de populatie. Merk op dat we de beslissingsregels meer compact kunnen schrijven als:

- als $|g| \leq t_{n-1;\alpha/2}$ verwerpen we H_0 niet.
- als $|g| > t_{n-1;\alpha/2}$ verwerpen we H_0 en besluiten we H_a.

7.3.3 Type I en type II fout

Als we op basis van een statistische toets een besluit formuleren omtrent de hypotheses, bestaat er altijd een mogelijkheid dat we fout zijn. Er zijn 4 scenario's mogelijk waarvan er twee resulteren in een correcte conclusie en twee in een foutieve conclusie. De 4 scenario's zijn:

A De nulhypothese is waar (zodat in werkelijkheid $\mu = \mu_0$) en we verwerpen H_0 niet.

B De nulhypothese is waar (zodat in werkelijkheid $\mu = \mu_0$) en we verwerpen H_0.

C De alternatieve hypothese is waar (zodat in werkelijkheid $\mu \neq \mu_0$) en we verwerpen H_0 niet.

D De alternatieve hypothese is waar (zodat in werkelijkheid $\mu \neq \mu_0$) en we verwerpen H_0.

Tabel 7.5 geeft deze 4 scenario's schematisch weer.

	In werkelijkheid is H_0	
	juist	fout
We verwerpen H_0 niet	Juiste beslissing (A) Betrouwbaarheid $(1-\alpha)$	Foute beslissing (C) Type II fout β
We verwerpen H_0	Foute beslissing (B) Type I fout α	Juiste beslissing (D) Onderscheidingsvermogen $(1-\beta)$

Tabel 7.5: Overzicht van de 4 scenario's.

In scenario's A en D is ons besluit op basis van de statistische toets correct, terwijl we in scenario's B en C een foutieve conclusie formuleren. Het is evident dat we de kans dat we foutief zijn, zo klein mogelijk wensen te houden. De kans om de fout te maken zoals beschreven in scenario B noteren we symbolisch als:

$$P(\text{verwerp } H_0 \mid \mu = \mu_0).$$

We lezen deze kans als: "de kans om H_0 te verwerpen terwijl in werkelijkheid $\mu = \mu_0$". De statistische toets op basis van de beslissingsregels uit paragraaf 7.3.2 garandeert dat deze kans gelijk is aan α:

$$P(\text{verwerp } H_0 \mid \mu = \mu_0) = \alpha. \tag{7.11}$$

De kans om de foutieve beslissing zoals beschreven in scenario B te maken, is dus gelijk aan α. Dit komt net door de specifieke keuze van de kritische waarden; zie Figuur 7.12. De totale oppervlakte onder de curve in de verwerpingsgebieden is $\alpha/2 + \alpha/2 = \alpha$.

De kans (7.11) moeten we opnieuw interpreteren via de herhaalde steekproeftrekking. Stel dat in werkelijkheid $\mu = \mu_0$ en we herhalen het experiment vele malen op basis van nieuwe steekproeven (theoretisch gezien herhalen we het experiment een oneindig aantal keer). Per steekproef kunnen we het gemiddelde, de standaarddeviatie en de toetsingsgrootheid g berekenen. Vervolgens kunnen we via de beslissingsregels een conclusie formuleren (al dan niet H_0 verwerpen). De uitdrukking (7.11) stelt dat de proportie van steekproeven waarvoor we H_0 verwerpen (en we bijgevolg een fout maken), gelijk is aan α. Vaak kiezen we $\alpha = 0.05$, zodat deze kans klein is, namelijk 5%.

We kunnen dit toepassen op het voorbeeld rond IQ: indien in werkelijkheid $\mu = 115$ en we herhalen het experiment vele malen, zullen we voor slechts 5% van de steekproeven een foutief besluit formuleren. Dit impliceert dat, indien $\mu = 115$, we in 95% van de gevallen een correct besluit over de populatie zullen formuleren. Wij zijn niet 100% zeker van ons besluit, maar 95% en we kunnen de kans dat we foutief zijn, indien $\mu = 115$, controleren (via de keuze van α).

De kans (7.11) noemen we ook de kans op een *type I fout* en α wordt het *significatieniveau* genoemd. De kans dat we een correcte conclusie bekomen indien $\mu = \mu_0$:

$$P(\text{verwerp } H_0 \text{ niet} \mid \mu = \mu_0) = 1 - \alpha,$$

wordt de *betrouwbaarheid* genoemd.

Opgelet: de kans op een type I fout is *exact* gelijk aan α als X uit een normale verdeling komt. In deze cursus zullen we echter geen methodes zien om na te gaan of X uit een normale verdeling komt. Indien X niet uit een normale verdeling komt, dan garandeert de centrale limietstelling dat de kans op een type I fout *bij benadering* gelijk is aan α indien de steekproef groot is. Indien X niet uit een normale verdeling komt en de steekproef klein is, kan de kans op een type I fout sterk verschillen van α en zullen we de toets niet gebruiken: als we niet weten wat de kans op een fout is, heeft de statistische toets geen meerwaarde. In deze cursus zullen we geen toets behandelen voor dit geval.

Nu bestaat er ook een kans dat we een ander type fout maken, namelijk die beschreven in scenario C: in werkelijkheid is $\mu \neq \mu_0$, maar op basis van de statistische toets besluiten

we H_0 niet te verwerpen. Dit wordt een kans op een *type II fout* genoemd en duiden we aan met β (uitspraak *bèta*):

$$P(\text{verwerp } H_0 \text{ niet} \mid \mu \neq \mu_0) = \beta. \tag{7.12}$$

De kans dat we een correcte conclusie formuleren terwijl in werkelijkheid $\mu \neq \mu_0$ is, wordt de *onderscheidingskans* (of *power*) genoemd en is gelijk aan $1 - \beta$:

$$P(\text{verwerp } H_0 \mid \mu \neq \mu_0) = 1 - \beta.$$

De kans op een type I fout is gelijk aan α en kunnen we controleren via de statistische toets (omdat we α vrij kunnen kiezen bij de beslissingsregels). Voor de kans op een type II fout is dit niet zo, we kunnen ze niet exact controleren via de statistische toets. De kans op een type II fout (dus β) hangt onder andere af van volgende factoren:

- het significantieniveau α: β stijgt als α daalt.
- de steekproefgrootte n: β daalt als n stijgt.

De kans op een type I fout kunnen we zelf controleren en zetten we typisch op 5% (maar andere keuzes zijn ook mogelijk). Maar waarom zetten we deze kans niet lager, bijvoorbeeld op 0.01%? Op het eerste zicht lijkt dit logisch, omdat de kans op een type I fout dan kleiner wordt (en hoe kleiner de kans op een foutief besluit, hoe beter). Dit zal echter resulteren in een verhoogde kans op een type II fout, wat niet wenselijk is. De algemene conventie is om $\alpha = 0.05$ te kiezen, maar men kan daar van afwijken.

De kans op een type II fout kan men inschatten via een specifieke analyse (een power-analyse). Als de kans op een type II fout te groot is, kan men ervoor opteren om een grotere steekproef te nemen (dan zal deze kans dalen). Dit zullen we niet behandelen in deze cursus, maar zal aan bod komen in de cursus 'Statistiek II'.

Indien we H_0 niet kunnen verwerpen op basis van de toets wil dit niet zeggen dat H_0 waar is. Het kan zijn dat de power van de test zeer laag is (doordat we bijvoorbeeld een kleine steekproef hebben) en dat we daardoor H_0 niet kunnen verwerpen. Daarom spreken we over het *niet verwerpen* van H_0, wat minder sterk geformuleerd is dan bijvoorbeeld het *besluiten* van H_0.

7.3.4 Beslissingsregels op basis van het betrouwbaarheidsinterval

Men kan aantonen dat de beslissingsregels voor een toets op het α significatieniveau (zie paragraaf 7.3.2) equivalent kunnen worden uitgedrukt door gebruik te maken van

een $(1-\alpha)100\%$ betrouwbaarheidsinterval:

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

Deze regels zijn:

- als μ_0 in het betrouwbaarheidsinterval ligt, verwerpen we H_0 niet.
- als μ_0 niet in het betrouwbaarheidsinterval ligt, verwerpen we H_0 en besluiten we $H_a : \mu \neq \mu_0$.

In plaats van deze gelijkheid wiskundig te bewijzen, zullen we een meer intuïtieve verklaring geven aan deze beslissingsregels. Voor de eenvoud kiezen we $\alpha = 0.05$.

Veronderstel dat het betrouwbaarheidsinterval μ_0 niet bevat. Dan zijn we 95% zeker dat $\mu \neq \mu_0$. Dit impliceert dat er een kans is van 5% dat we verkeerdelijk zeggen dat $\mu \neq \mu_0$. Dit is gelijk aan de kans om ten onrechte $\mu = \mu_0$ te verwerpen (dit is een type I fout). Er is bijgevolg 5% kans om een type I fout te maken. Deze laatste uitdrukking is gelijk aan het verwerpen van H_0 op het 5% significatieniveau.

Illustratie in R

De t-toets voor één steekproef kan in R uitgevoerd worden via `t.test(data, mu = 115)`, waar `mu = 115` staat voor $\mu_0 = 115$.

```
> IQ <- c(114, 137, 105, 123, 117,  91, 110, 126, 108,  94, 123,
+         105, 112,  91, 132, 110, 110, 110,  92, 102, 131,
+                 131, 104, 103, 125, 116, 121, 100, 115, 124)
> t.test(IQ,  mu = 115)


        One Sample t-test

data:  IQ
t = -0.9709, df = 29, p-value = 0.3396
alternative hypothesis: true mean is not equal to 115
95 percent confidence interval:
 107.9585 117.5082
```

```
sample estimates:
mean of x
 112.7333
```

`t = -0.9709` geeft de toetsingsgrootheid $g = -0.9709$ (R gebruikt de conventie om de toetsingsgrootheid aan te duiden met `t` in plaats van `g`; ook zal R tussenresultaten niet afronden).

Deze waarde moeten we vergelijken met de kritische waarde $t_{29;0.025}$. Herinner je dat $t_{29;0.025}$ staat voor de waarde zodat de oppervlakte *rechts* gelijk is aan 0.025:

$$P(T > t_{29;0.025}) = 0.025, \quad T \sim t_{n-1},$$

terwijl het commando `qt()` in R de oppervlakte *links* uitwerkt (dus kansen van de vorm $P(T \leq t)$). We kunnen nu de kritische waarde bekomen door gebruik te maken van de eigenschap:

$$P(T \leq t_{29;0.025}) = 1 - 0.025 = 0.975, \quad T \sim t_{n-1}.$$

In R wordt dit

```
> qt(0.975, 29)
```

```
[1] 2.04523
```

Anderzijds kunnen we ook de optie `lower.tail = FALSE` gebruiken om aan te geven aan R dat we de oppervlakte rechts bedoelen:

```
> qt(0.025, 29, lower.tail = FALSE)
```

```
[1] 2.04523
```

Uiteraard zijn beide kritische waarden aan elkaar gelijk. We bekomen $t_{29;0.025} = 2.04523$.

Zoals eerder aangegeven aanvaarden we H_0 omdat $-t_{n-1;\alpha/2} \leq g \leq t_{n-1;\alpha/2}$. We kunnen dan als volgt een besluit formuleren:

Op het 5% significantieniveau hebben we geen bewijs gevonden dat het populatiegemiddelde verschillend is van 115.

Merk op dat we het besluit genuanceerd formuleren: we schrijven bijvoorbeeld niet dat we besluiten dat het populatiegemiddelde gelijk is aan 115. Deze formulering is te sterk, wetende dat er altijd een kans is dat we fout zijn en dat we de kans op een type II fout niet kennen. We vermelden ook steeds het significantieniveau, want het besluit van de test hangt af van α. Indien we bijvoorbeeld $\alpha = 0.40$ nemen, is

```
> qt(0.20, 29, lower.tail = FALSE)
```

```
[1] 0.854192
```

zodat $t_{n-1;\alpha/2} = t_{29;0.20} = 0.85$ en zouden we H_0 verwerpen (omdat $g < -t_{29;0.20}$). Dus op basis van een andere keuze van α kunnen we een ander besluit bekomen. Dit is geen tegenstrijdigheid, want door α te wijzigen, wijzigt ook de kans op type I en type II fouten. Bij $\alpha = 0.40$ is er 40% kans om een type I fout te maken. Dit is zeer hoog en zeer ongebruikelijk. Zoals eerder aangeven kiest men vaak voor $\alpha = 0.05$.

Merk op dat `t.test()` ook het 95% betrouwbaarheidsinterval geeft $[107.9585, 117.5082]$ en het steekproefgemiddelde $\bar{x} = 112.7333$. Zoals verwacht ligt $\mu_0 = 115$ in het betrouwbaarheidsinterval (zie paragraaf 7.3.4).

De waarde `df` in de output komt overeen met de vrijheidsgraden (*degrees of freedom*) van de t_{n-1}-verdeling. Hier is dit $n - 1$, dus $30 - 1 = 29$ voor deze steekproef. `p-value = 0.3396` geeft de p-waarde (of overschrijdingskans) dewelke we verder in de syllabus bespreken.

7.3.5 Eenzijdige en tweezijdige toetsen

De statistische toets die we in de vorige paragrafen hebben besproken, wordt ook de *tweezijdige* t-toets voor één steekproef genoemd. Dit komt doordat de alternatieve hypothese $H_a : \mu \neq \mu_0$ toelaat dat het populatiegemiddelde zowel kleiner als groter kan zijn dan μ_0. Het kan echter zijn dat de onderzoeker enkel geïnteresseerd is in een *eenzijdige* alternatieve hypothese: $H_a : \mu < \mu_0$ (ook de *linkszijdige* alternatieve genoemd) of $H_a : \mu > \mu_0$ (ook de *rechtszijdige* alternatieve genoemd).

Rechtszijdig

Stel dat we voor het voorbeeld rond de IQ-scores volgende vraag wensen te beantwoorden:

Is het gemiddeld IQ van de populatie groter dan 115?

We kunnen deze vraag formuleren door middel van hypotheses:

$$H_0 : \mu = 115 \quad \text{en} \quad H_a : \mu > 115.$$

De nulhypothese is nog steeds een gelijkheid, net als bij de tweezijdige toets. De alternatieve hypothese is nu echter eenzijdig: uit de onderzoeksvraag halen we dat men enkel geïnteresseerd is of het gemiddelde groter is dan 115. Het uitvoeren van een dergelijke eenzijdige toets is zeer gelijkaardig aan de tweezijdige: we starten met het berekenen van de toetsingsgrootheid (7.9). Vervolgens hebben we beslissingsregels nodig om H_0 al dan niet te verwerpen. Deze beslissingsregels zijn *verschillend* van die van de tweezijdige toets uit paragraaf 7.3.2. Dit komt doordat de beslissingsregels afhangen van de keuze van alternatieve hypothese. Meer specifiek zijn de beslissingsregels:

- als $g \leq t_{n-1;\alpha}$ verwerpen we H_0 niet.
- als $g > t_{n-1;\alpha}$ verwerpen we H_0 en besluiten we H_a.

Figuur 7.13 visualiseert deze beslissingsregels. Omdat het verwerpingsgebied rechts ligt, wordt de toets horende bij de eenzijdige alternatieve $H_a : \mu > \mu_0$ ook de *rechtszijdige toets* genoemd.

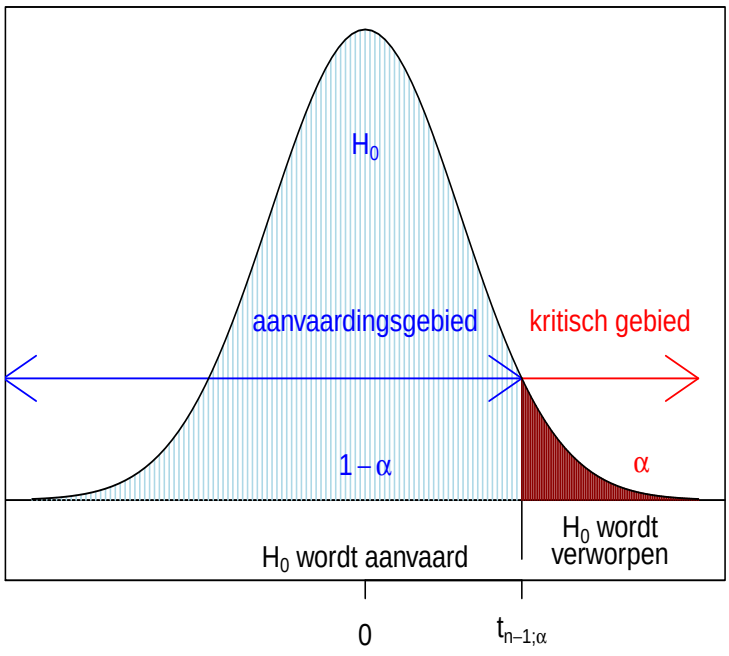
Ook voor eenzijdige toetsen kunnen we een betrouwbaarheidsinterval opstellen. Het rechtszijdig betrouwbaarheidsinterval wordt gegeven door:

$$[\bar{X} - t_{n-1;\alpha}\frac{S_X}{\sqrt{n}}, +\infty[,$$

waarbij $+\infty$ staat voor 'plus oneindig'. Analoog als bij de tweezijdige toets, kunnen we het betrouwbaarheidsinterval gebruiken om conclusies te formuleren:

- als μ_0 in het betrouwbaarheidsinterval ligt, verwerpen we H_0 niet.
- als μ_0 niet in het betrouwbaarheidsinterval ligt, verwerpen we H_0 en besluiten we $H_a : \mu > \mu_0$.

Merk op: het aanvaardingsgebied bepaald door de beslissingsregels heeft een geschatte bovengrens, terwijl het betrouwbaarheidsinterval een geschatte ondergrens heeft. Toch leiden beide regels tot dezelfde conclusie. Dit zal in meer detail besproken worden tijdens de oefeningensessies.



Figuur 7.13: Visualisatie van de eenzijdige beslissingsregel bij $H_a : \mu > \mu_0$.

Linkszijdig

Veronderstel nu dat we voor het voorbeeld rond de IQ-scores volgende vraag wensen te beantwoorden:

Is het gemiddeld IQ van de populatie kleiner dan 115?

Dit kunnen we vertalen in volgende hypotheses:

$$H_0 : \mu = 115 \quad \text{en} \quad H_a : \mu < 115.$$

Dit is ook een eenzijdige alternatieve hypothese. De beslissingsregels zijn:

- als $g > -t_{n-1;\alpha}$ verwerpen we H_0 niet.
- als $g < -t_{n-1;\alpha}$ verwerpen we H_0 en besluiten we H_a.

Figuur 7.14 visualiseert deze beslissingsregels. Omdat het verwerpingsgebied links ligt, wordt de toets horende bij de eenzijdige alternatieve $H_a : \mu < \mu_0$ ook de *linkszijdige toets* genoemd.

Het linkszijdig betrouwbaarheidsinterval wordt gegeven door:

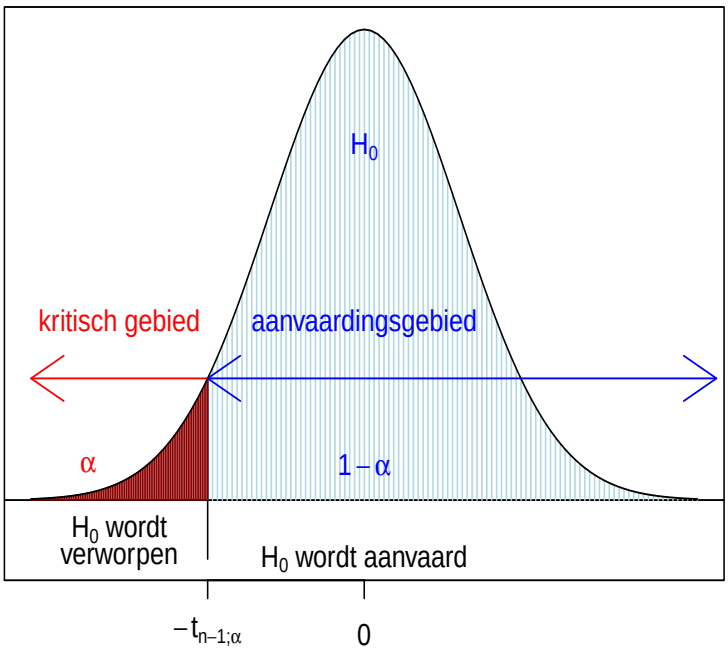
$$]-\infty, \bar{X} + t_{n-1;\alpha}\frac{S_X}{\sqrt{n}}],$$

waarbij $-\infty$ staat voor 'min oneindig'. We kunnen opnieuw het betrouwbaarheidsinterval gebruiken om conclusies te formuleren:

- als μ_0 in het betrouwbaarheidsinterval ligt, verwerpen we H_0 niet.
- als μ_0 niet in het betrouwbaarheidsinterval ligt, verwerpen we H_0 en besluiten we $H_a : \mu < \mu_0$.

Eenzijdig of tweezijdig toetsen?

Eenzijdige en tweezijdige toetsen hebben elk hun voor- en nadelen. Voor beide types van toetsen is de kans op een type I fout gelijk aan α, maar de power zal verschillen. Indien de nulhypothese niet opgaat, zijn er twee mogelijkheden: ofwel is $\mu < \mu_0$ ofwel is $\mu > \mu_0$. We bekijken de power van de verschillende toetsen voor beide situaties (deze eigenschappen zullen we niet bewijzen).



Figuur 7.14: Visualisatie van de eenzijdige beslissingsregel bij $H_a : \mu < \mu_0$.

- In werkelijkheid is $\mu < \mu_0$.
  - De tweezijdige toets met $H_a : \mu \neq \mu_0$ zal een specifiek alternatief kunnen detecteren met een bepaalde power.
  - De linkszijdige toets met $H_a : \mu < \mu_0$ zal ook een specifiek alternatief kunnen detecteren met een *hogere* power dan de tweezijdige toets.
  - De rechtszijdige toets met $H_a : \mu > \mu_0$ heeft een power van maximaal α (dus een zeer lage power).
- In werkelijkheid is $\mu > \mu_0$.
  - De tweezijdige toets met $H_a : \mu \neq \mu_0$ zal een specifiek alternatief kunnen detecteren met een bepaalde power.
  - De linkszijdige toets met $H_a : \mu < \mu_0$ heeft een power van maximaal α (dus een zeer lage power).
  - De rechtszijdige toets met $H_a : \mu > \mu_0$ zal ook een specifiek alternatief kunnen detecteren met een *hogere* power dan de tweezijdige toets.

In de praktijk wordt vaak gekozen voor de tweezijdige toets omdat ze voor zowel $\mu < \mu_0$ als $\mu > \mu_0$ power zal hebben. Als men anderzijds *op voorhand* weet dat $\mu \leq \mu_0$ of men is enkel geïnteresseerd in de alternatieve $H_a : \mu < \mu_0$, kiest men best voor een linkzijdige toets omdat deze een hogere power zal hebben. Analoog, als men anderzijds *op voorhand* weet dat $\mu \geq \mu_0$ of men is enkel geïnteresseerd in de alternatieve $H_a : \mu > \mu_0$, kiest men best voor een rechtszijdige toets omdat deze een hogere power zal hebben. Bij de eenzijdige toetsen bestaat echter de mogelijkheid dat je toets quasi geen power zal hebben: dit komt voor wanneer de richting van je alternatieve hypothese niet overeenstemt met de werkelijkheid.

Tabel 7.6 geeft een overzicht van de verschillende stappen die je moet volgen bij de t-toets voor één steekproef. De keuze om eenzijdig of tweezijdig te toetsen zal je tijdens de oefeningen kunnen afleiden uit de opgave.

7.3.6 p-waarde

In de voorgaande paragrafen hebben we gezien hoe we op basis van beslissingsregels een uitspraak kunnen doen over H_0 (al dan niet verwerpen). Deze beslissingsregels maken gebruik van kritische waarden die we kunnen opzoeken in de tabel of kunnen berekenen via R. Anderzijds kunnen we ook gebruik maken van de beslissingsregels op basis van

$H_0 : \mu = \mu_0$		
Kies de alternatieve hypothese H_a		
Bepaal het significantieniveau α		
Bereken de toetsingsgrootheid g		
Besluit op basis van de gekozen H_a:		
Indien linkszijdig	Indien rechtszijdig	Indien tweezijdig
$H_a : \mu < \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu \neq \mu_0$
Verwerp H_0 als	Verwerp H_0 als	Verwerp H_0 als
$g < -t_{n-1;\alpha}$	$g > t_{n-1;\alpha}$	$\lvert g \rvert > t_{n-1;\alpha/2}$

Tabel 7.6: Overzicht te volgen stappen bij de t-toets voor één steekproef.

een betrouwbaarheidsinterval. Er bestaat ook nog een derde mogelijkheid om tot een besluit te komen: via de *p-waarde* (ook de *overschrijdingskans* genoemd).

Niettegenstaande beslissingsregels op basis van de p-waarde eenvoudig zijn, is de p-waarde echter moeilijk te berekenen en te interpreteren. We starten met het gebruik van de beslissingsregels, waarbij we de p-waarde aanduiden met p:

- als $p \geq \alpha$ verwerpen we H_0 niet.
- als $p < \alpha$ verwerpen we H_0 en besluiten we H_a.

De p-waarde moeten we enkel vergelijken met het significantieniveau α om tot een besluit te komen.

Wat is deze p-waarde precies en hoe kunnen we ze berekenen? Formeel kunnen we de p-waarde als volgt omschrijven:

De p-waarde is de kans om een toetsingsgrootheid te observeren die minstens even extreem[b] *is als deze die waargenomen is, berekend in de veronderstelling dat de nulhypothese waar is.*

Deze omschrijving is vrij abstract, maar omvat volgende informatie:

- de p-waarde is een *kans*. Ze kan bijgevolg nooit kleiner zijn dan 0 en nooit groter zijn dan 1 (of 100 als je ze uitdrukt in percent).

[b] Het woord 'extreem' duidt op de richting waarvoor de teststatistiek onder de alternatieve hypothese meer waarschijnlijk is.

- de p-waarde wordt berekend in de veronderstelling dat H_0 waar is.
- de p-waarde hangt af van de alternatieve hypothese.

Voor de 3 soorten alternatieve hypotheses bespreken we de p-waarde.

Linkszijdige alternatieve hypothese $H_a : \mu < \mu_0$

Figuur 7.11 op pagina 229 illustreert welke waarden G kan aannemen wanneer $\mu < \mu_0$ (voor deze figuur is $\mu = 105$ en $\mu_0 = 115$): de toetsingsgrootheid neemt waarden links van nul aan (dus negatieve waarden). Meer extreem in de richting waarvoor de toetsingsgrootheid onder de alternatieve hypothese meer waarschijnlijk is, duidt hier dus op waarden links van de geobserveerde toetsingsgrootheid (dus kleinere waarden). Bijgevolg wordt de p-waarde gegeven door de kans:

$$P(G < g \mid \mu = \mu_0).$$

We illustreren dit aan de hand van het voorbeeld rond de IQ-scores en $H_a : \mu < 115$. De geobserveerde toetsingsgrootheid was $g = -0.98$ (zie pagina 225) en we weten dat G onder de nulhypothese een t_{n-1}-verdeling volgt. Bijgevolg is de p-waarde gelijk aan de kans

$$P(T < -0.98), \quad T \sim t_{n-1}.$$

Figuur 7.15 geeft deze kans grafisch weer. We kunnen gebruik maken van R om deze kans te berekenen:

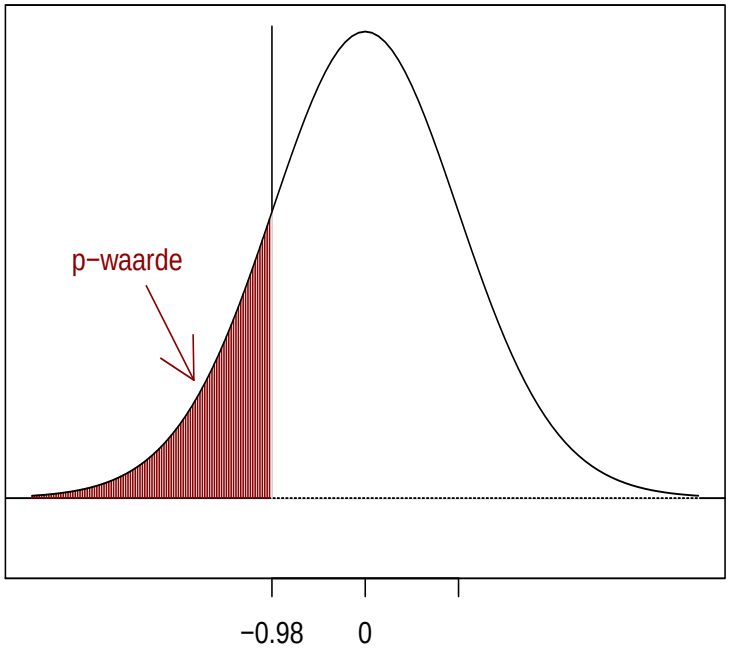
```
> pt(-0.98, 30-1)

[1] 0.1675957
```

De p-waarde is dus afgerond gelijk aan $p = 0.17$.

Rechtszijdige alternatieve hypothese $H_a : \mu > \mu_0$

Figuur 7.10 op pagina 228 illustreert welke waarden G kan aannemen wanneer $\mu > \mu_0$ (voor deze figuur is $\mu = 125$ en $\mu_0 = 115$): de toetsingsgrootheid neemt waarden rechts van nul aan (dus positieve waarden). Meer extreem in de richting waarvoor de toetsingsgrootheid onder de alternatieve hypothese meer waarschijnlijk is, duidt hier dus op waarden rechts van de geobserveerde toetsingsgrootheid (dus grotere waarden). De p-waarde wordt gegeven door de kans:

$$P(G > g \mid \mu = \mu_0),$$



Figuur 7.15: Visualisatie van de p-waarde horende bij de linkszijdige alternatieve hypothese. De p-waarde komt overeen met het rode gearceerde oppervlakte.

wat in het voorbeeld overeenkomt met

$$P(T > -0.98), \quad T \sim t_{n-1}.$$

Figuur 7.16 geeft deze kans grafisch weer. In R wordt dit:

```
> pt(-0.98, 30-1, lower.tail = FALSE)
```

```
[1] 0.8324043
```

waarbij we de optie `lower.tail = FALSE` hebben gebruikt om aan te geven dat het de oppervlakte rechts is die we willen berekenen. Anderzijds kan je ook beroep doen op formule (5.9) om deze kans te berekenen:

```
> 1 - pt(-0.98, 30-1)
```

```
[1] 0.8324043
```

De p-waarde is dus afgerond gelijk aan $p = 0.83$.

Tweezijdige alternatieve hypothese $H_a : \mu \neq \mu_0$

De tweezijdige alternatieve hypothese laat zowel $\mu < \mu_0$ als $\mu > \mu_0$ toe. Het berekenen van de p-waarde hangt af van het teken van g:

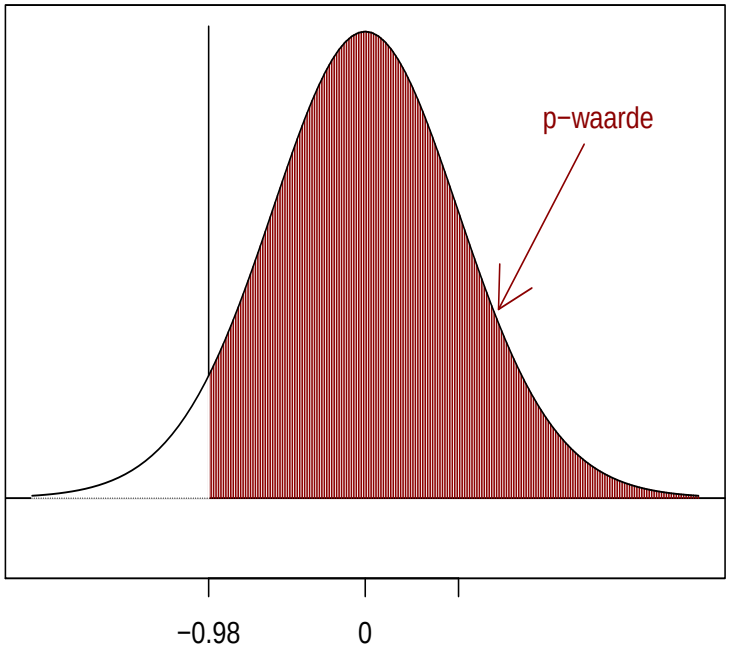
- als $g > 0$ dan is de p-waarde gelijk aan $2 \times P(G > g \mid \mu = \mu_0)$.
- als $g \leq 0$ dan is de p-waarde gelijk aan $2 \times P(G < g \mid \mu = \mu_0)$.

In het voorbeeld is $g = -0.98$, dus negatief en wordt de p-waarde gegeven door

$$2 \times P(T < -0.98), \quad T \sim t_{n-1}.$$

De p-waarde is twee maal het rode gearceerde gedeelte van Figuur 7.15. Omdat de t-verdeling symmetrisch is rond nul, is dit gelijk aan het gearceerde gebied in Figuur 7.17; ze geeft visueel weer dat dit inderdaad een tweezijdige p-waarde is.

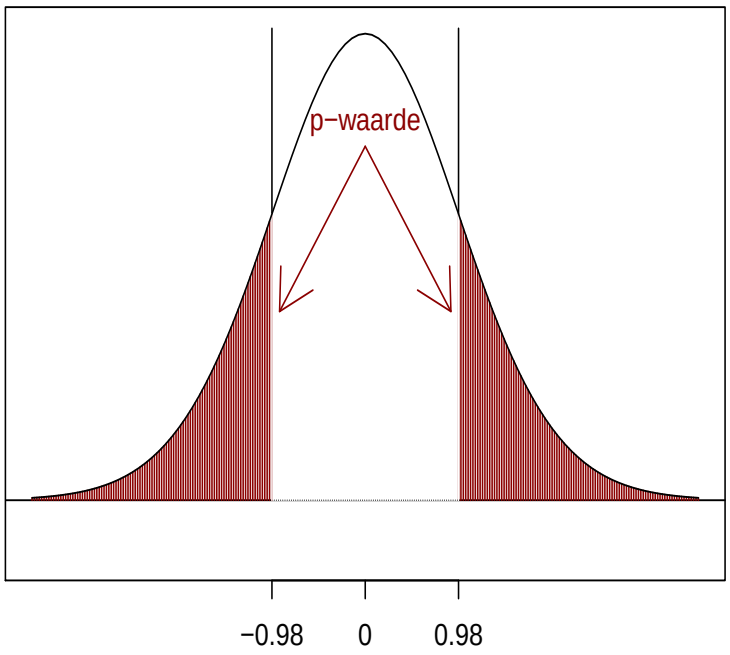
In R berekenen we dit als volgt:



Figuur 7.16: Visualisatie van de p-waarde horende bij de rechtszijdige alternatieve hypothese. De p-waarde komt overeen met het rode gearceerde oppervlakte.

```
> 2*pt(-0.98, 30-1)

[1] 0.3351915
```

De p-waarde is afgerond gelijk aan $p = 0.34$.



Figuur 7.17: Visualisatie van de p-waarde horende bij de tweezijdige alternatieve hypothese. De p-waarde komt overeen met de som van de rode gearceerde oppervlaktes.

Interpretatie van de p-waarde

De p-waarde is een kans die we kunnen interpreteren via de herhaalde steekproeftrekking. We illustreren de interpretatie aan de hand van de linkszijdige toets met $H_a : \mu < 115$ waarvoor $g = -0.98$ en de p-waarde gelijk is aan $p = 0.17$. Stel dat in werkelijkheid $\mu = 115$ (dus H_0 is correct) en we herhalen het experiment vele malen op basis van nieuwe steekproeven (theoretisch gezien herhalen we het experiment een

oneindig aantal keer) en per steekproef berekenen we de toetsingsgrootheid g. Een p-waarde van 0.17 drukt uit dat 17% van die toetsingsgrootheden kleiner zullen zijn dan -0.98.

Deze interpretatie geeft aan dat hoe kleiner de p-waarde is, hoe meer bewijskracht we tegen de nulhypothese hebben: de p-waarde meet de bewijskracht tegen H_0 in de richting van H_a.

Illustratie in R

De linkszijdige toets met $H_a : \mu < 115$ kan bekomen worden met de optie `alternative = "less"` (indien we deze optie niet aangeven, zal R automatisch de tweezijdige toets uitvoeren):

```
> t.test(IQ,  mu = 115, alternative = "less")


        One Sample t-test

data:  IQ
t = -0.9709, df = 29, p-value = 0.1698
alternative hypothesis: true mean is less than 115
95 percent confidence interval:
     -Inf 116.7001
sample estimates:
mean of x
 112.7333
```

De bovengrens van het linkszijdig betrouwbaarheidsinterval kunnen we narekenen ter controle:

```
> mean(IQ) + qt(0.95, 29)*sd(IQ)/sqrt(30)


[1] 116.7001
```

De rechtszijdige toets $H_a : \mu > 115$ kan bekomen worden met de optie `alternative = "greater"`

```
> t.test(IQ,  mu = 115, alternative = "greater")


        One Sample t-test

data:  IQ
t = -0.9709, df = 29, p-value = 0.8302
alternative hypothesis: true mean is greater than 115
95 percent confidence interval:
 108.7665      Inf
sample estimates:
mean of x
 112.7333
```

De ondergrens van het rechtzijdig betrouwbaarheidsinterval bekomen we via

```
> mean(IQ) - qt(0.95, 29)*sd(IQ)/sqrt(30)


[1] 108.7665
```

Merk op dat de toetsingsgrootheid voor de eenzijdige en tweezijdige toetsen dezelfde is. De p-waarde en de betrouwbaarheidsintervallen zijn echter verschillend omdat ze afhangen van H_a.

7.3.7 Overzicht en opmerkingen

Het toetsen van hypotheses via statistische toetsen zoals beschreven in de hoofdstuk wordt in het Engels *Null Hypothesis Significance Testing (NHST)* genoemd. Ze vormt een combinatie van de theorie ontwikkeld door Ronald Fisher (1890-1962) enerzijds en Jerzy Neyman (1894-1981) en Egon Pearson (1895-1980) anderzijds. De tabel onderaan geeft een schematisch overzicht van de te volgen stappen bij de toetsingsprocedure.

Overzicht te volgen stappen toetsingsprocedure

1. Formuleer $H_0 : \mu = \mu_0$ en kies een alternatieve hypothese H_a (linkszijdig, rechtszijdig of tweezijdig).
2. Leg het significantieniveau α vast.
3. Bereken de toetsingsgrootheid g.
4. Formuleer een beslissing (H_0 *niet verwerpen* of H_0 *verwerpen en* H_a *besluiten*)
   (a) met behulp van de kritieke waarden
   (b) met behulp van de p-waarde
   (c) met behulp van het betrouwbaarheidsinterval

In deze paragraaf geven we ook enkele belangrijke opmerkingen/misvattingen rond het gebruik van statistische toetsen[c].

- De logica van NHST is tot op zeker hoogte gelijkaardig met diegene in een rechtbank: men vertrekt vanuit de assumptie dat een beklaagde onschuldig is (nulhypothese) totdat voldoende bewijs van schuld kan worden aangetoond; de uitkomst is ofwel schuld, ofwel onschuld. Bemerk dat indien iemand onschuldig wordt bevonden dit niet betekent dat het bewezen werd dat de persoon inderdaad onschuldig is; er is enkel onvoldoende evidentie gevonden om schuld aan te tonen[d].
- De p-waarde die we bekomen op basis van de geobserveerde data kunnen we symbolisch noteren als $P(Data \mid H_0)$: de kans om een bepaalde waarde voor een steekproefgrootheid G te observeren, onder de assumptie dat de nulhypothese waar is. Echter, deze p-waarde bevat geen enkele informatie over $P(Data \mid H_a)$: de kans dat we G observeren mocht H_a waar zijn.
- Het strikte onderscheid tussen H_0 niet dan wel verwerpen is misschien te scherp. Wat indien $p = 0.049$ of $p = 0.051$ of zelfs $p = 0.045$ of $p = 0.06$ als $\alpha = 0.05$? Sommige onderzoekers spreken in dit verband over een 'trend-effect' of in het Engels van een 'marginally (in)significant effect'. Het geniet in ieder geval de

[c]De meeste van deze opmerkingen zijn afkomstig uit de cursus Statistiek II Academiejaar 2013-2014.

[d]Merk ook op dat de overeenkomst met de rechtbank hier ophoudt: bij een rechtszaak zal men vaak op een *subjectieve* wijze data verzamelen, terwijl we binnen empirische onderzoek juist op een *objectieve* wijze data trachten te verzamelen om de hypotheses te toetsen.

voorkeur om de exacte p-waarde altijd te vermelden (in plaats van bijvoorbeeld te noteren $p < 0.05$). De lezer kan dan zelf zijn/haar conclusies trekken of de evidentie tegen H_0 (zoals gereflecteerd in de p-waarde) overtuigend is of niet. Zeer kleine p-waarden kunnen worden genoteerd als $p < 0.0001$.

- De p-waarden worden bekomen op basis van theoretische verdelingen die enkel geldig zijn wanneer aan alle assumpties is voldaan (bv. variabele is normaal verdeeld of de steekproef is groot genoeg). Het verdient aanbeveling deze assumpties te controleren vooraleer men van start gaat (dit wordt gezien in de cursus 'Statistiek II'). Indien blijkt dat de assumpties niet gelden dienen alternatieve methodes te worden gebruikt. Dergelijke alternatieve methodes worden echter niet gezien in deze cursus.

Misvattingen rond de p-waarde

- *"De p-waarde is de kans dat H_0 waar is en $1-p$ is de kans dat H_a waar is."* Tweemaal fout. De p-waarde is de overschrijdingskans die we bekomen op basis van de geobserveerde data (i.e., de waarde g) *op voorwaarde dat de nulhypothese waar is*. Of nog, de p-waarde is een conditionele kans gebaseerd op de geobserveerde data $P(Data \mid H_0)$ wat niet hetzelfde is als $P(H_0 \mid Data)$.

- *"In het algemeen: hoe kleiner de p-waarde, hoe groter het verschil tussen μ en μ_0."* Fout. Dit geldt enkel indien de steekproefgrootte en variabiliteit constant blijven. Echter met een grote steekproef en weinig variabiliteit kan zelfs een klein verschil een kleine p-waarde opleveren.

- *"Een statistisch significant verschil tussen μ en μ_0 is voor de theorie of voor de praktijk ook significant."* Niet noodzakelijk. Met een grote steekproef bijvoorbeeld is een klein (onbenullig) verschil ook significant, doch niet van enige praktische waarde. Bemerk dat ook het omgekeerde niet geldt: *"indien er geen significant verschil gevonden wordt is er ook geen theoretisch of praktisch verschil."* Niet noodzakelijk. Misschien is de steekproef te klein en is de power daardoor laag.

- *"Als ik geen significant verschil vind, is mijn onderzoek nutteloos."* Fout. Het niet vinden van een significant verschil kan bijzonder informatief zijn op voorwaarde dat een gepast onderzoeksopzet werd gehanteerd wat vervolgens correct werd uitgevoerd en geanalyseerd.

7.4 Samenvatting

In dit hoofdstuk hebben we statistische methodes gebruikt om op basis van een steekproef een besluit te formuleren over de populatie. We kunnen een betrouwbaarheidsinterval opstellen en een statistische toets uitvoeren. Indien we een besluit formuleren over de populatie is er altijd een kans dat we een foute beslissing nemen. Deze kans kunnen we echter controleren, meten en communiceren.

In dit hoofdstuk hebben we voornamelijk een uitspraak willen doen over het populatiegemiddelde op basis van een aselecte steekproef. Er bestaan ook betrouwbaarheidsintervallen en statistische toetsen voor andere populatieparameters (vb. de variantie, de mediaan en de correlatiecoëfficiënt) alsook voor verschillende types van steekproeftrekkingen (verschillend van de aselecte steekproeftrekking). Dit zal worden behandeld in vervolgcursussen.

Bibliografie

Buchanan, M. (2007). Statistics: Conviction by numbers. *Nature*, 445(7125):254–255.

Davidian, M. and Louis, T. A. (2012). Why statistics? *Science*, 336(6077):12–12.

De Houwer, J. (2006). What are implicit measures and why are we using them. *The handbook of implicit cognition and addiction*, pages 11–28.

Meester, R., Collins, M., Gill, R., and Van Lambalgen, M. (2006). On the (ab) use of statistics in the legal case against the nurse lucia de b. *Law, Probability & Risk*, 5(3-4):233–250.

Willerman, L., Schultz, R., Rutledge, J. N., and Bigler, E. D. (1991). In vivo brain size and intelligence. *Intelligence*, 15(2):223–228.

Enkele vrijblijvende referenties naar handboeken statistiek:

Moore D.S. and McCabe G.P. (2006) *Statistiek in de Praktijk*. Academic Service.

Ellis, J. L. (2003). *Statistiek voor de Psychologie*. Boom.

Buhrman, J. M. (1996). *Basisboek Statistiek*. Wolters-Noordhoff.

Een handboek om je voorkennis wiskunde op te frissen:

Flohr, R. (2007). *Basiswiskunde voor Statistiek*, Academic Service.