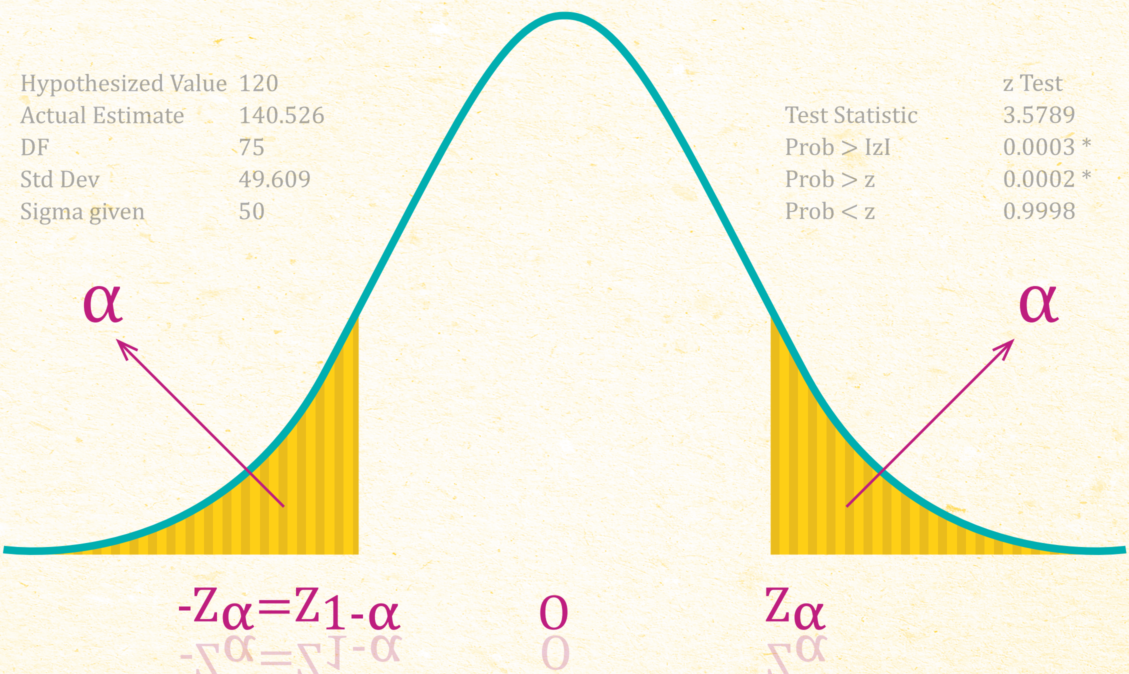
PETER GOOS

VERKLARENDE STATISTIEK

Schatten en toetsen



acco

Verklarende Statistiek
Schatten en Toetsen

Peter Goos

acco
Leuven/Den Haag

Eerste druk: 2014

Gepubliceerd door
Uitgeverij Acco, Blijde Inkomststraat 22, 3000 Leuven, België
E-mail: uitgeverij@acco.be
Website: www.uitgeverijacco.be

Voor Nederland:
Acco Nederland, Westvlietweg 67 F, 2495 AA Den Haag, Nederland
E-mail: info@uitgeverijacco.nl
Website: www.uitgeverijacco.nl

Omslagontwerp: www.frisco-ontwerpbureau.be

D/2014/0543/2 NUR 785 ISBN 978-90-334-9533-5

Voorwoord

Dit boek is het resultaat van een grondige herwerking van de cursusinhoud van het vak *Statistiek met (bedrijfs)economische toepassingen* 2 voor studenten handelsingenieur en toegepaste economische wetenschappen aan de Universiteit Antwerpen in het academiejaar 2012–2013. In het vak wordt het onderwerp van de verklarende statistiek behandeld en wordt de aandacht dus toegespitst op puntschatters, intervalschatters en hypothesetoetsen. In vergelijking met andere basishandboeken over verklarende statistiek komen een aantal extra onderwerpen aan bod. Er wordt bijvoorbeeld ruime aandacht besteed aan niet-parametrische toetsen, zoals de tekentoets, de rangtekentoets en de Kruskal-Wallis-toets. Daarnaast worden toetsen en betrouwbaarheidsintervallen besproken voor de correlatiecoëfficiënten van Pearson en Spearman, wordt het concept van equivalentietoetsen geïntroduceerd en is er een hoofdstuk over de beginselen van het optimaal ontwerp van experimenten of optimale statistische proefopzet. Bij de niet-parametrische toetsen worden exacte p-waarden uitvoerig belicht, naast de bekendere benaderende p-waarden.

Het boek besteed op een evenwichtige manier aandacht aan de wiskundige aspecten, de interpretatie van alle statistische concepten en het gebruik van schatters en toetsen in de praktijk. Om het begrijpen van alle nieuwe methodes en technieken te vergemakkelijken en het nut ervan in te zien, is het boek doorspekt van voorbeelden. Het brede toepassingsgebied van de verklarende statistiek wordt in de verf gezet door middel van voorbeelden uit tal van toepassingsgebieden: bedrijfskunde, economie, sport, ingenieurswetenschappen en natuurwetenschappen.

Er waren twee belangrijke drijfveren voor het schrijven van dit boek. Eerst en vooral wilde ik een boek schrijven dat verder gaat dan veel bestaand lesmateriaal wat betreft de technische en wiskundige achtergrond. Er is immers een toenemende tendens om technische aspecten en wiskundige afleidingen onder de mat te vegen.

Met dit boek roei ik dus tegen de stroom in. Ik ben er immers van overtuigd dat studenten aangemoedigd moeten worden om hun wiskundige kennis toe te passen en dat dit hen helpt om de statistische concepten beter te begrijpen. Het gebruik van dit boek vergt hierdoor uiteraard enige wiskundige voorkennis. Gezien de aandacht die aan wiskunde besteed wordt in het secundair onderwijs en in inleidende vakken wiskunde aan de universiteit zou die voorkennis bij elke student handelsingenieur, bio-ingenieur, toegepaste economische wetenschappen, burgerlijk en industrieel ingenieur aanwezig moeten zijn.

Een tweede drijfveer voor het schrijven van het boek is dat ik ervan overtuigd ben dat studenten moeten leren om de statistische concepten in de praktijk om te zetten. Daarom wordt in dit boek uitvoerig beschreven hoe alle hypothesetoetsen uitgevoerd kunnen worden en alle populatieparameters geschat kunnen worden met behulp van een gebruiksvriendelijk statistisch pakket, namelijk JMP (spreek uit als *jump*). Ik heb gekozen voor JMP omdat het een krachtig softwarepakket is dat toch gebruiksvriendelijk is, dat beschikbaar is voor Windows en Mac, en dat voor een brede waaier aan statistisch getinte vakken gebruikt kan worden (beschrijvende statistiek, kansrekenen, regressie, variantieanalyse, multivariate statistiek, statistische proefopzet of experimenteel ontwerp, ...). Ik ben van mening dat inleidende cursussen statistiek zoveel mogelijk gebruik dienen te maken van gebruiksvriendelijke software. Mits demonstraties in hoorcolleges en het gebruik ervan in werkcolleges kan een dergelijke software studenten enthousiasmeren voor statistiek, terwijl een ongebruiksvriendelijk pakket als R elk potentieel enthousiasme voor statistiek met grote waarschijnlijkheid in de kiem smoort. De kans dat een student in zijn latere beroepscarrière statistiek gebruikt zal veel groter zijn wanneer het gebruik van statistiek tijdens zijn opleiding een plezier was dan wanneer het een moeizaam traject was.

Dit boek bouwt voort op het boek *Beschrijvende Statistiek en Kansrekenen*, dat in 2013 verscheen bij ACCO. Het boek *Beschrijvende Statistiek en Kansrekenen* volgt dezelfde filosofie als het huidige boek en gebruikt dezelfde software. De enorme grafische mogelijkheden van het pakket JMP worden in *Beschrijvende Statistiek en Kansrekenen* uitvoerig belicht.

Aan veel onderwijsinstellingen is JMP vrij beschikbaar voor lesgevers en studenten, zowel voor thuisgebruik als in pc-klassen. Te allen tijde kan ook een gratis proefversie van JMP gedownload worden van www.jmp.com. U moet hiertoe wel eerst een korte registratieprocedure doorlopen. De proefversie werkt gedurende 30 dagen. Hoe gebruiksvriendelijk JMP ook is, het gebruik ervan zal enige oefening vergen. Om vertrouwd te raken met het softwarepakket kunt u enkele *webcasts* bekijken op www.jmp.com. De belangrijkste *webcasts* zijn misschien wel de *on-demand webcasts* met als titel “JMP for Students 1: Navigation and Use” en “JMP for Students 2: Basic Statistics”. U kunt deze *webcast* zien door te klikken op “News and Events”, vervolgens op “On-Demand Webcasts” en ten slotte op het tabblad “Academic” te klikken. Voor docenten zijn er ook enkele interessante *webcasts* getiteld “JMP for Professors: Tips for Teaching”' en “JMP for Professors: Resources for Teaching JMP”. Ook de *webcast* “Getting Started with JMP” is het bekijken waard.

Bij dit boek horen diverse datasets en JMP scripts. Deze datasets en scripts zijn beschikbaar via mijn webstek en via de website van uitgever ACCO. Docenten die het boek gebruiken kunnen via uitgever ACCO de LaTeX slides die bij het boek horen verkrijgen.

Ik heb tijdens het schrijven van dit boek tal van bronnen geraadpleegd. Ik wil in dit voorwoord graag de meest belangrijke vermelden. Een bron over niet-parametrische technieken die ik extreem waardevol vond, gezien zijn diepgang en bevattelijke uitleg, is het boek *Nonparametric Statistical Methods* van Hollander, Wolfe en Chicken. Een algemener boek dat ik zeer nuttig vond is *Handbook of Parametric and Nonparametric Statistical Procedures* van Sheskin. Hetzelfde geldt voor *Biostatistical Analysis* van Zar.

Ten slotte wil ik bij de uitgave van dit boek graag Kris Annaert, Stefan Becuwe, Marco Castro, Filip De Baerdemaeker, Eva Engelen, Hajar Hamidouche, Roselinde Kessels, Ida Ruts, Bagus Sartono, Daniel Palhazi Cuervo, Evelien Stoffels, Anja Struyf, Yahri Tillemans, Ellen Vandervieren, Katrien Van Driessen, Diane Verbiest, Tom Vermeire en Sara Weyns bedanken voor hun gedetailleerde opmerkingen en constructieve suggesties, en voor de technische ondersteuning bij het creëren van figuren en tabellen. Tevens gaat mijn dank uit naar Kristel Van Rompay en Peter Willemé voor de creatie van enkele tabellen en figuren, en voor de inspiratie bij de uitwerking van enkele voorbeelden. Katrien Van Driessen, Ellen Vandervieren en Stefan Becuwe verdienen eveneens een speciaal bedankje voor hun werk aan de LaTeX slides. Ik wil ook Bradley Jones, Volker Kraft en Brady Brady van de JMP-divisie in het SAS Institute bedanken voor hun hulp met betrekking tot het softwarepakket JMP. Brady Brady en Bradley Jones ontwikkelden diverse JMP scripts speciaal voor dit boek. Ten slotte wil ik ook prof. dr. Willy Gochet van de KU Leuven bedanken voor de initiële inspiratie die hij bood bij de ontwikkeling van deze cursus.

Inhoudsopgave

Deel I

Schatters en toetsen

Hoofdstuk 1

Het schatten van populatieparameters

Hoe lang ik daar sta weet ik niet, of ik al een keer in een striemende regen trouw mijn rouw heb gedragen geloof ik niet, maar een motregentje, een saladevlaagje, moet er statistisch gezien wel een keer bij geweest zijn.
(uit De Helaasheid der Dingen, Dimitri Verhulst, p. 135)

1.1 Inleiding: schatter versus schatting

De populatieparameters μ, σ^2, π en λ uit het boek *Beschrijvende Statistiek en Kansrekenen* zijn in de praktijk zelden of nooit bekend. Wanneer u bijvoorbeeld het aankomstenpatroon van klanten in een bankkantoor bestudeert, dan weet u uit het boek *Beschrijvende Statistiek en Kansrekenen* dat het aantal aankomsten per tijdseenheid vaak Poisson[1]-verdeeld is. De precieze waarde van de parameter λ, die bij deze kansverdeling hoort, kent u dan natuurlijk niet. Op de een of andere manier zult u deze parameter dan proberen te schatten. Deze schatting zal dan wellicht gebaseerd zijn op een aantal metingen of waarnemingen $x_1, x_2, \ldots, x_n$ die u in de bank uitvoert, met andere woorden: op de steekproefgegevens die u verzamelt.

De **schatting** voor de onbekende λ zal een functie zijn van de verzamelde steekproefwaarden $x_1, x_2, \ldots, x_n$, bijvoorbeeld het steekproefgemiddelde $\overline{x}$. Het is natuurlijk zo dat elke onderzoeker die zich op hetzelfde probleem, het bestuderen van het

[1]De Poisson-verdeling wordt vaak gebruikt voor kansvariabelen die een aantal gebeurtenissen per tijdseenheid, per lengte-eenheid, per volume-eenheid, . . . weergeven. De Poisson-verdeling bezit één parameter, λ, die het gemiddeld aantal gebeurtenissen per tijdseenheid, per lengte-eenheid, per volume-eenheid, . . . weergeeft.

aankomstenpatroon van klanten, stort, andere steekproefwaarden verkrijgt, en bijgevolg ook een ander steekproefgemiddelde of een andere schatting. De reden hiervoor is dat het aantal aankomsten in het kantoor in een bepaald tijdsinterval een kansvariabele is. Dit kunnen we expliciet weergeven door hoofdletters te gebruiken voor de steekproefwaarnemingen $X_1, X_2, \ldots, X_n$. Het feit dat elke onderzoeker een andere schatting voor λ verkrijgt, kunnen we eveneens expliciteren door een hoofdletter te gebruiken om het steekproefgemiddelde aan te duiden: $\overline{X}$. Het steekproefgemiddelde wordt dan geïnterpreteerd als een kansvariabele en dan spreekt men van een **schatter** in plaats van een schatting. Kortom, een schatting is altijd een reëel getal terwijl een schatter een kansvariabele is waarvan de waarde nog niet bekend is.

Het steekproefgemiddelde is natuurlijk maar een van de vele mogelijke functies van de steekproefwaarnemingen $X_1, X_2, \ldots, X_n$ en dus maar een van de vele mogelijke schatters. Een onderzoeker is vanzelfsprekend niet geïnteresseerd in om het even welke functie van de steekproefwaarnemingen, maar hij/zij wil een goed idee krijgen van de onbekende parameter. De onderzoeker wil met andere woorden een schatting verkrijgen die gemiddeld gelijk is aan de onbekende parameter en liefst gegarandeerd dicht bij de onbekende parameter ligt. Statistici vertalen deze vereiste in "de schatter moet onvertekend zijn" en "de schatter moet een kleine variantie hebben". Deze vereisten worden in de volgende paragraaf duidelijk gemaakt.

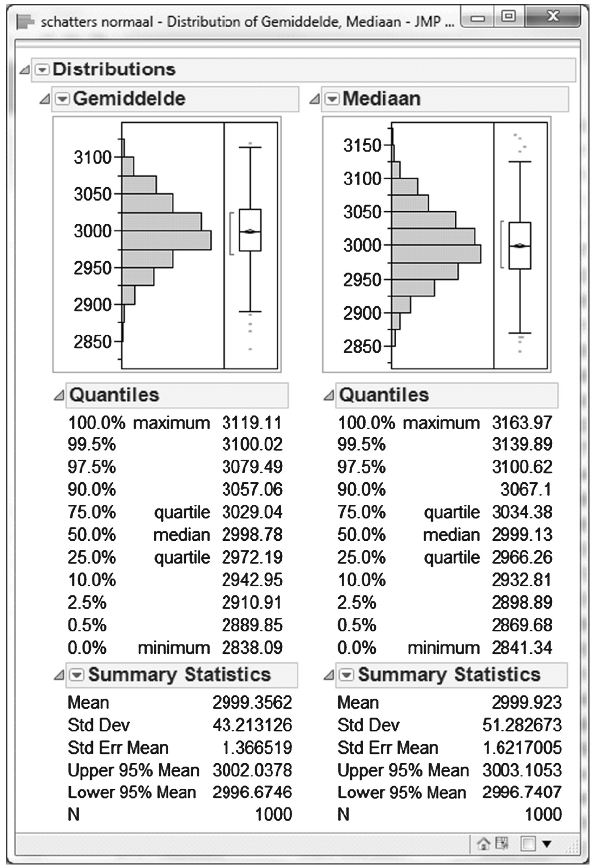
1.2 Het schatten van een gemiddelde

De eisen die aan een goede schatter gesteld worden, kunnen het beste geïllustreerd worden aan de hand van twee simulatiestudies. In een eerste simulatiestudie wordt nagebootst dat een normaal verdeelde populatie bestudeerd wordt. In de tweede studie wordt nagebootst dat een populatie met een exponentiële kansdichtheid bestudeerd wordt.

1.2.1 Gemiddelde van een normaal verdeelde populatie

We veronderstellen eerst dat een normaal verdeelde populatie met gemiddelde $\mu = 3000$ en standaarddeviatie $\sigma = 100$ bestudeerd wordt door 1000 (fictieve) studenten. Elk van deze studenten verricht 5 metingen. Een eerste optie voor deze studenten om de voor hen onbekende waarde μ te schatten is het steekproefgemiddelde te berekenen. Op die manier verkrijgen we 1000 steekproefgemiddeldes, waarvan een histogram is afgebeeld in Figuur 1.1 linksboven. Het gemiddelde van deze 1000 steekproefgemiddeldes bedraagt 2999.3562, terwijl de standaarddeviatie 43.2131 bedraagt.

Een andere mogelijkheid om de "onbekende" μ te schatten is om de mediaan te berekenen. Voor een normaal verdeelde populatie geldt dat zowel de mediaan als de verwachte waarde gelijk is aan de parameter μ, zodat dit niet eens zo gek is. Op



Figuur 1.1

Histogrammen en beschrijvende statistieken van 1000 steekproefgemiddeldes en -medianen berekend op basis van steekproeven van 5 waarnemingen uit een normaal verdeelde populatie met gemiddelde 3000 en standaarddeviatie 100.

basis van de steekproeven die de studenten verzamelden, kunnen dan 1000 medianen berekend worden, waarmee ook weer een histogram opgesteld kan worden. Dit histogram is te zien in Figuur 1.1 rechtsboven.[2] De aandachtige lezer zal onmiddellijk opmerken dat het nieuwe histogram een tikkeltje breder uitvalt dan het eerste. Dit wordt onder meer weerspiegeld in het feit dat de standaarddeviatie van de 1000 medianen 51.2827 bedraagt. Het gemiddelde van de 1000 medianen is gelijk aan 2999.9230. In Figuur 1.1 is ook te zien dat het minimum (2841.34) en het eerste kwartiel (2966.26) van de steekproefmedianen kleiner zijn dan het minimum (2838.09) en het eerste kwartiel (2972.19) van de steekproefgemiddeldes. Ook zijn het maximum (3163.97) en het derde kwartiel (3034.38) van de steekproefmedianen groter dan het maximum (3119.11) en het derde kwartiel (3029.04) van de steekproefgemiddeldes. Dit betekent dat de steekproefmedianen over het algemeen verder af liggen van het populatiegemiddelde $\mu = 3000$ dan de steekproefgemiddeldes.

[2]Outputs zoals deze in Figuren 1.1 en 1.2 kunnen gecreëerd worden door in JMP het menu "`Analyze`" te kiezen, en vervolgens voor "`Distribution`" te opteren.

Het is opvallend dat zowel het gemiddelde van de 1000 steekproefgemiddeldes (2999.3562) als dat van de 1000 medianen (2999.9230) bijzonder dicht bij 3000 ligt. Indien het aantal steekproeven gevoelig opgedreven zou worden (in theorie tot er oneindig veel steekproeven genomen zijn), dan zouden de gemiddeldes van de steekproefgemiddeldes en de steekproefmedianen gelijk worden aan de "onbekende" $\mu = 3000$. Daarom zegt men dat zowel het steekproefgemiddelde als de mediaan **zuivere schatters** of **onvertekende schatters** zijn van het gemiddelde van een normaal verdeelde populatie.

Het feit dat de spreidingsbreedte, de interkwartielbreedte, de standaarddeviatie, en dus ook de variantie, van de 1000 steekproefgemiddeldes kleiner zijn dan deze van de 1000 steekproefmedianen betekent eigenlijk dat het steekproefgemiddelde een betrouwbaardere schatter is voor het onbekende populatiegemiddelde dan de mediaan. De grotere variantie van de medianen geeft aan dat de medianen doorgaans verder af liggen van hun gemiddelde dan de steekproefgemiddeldes. Kortom, een onderzoeker zal meer vertrouwen hebben in het steekproefgemiddelde omdat dat doorgaans dichter bij de onbekende μ ligt. Men zegt in zo'n geval dat de ene schatter (hier het steekproefgemiddelde) een **efficiëntere** of een **preciezere** schatter is dan de andere (hier de mediaan).

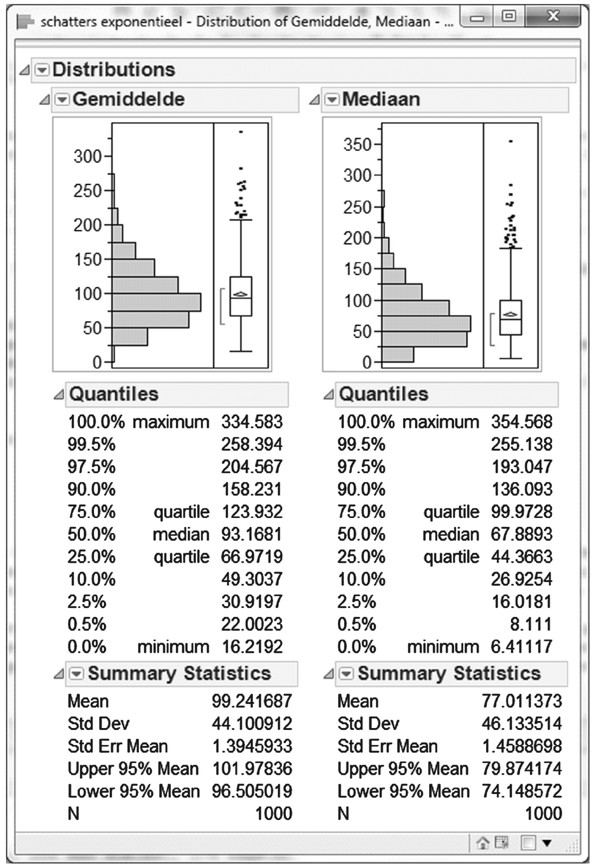
1.2.2 Gemiddelde van een exponentieel verdeelde populatie

Dit keer veronderstellen we dat een exponentieel verdeelde populatie met parameter $\lambda = 1/100$ wordt onderzocht. Het "onbekende" populatiegemiddelde bedraagt bijgevolg $\mu = 1/\lambda = 100$ (zie het boek *Beschrijvende Statistiek en Kansrekenen*). Elk van de fictieve studenten verricht vijf metingen. Een eerste optie voor deze studenten om de voor hen onbekende waarde μ te schatten is het steekproefgemiddelde te berekenen. Een histogram van de 1000 steekproefgemiddeldes is afgebeeld in Figuur 1.2 linksboven. Het gemiddelde van deze 1000 steekproefgemiddeldes bedraagt 99.2417, terwijl de standaarddeviatie 44.10 bedraagt.

Op basis van de steekproeven die de studenten verzamelden, kunnen opnieuw 1000 medianen berekend worden, waarmee ook weer een histogram opgesteld kan worden. Dit histogram is te zien in Figuur 1.2 rechtsboven. Het gemiddelde van de 1000 medianen is gelijk aan 77.0114.

Uit deze berekeningen blijkt dat het populatiegemiddelde $\mu = 1/\lambda = 100$ vrij goed benaderd wordt met behulp van de steekproefgemiddeldes, die gemiddeld 99.2417 bedragen. Dit is niet het geval voor de medianen, die gemiddeld verre van gelijk zijn aan μ. Dit zal nog steeds zo zijn als het aantal steekproeven verhoogd wordt. Men zegt daarom dat de mediaan in dit voorbeeld, bij een exponentieel verdeelde populatie, geen zuivere maar wel een **vertekende** schatter is van het populatiegemiddelde.

Figuur 1.2 toont verder ook nog dat de standaarddeviatie van de steekproefmedianen (46.13) groter is dan die van de steekproefgemiddeldes (44.10).



Figuur 1.2

Histogrammen en beschrijvende statistieken van 1000 steekproefgemiddeldes en -medianen berekend op basis van steekproeven van 5 waarnemingen uit een exponentieel verdeelde populatie met parameter $\lambda = 1/100$.

1.3 Criteria voor schatters

1.3.1 Een onvertekende of zuivere schatter

Een ideale schatter die gegarandeerd aangeeft wat de precieze waarde van een onbekende populatieparameter is, bestaat niet. Uit het voorgaande blijkt wel dat sommige schatters, namelijk zuivere of onvertekende schatters, gemiddeld gezien gelijk zijn aan de onbekende populatieparameters, terwijl andere systematisch een populatieparameter onder- of overschatten. Dit laatste is niet meteen wat een onderzoeker wenst. Formeel luidt de definitie van een onvertekende schatter $\hat{\theta}$ voor een onbekende populatieparameter θ:

Definitie 1.3.1 Een schatter $\hat{\theta}$ voor een populatieparameter θ is zuiver of onvertekend indien

$$E(\hat{\theta}) = \theta.$$

De **vertekening** van een schatter is het verschil $V(\hat{\theta}) = |E(\hat{\theta}) - \theta|$. Een zuivere schatter heeft een vertekening van nul. Bij een onvertekende schatter verwacht men, vooraleer de steekproefgegevens verzameld zijn, dat de uiteindelijke schatting precies gelijk zal zijn aan de gezochte populatieparameter. De histogrammen voor de steekproefgemiddeldes in de linkergedeeltes van Figuren 1.1 en 1.2 tonen dat, zodra de steekproefgegevens verzameld zijn, de schatting wel in de buurt van de onbekende populatieparameter zal liggen, maar er niet precies op.

Bemerk dat het symbool $\hat{\theta}$ hier wordt gebruikt om een schatter aan te duiden voor de onbekende populatieparameter θ. We gebruiken in deze cursus, zoals gebruikelijk is in de statistiek, een Griekse letter om een onbekende populatieparameter (zoals een populatiegemiddelde, een populatieproportie en een populatievariantie) aan te duiden. Wanneer we die onbekende populatieparameter willen schatten, dan gebruiken we een schatter, wat een synoniem is voor een schattingsmethode. Deze schatter duidt men in de statistiek over het algemeen aan met het symbool $\hat{\theta}$ (spreek uit als “thèta hoed”). In deze cursus spitsen we de aandacht hoofdzakelijk toe op drie specifieke schatters, namelijk het steekproefgemiddelde, de steekproefproportie en de steekproefvariantie. Voor deze drie schatters worden om historische redenen de symbolen $\overline{X}$, $\hat{P}$ en S^2 gebruikt in plaats van $\hat{\mu}$, $\hat{\pi}$ en $\hat{\sigma}^2$.

Het steekproefgemiddelde $\overline{X}$ is steevast een onvertekende of zuivere schatter van een populatiegemiddelde (een bewijs hiervan wordt in Stelling 1.1 gegeven). Dit geldt eigenlijk voor alle mogelijke lineaire functies $Y = \sum_{i=1}^{n} \alpha_i X_i$ van de steekproefwaarnemingen waarvoor $\sum_{i=1}^{n} \alpha_i = 1$, en het steekproefgemiddelde is een speciaal geval van zo’n lineaire combinatie waarbij elke $\alpha_i = 1/n$:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n.$$

Het kan aangetoond worden dat het steekproefgemiddelde van alle mogelijke lineaire functies van $X_1, X_2, \ldots, X_n$ de kleinste variantie[3] heeft. Met andere woorden, het steekproefgemiddelde zal de onderzoeker doorgaans een schatting opleveren die dichter bij het populatiegemiddelde ligt dan om het even welke andere lineaire functie Y van $X_1, X_2, \ldots, X_n$.

In Stelling 1.8 wordt bewezen dat de steekproefvariantie

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

een onvertekende schatter is van een populatievariantie σ^2. Deze stelling toont daarmee eigenlijk aan waarom er gedeeld wordt door $n - 1$ bij de berekening van de

[3] In de meeste handboeken wordt het steekproefgemiddelde om deze reden de beste lineaire onvertekende schatter genoemd. In het Engels spreekt men van “best linear unbiased estimator” of “BLUE”.

steekproefvariantie en niet door n. Belangrijk is op te merken dat de steekproefstandaarddeviatie S wel een vertekende schatter van de populatiestandaarddeviatie σ is.

Ten slotte blijkt uit Paragraaf 1.6 dat een steekproefproportie $\hat{P}$ een speciaal geval is van een steekproefgemiddelde. De verwachte waarde ervan is gelijk aan de populatieproportie π, zodat ook $\hat{P}$ een onvertekende schatter is.

1.3.2 *Precisie of efficiëntie van een schatter*

Het is wenselijk dat een schatter zo betrouwbaar mogelijk is en schattingen oplevert die doorgaans dicht bij de onbekende populatieparameter liggen. Kortom, de schatter moet een kleine variantie of standaarddeviatie bezitten. Een schatter met een kleine variantie noemt men een precieze of een efficiënte schatter.

Wanneer $\hat{\theta}_1$ en $\hat{\theta}_2$ twee onvertekende of zuivere schatters voor eenzelfde onbekende populatieparameter θ zijn, dan wordt de relatieve efficiëntie van $\hat{\theta}_2$ ten opzichte van $\hat{\theta}_1$ berekend als $\mathrm{var}(\hat{\theta}_1)/\mathrm{var}(\hat{\theta}_2)$.

Soms heeft een onderzoeker de keuze tussen een schatter die onvertekend is maar een grote variantie bezit en een schatter die vertekend is maar een kleine variantie bezit. In dit geval is het niet meteen duidelijk welke schatter het best gebruikt kan worden. Om een beslissing te nemen, kiest men in dergelijke situaties voor de schatter die de kleinste gemiddelde gekwadrateerde afwijking $\mathrm{GGA}(\hat{\theta})$ bezit:

Definitie 1.3.2 De gemiddelde gekwadrateerde afwijking van een schatter $\hat{\theta}$ is de som van zijn variantie en het kwadraat van zijn vertekening:

$$\mathrm{GGA}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + \left[V(\hat{\theta})\right]^2.$$

Ten slotte is het ook wenselijk dat de nauwkeurigheid of de precisie van een schatter toeneemt naarmate het aantal waarnemingen toeneemt. Meer waarnemingen leveren immers meer informatie, zodat betere schattingen verwacht kunnen worden. Stelling 1.2 geeft bijvoorbeeld aan dat de variantie van het steekproefgemiddelde gelijk is aan σ^2/n. De variantie neemt dus af naarmate de steekproefgrootte n toeneemt. De precisie van het steekproefgemiddelde neemt toe naarmate meer gegevens beschikbaar zijn.

1.4 Methoden voor het berekenen van schatters

Het vinden van schatters met goede eigenschappen is niet altijd even eenvoudig. In de statistische literatuur zijn er drie methoden die frequent gebruikt worden:

1. de methode van de momenten,
2. de methode van de kleinste kwadraten, en
3. de methode van de grootste aannemelijkheid (*maximum likelihood method*).

Deze methoden vallen buiten het bestek van dit boek maar komen aan bod in, bijvoorbeeld, boeken over econometrie, kwantitatieve beleidsmethoden en gevorderde statistiek. De schatters die in dit boek aan bod komen, zijn voornamelijk steekproefgemiddeldes, steekproefproporties en steekproefvarianties. In wat volgt, wordt van elk van deze schatters aangetoond dat ze onvertekend zijn. Daarnaast worden ook hun kansdichtheden afgeleid.

1.5 Het steekproefgemiddelde

1.5.1 Verwachte waarde en variantie

Indien het steekproefgemiddelde als een schatter beschouwd wordt en dus als een kansvariabele aangezien wordt, dan kunnen we de verwachte waarde, de variantie en zelfs de kansdichtheid ervan bepalen. Het steekproefgemiddelde wordt dan met behulp van een hoofdletter genoteerd als

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

om dit expliciet aan te geven. We beschouwen het steekproefgemiddelde als een kansvariabele of schatter zolang we geen data verzameld hebben, dus zolang de individuele waarnemingen $X_1, X_2, \ldots, X_n$ niet bekend zijn. Zodra de gegevens verzameld zijn, gebruiken we voor de individuele waarnemingen kleine letters: $x_1, x_2, \ldots, x_n$. Voor het steekproefgemiddelde dat we berekenen op basis van de waargenomen waarden $x_1, x_2, \ldots, x_n$ gebruiken we dan ook een kleine letter:

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

Stelling 1.1 *Voor een lukrake steekproef uit een populatie met verwachte waarde μ geldt dat*

$$E(\overline{X}) = \mu.$$

Het bewijs[4] hiervan verloopt als volgt:

$$\begin{aligned} E(\overline{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \end{aligned}$$

[4]Er dient voor de volledigheid opgemerkt te worden dat dit bewijs impliciet gebruik maakt van enkele eigenschappen van lukrake steekproeven met zogenaamde teruglegging. In dit boek wordt hier evenwel niet dieper op ingegaan.

$$= \frac{1}{n}(\mu + \mu + \cdots + \mu)$$
$$= \frac{n\mu}{n}$$
$$= \mu.$$

Deze stelling geeft aan dat, voordat de steekproefgegevens verzameld worden, de verwachte waarde van het steekproefgemiddelde gelijk is aan het populatiegemiddelde. Met andere woorden, deze stelling toont aan dat het steekproefgemiddelde een onvertekende of zuivere schatter is van het populatiegemiddelde.

Eens de steekproef uitgevoerd, vinden we voor de steekproefwaarnemingen $x_1, x_2, \ldots, x_n$ een steekproefgemiddelde $\overline{x}$. Dat steekproefgemiddelde zal natuurlijk niet precies gelijk zijn aan μ. Dit bleek ook al uit de berekeningen in Paragraaf 1.2, waar elke student een ander steekproefgemiddelde vond. Om een idee te krijgen van de grootte van de mogelijke afwijkingen tussen het steekproefgemiddelde $\overline{X}$ en het populatiegemiddelde μ is het interessant de variantie en de standaarddeviatie van $\overline{X}$ te bestuderen. In Figuur 1.1 konden we al zien dat de standaarddeviatie van de steekproefgemiddeldes van 1000 (fictieve) studenten gelijk was aan 43.21, terwijl de oorspronkelijke populatie een standaarddeviatie heeft van 100 en dus elke individuele waarneming van de studenten ook een standaarddeviatie heeft van 100. De standaarddeviatie van een steekproefgemiddelde ligt dus lager dan de standaarddeviatie van een individuele waarneming. Bijgevolg zal ook de variantie van een steekproefgemiddelde kleiner zijn dan de variantie van een individuele waarneming. Hoeveel kleiner de variantie van een steekproefgemiddelde is, wordt duidelijk in de volgende stelling:

Stelling 1.2 *Voor een lukrake steekproef van n waarnemingen uit een populatie met variantie* σ^2 *geldt dat*

$$\sigma^2_{\overline{X}} = \text{var}(\overline{X}) = \frac{\sigma^2}{n}$$

en

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Het bewijs omvat de volgende stappen:

$$\sigma^2_{\overline{X}} = \text{var}(\overline{X}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i)$$

$$= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \cdots + \sigma^2\right)$$
$$= \frac{n\sigma^2}{n^2}$$
$$= \frac{\sigma^2}{n}.$$

In de tweede stap van dit bewijs wordt verondersteld dat de covariantie tussen twee verschillende steekproefwaarnemingen X_i en X_j gelijk is aan nul. Dan is de variantie van een lineaire combinatie van kansvariabelen immers gelijk aan een lineaire combinatie van de varianties, met gekwadrateerde coëfficiënten.[5] Het bewijs geeft aan dat de variantie van het steekproefgemiddelde lineair afneemt wanneer de steekproefomvang n toeneemt. Dit betekent dat, naarmate een steekproef groter wordt, de kans groter wordt dat een steekproefgemiddelde $\overline{x}$ dicht bij (de onbekende) μ zal liggen.

De vierkantswortel van deze variantie, namelijk $\sigma_{\overline{X}}$, wordt **standaardfout** of **standard error** genoemd. De geschatte versie van dit kengetal, namelijk $s/\sqrt{n}$, is terug te vinden in de outputs van heel wat statistische pakketten. Figuren 1.1 en 1.2 illustreren dat de standaardfout of *standard error* ook gerapporteerd wordt door JMP,[6] namelijk onder "Std Err Mean". Het is niet moeilijk te verifiëren dat de standaardfout in deze figuren een factor $\sqrt{n} = \sqrt{1000} = 31.62$ kleiner is dan de overeenkomstige gerapporteerde standaarddeviatie (*standard deviation*), die afgekort wordt tot "Std Dev" in de JMP-outputs.

1.5.2 Kansdichtheid van het steekproefgemiddelde uit een normaal verdeelde populatie

In het geval waarbij de steekproef getrokken wordt uit een normaal verdeelde populatie, kunnen we van Stelling 1.3 gebruik maken.

[5]Uit het boek *Beschrijvende Statistiek en Kansrekenen* weten we dat

$$\text{var}(aX + bY) = a^2\,\text{var}(X) + b^2\,\text{var}(Y) + 2ab\;\text{cov}(X, Y),$$

wat vereenvoudigd kan worden tot

$$\text{var}(aX + bY) = a^2\,\text{var}(X) + b^2\,\text{var}(Y)$$

indien de kansvariabelen X en Y onafhankelijk of ongecorreleerd zijn (en dus $\text{cov}(X, Y) = 0$).

[6]Zoals eerder aangegeven, kan de standaardfout in JMP berekend worden via het menu "`Analyze`" en de optie "`Distribution`". Een alternatieve manier om de standaardfout te verkrijgen voor een bepaalde variabele in een gegevenstabel is door gebruik te maken van het menu "`Tables`", en vervolgens voor "`Summary`" te kiezen. Ten slotte dient u bij de knop "`Statistics`" voor "`Std Err`" te opteren.

Stelling 1.3 *Indien $X_1, X_2, \ldots, X_k$ onafhankelijke normaal verdeelde kansvariabelen zijn met verwachte waarden respectievelijk gelijk aan $E(X_1) = \mu_1, E(X_2) = \mu_2, \ldots,$ $E(X_k) = \mu_k$ en varianties respectievelijk gelijk aan $\mathrm{var}(X_1) = \sigma_1^2$, $\mathrm{var}(X_2) = \sigma_2^2$, $\ldots, \mathrm{var}(X_k) = \sigma_k^2$, dan is een lineaire functie $Y = \alpha_0 + \sum_{i=1}^{k} \alpha_i X_i$ eveneens normaal verdeeld met verwachte waarde $E(Y) = \alpha_0 + \sum_{i=1}^{k} \alpha_i \mu_i$ en variantie $\mathrm{var}(Y) = \sum_{i=1}^{k} \alpha_i^2 \sigma_i^2$.*

Uit deze stelling volgt dat het gemiddelde van een aantal normaal verdeelde kansvariabelen met eenzelfde variantie σ^2 eveneens normaal verdeeld is. Immers, het gemiddelde van een aantal variabelen $X_1, X_2, \ldots, X_n$ is een lineaire functie waarbij $\alpha_0 = 0$ en $\alpha_i = 1/n$. Bijgevolg hebben we in dit geval dat het steekproefgemiddelde $\overline{X}$ normaal verdeeld is met gemiddelde μ en variantie σ^2/n. In het vervolg van het boek zullen we dit aanduiden met de notatie

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Dit resultaat is geldig voor elke steekproefgrootte, hoe klein ook, en wordt geïllustreerd in Paragraaf 1.5.4.

1.5.3 *Kansverdeling of -dichtheid van het steekproefgemiddelde uit een niet-normaal verdeelde populatie*

In het geval waarbij de populatie een onbekende kansdichtheid of kansverdeling bezit, kan de kansverdeling van het steekproefgemiddelde in de meeste gevallen niet exact worden bepaald. In dit geval kan een grote steekproef evenwel redding brengen omdat, voor grote steekproeven, de centrale limietstelling gebruikt kan worden. Eén versie van deze stelling, namelijk Stelling 1.6, geeft immers aan dat het steekproefgemiddelde bij grote n benaderend normaal verdeeld is met gemiddelde μ en variantie σ^2/n.

1.5.3.1 Centrale limietstelling

De zogenaamde **centrale limietstelling** of *central limit theorem* is een van de belangrijkste stellingen uit de statistiek. Deze stelling verklaart ook voor een groot gedeelte waarom de normale kansdichtheid van cruciaal belang is in de statistiek. Van de stelling bestaan verschillende versies.

Stelling 1.4 *Indien $X_1, X_2, \ldots, X_n$ onafhankelijke kansvariabelen zijn met verwachte waarden $E(X_i) = \mu_i$ en varianties $var(X_i) = \sigma_i^2$, dan geldt onder heel algemene voorwaarden en voor een voldoende grote waarde n dat*

1. *de nieuwe kansvariabele $Y = \sum_{i=1}^{n} X_i$ benaderend normaal verdeeld is met verwachte waarde $\mu_Y = \sum_{i=1}^{n} \mu_i$ en variantie $\sigma_Y^2 = \mathrm{var}(Y) = \sum_{i=1}^{n} \sigma_i^2$,*

2. *en bijgevolg dat de nieuwe kansvariabele*

$$\frac{Y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$$

benaderend standaardnormaal verdeeld is.

De algemene voorwaarden waarvan sprake is in de stelling slaan op het feit dat geen van de individuele varianties σ_i^2 een dominante bijdrage levert tot de totale variantie van Y. In veel praktische toepassingen van de centrale limietstelling hebben alle kansvariabelen $X_1, X_2, \ldots, X_n$ evenwel dezelfde verdeling of dichtheid en bijgevolg dezelfde σ_i^2, zodat aan deze voorwaarde voldaan is. Indien alle kansvariabelen $X_1, X_2, \ldots, X_n$ inderdaad dezelfde verdeling of dichtheid bezitten, dan kan de centrale limietstelling als volgt herschreven worden:

Stelling 1.5 *Indien $X_1, X_2, \ldots, X_n$ onafhankelijke kansvariabelen zijn met verwachte waarde $E(X_i) = \mu$ en variantie $\mathrm{var}(X_i) = \sigma^2$, dan geldt voor een voldoende grote waarde n dat*

1. *de nieuwe kansvariabele $Y = \sum_{i=1}^{n} X_i$ benaderend normaal verdeeld is met verwachte waarde $\mu_Y = n\mu$ en variantie $\sigma_Y^2 = \mathrm{var}(Y) = n\sigma^2$,*
2. *en bijgevolg dat de nieuwe kansvariabele*

$$\frac{Y - n\mu}{\sigma\sqrt{n}}$$

benaderend standaardnormaal verdeeld is.

De centrale limietstelling kan ook geformuleerd worden in termen van het steekproefgemiddelde $\overline{X} = Y/n$:

Stelling 1.6 *Indien $X_1, X_2, \ldots, X_n$ onafhankelijke kansvariabelen zijn met verwachte waarde $E(X_i) = \mu$ en variantie $\mathrm{var}(X_i) = \sigma^2$, dan geldt voor een voldoende grote waarde n dat*

1. *de nieuwe kansvariabele $\overline{X} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$ benaderend normaal verdeeld is met verwachte waarde μ en variantie $\frac{\sigma^2}{n}$,*
2. *en bijgevolg dat de nieuwe kansvariabele*

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

benaderend standaardnormaal verdeeld is.

Een belangrijke praktische vraag is natuurlijk hoe groot de steekproefgrootte n dient te zijn vooraleer de centrale limietstelling geldig is. Hierop is geen algemeen antwoord te geven. De benodigde grootte van n hangt immers af van de verdeling of dichtheid van de individuele kansvariabelen X_i:

- Indien de kansdichtheid van elke waarneming X_i lijkt op de normale dichtheid, dan volstaat $n = 5$.
- Indien de kansverdeling of de kansdichtheid van elke waarneming X_i geen uitgesproken pieken vertoont, zoals bijvoorbeeld de uniforme dichtheid, dan volstaat doorgaans $n = 12$.
- Indien de kansverdeling of de kansdichtheid van elke waarneming X_i wel uitgesproken pieken vertoont, dan is het moeilijk om één specifieke waarde van n naar voor te schuiven. Een waarde van $n = 100$ zal meestal wel volstaan. Een voorbeeld van een gepiekte verdeling is $P(X = 1) = 0.06$ en $P(X = 10) = 1 - P(X = 1) = 0.94$.
- Voor de meeste kansvariabelen volstaat $n = 25$ of $n = 30$.

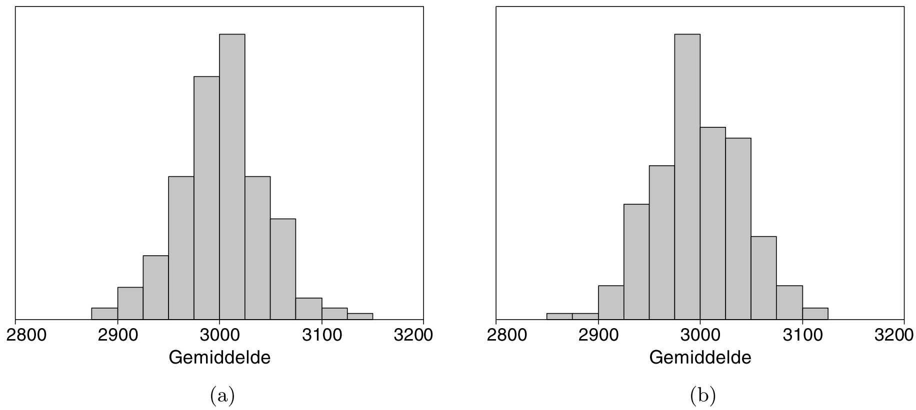
In de volgende paragraaf wordt de derde versie van de centrale limietstelling (Stelling 1.6) met behulp van simulaties in detail geïllustreerd.

1.5.4 Illustratie van de centrale limietstelling

Veronderstel dat een aantal studenten geïnteresseerd is in de waarde van de BEL20-index van de Brusselse beurs. Student 1 zal daarom een steekproef nemen van n waarnemingen van de BEL20-index en het gemiddelde daarvan berekenen, namelijk $\overline{X}_1$. Student 2 zal eveneens een steekproef nemen van n waarnemingen. Aangezien de BEL20-index van minuut tot minuut verandert, zal Student 2 natuurlijk andere waarden observeren van de BEL20-index (tenzij hij n keer net op hetzelfde moment zou observeren als Student 1). Ook Student 2 berekent het gemiddelde van zijn steekproef: $\overline{X}_2$. Op die manier verzamelen alle studenten n waarnemingen en berekenen zij elk van hun steekproef het gemiddelde. Als er 200 studenten zijn, dan krijgen we uiteindelijk 200 steekproefgemiddeldes $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_{200}$.

De derde versie van de centrale limietstelling zegt nu in feite niets anders dan dat deze 200 gemiddeldes uit een dichtheid komen die sterk gelijkt op de normale kansdichtheid. Met behulp van een histogram van deze 200 gemiddeldes is het eenvoudig om dit te verifiëren.

Dit is precies wat in deze paragraaf zal gebeuren. Bij gebrek aan studenten zullen we deze situatie simuleren of nabootsen in het pakket JMP. We zullen JMP 200 keer een steekproef van n waarnemingen laten nemen, de 200 steekproefgemiddeldes berekenen en vervolgens een histogram maken van deze steekproefgemiddeldes. Deze simulatie vereist wel dat we een kansverdeling of een kansdichtheid specificeren waaruit JMP waarnemingen kan genereren. We starten met een normale verdeling. We doen dus eerst alsof de BEL20-index een normaal verdeelde kansvariabele is. Als gemiddelde voor die normale verdeling gebruiken we $\mu = 3000$ (ongeveer de stand



Figuur 1.3

Twee mogelijke histogrammen van 200 steekproefgemiddeldes voor normaal verdeelde gegevens en steekproeven van 5 waarnemingen.

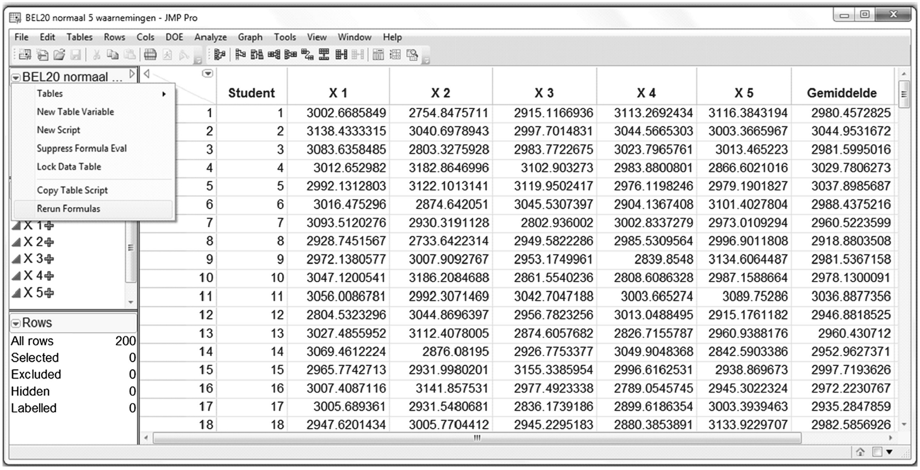
van de index toen de eerste versie van dit boek geschreven werd in 2004) en als standaarddeviatie $\sigma = 100$ (een keuze).

1.5.4.1 Normaal verdeelde X

Stel eerst dat elke student een steekproef van 5 waarnemingen neemt, met andere woorden, dat $n = 5$. Dan dienen we met behulp van JMP 200 keer 5 waarnemingen te simuleren. Daarvoor moeten we in JMP een gegevenstabel maken met 200 rijen van 5 kolommen met pseudo-willekeurige getallen getrokken uit een normale kansdichtheid met $\mu = 3000$ en $\sigma = 100$. De formule die we moeten invoeren voor elk van de 5 kolommen is “`Random Normal(3000, 100)`”. Van elke steekproef van 5 waarnemingen in iedere rij berekenen we het gemiddelde, en alle gemiddeldes worden vervolgens in een histogram gezet. Als we op dezelfde wijze een tweede gegevenstabel creëren, dan zetten we als het ware een tweede groep van 200 studenten aan het werk die ook steekproeven van 5 waarnemingen verzamelen. Twee mogelijke histogrammen, op die manier verkregen, zijn afgebeeld in Figuur 1.3. De verkregen histogrammen zijn behoorlijk klokvormig, wat illustreert dat de steekproefgemiddeldes normaal verdeeld zijn.

De snelste manier om 200 nieuwe steekproeven te genereren is om aan JMP te vragen de ingevoerde formule “`Random Normal(3000, 100)`” opnieuw te berekenen. Dit kan met behulp van het commando “`Rerun Formulas`”, dat verschijnt als u op het rode driehoekje klikt naast de naam van de gegevenstabel. Dit wordt geïllustreerd in Figuur 1.4.

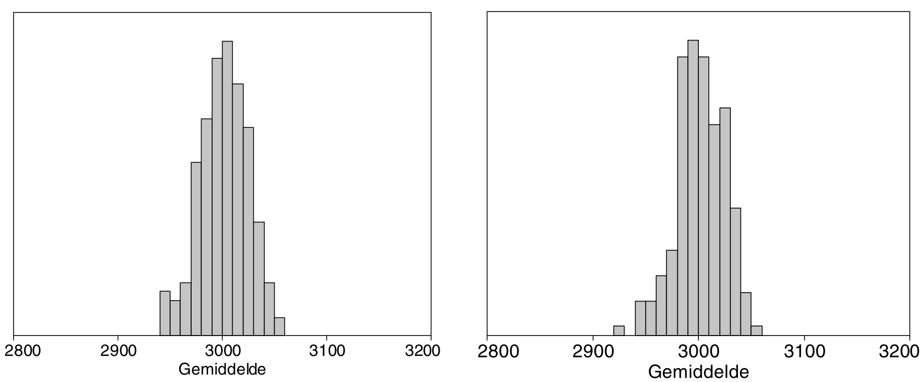
Wanneer de studenten elk 20 in plaats van 5 waarnemingen zouden verrichten, dan hebben de histogrammen een andere vorm: ze zijn nog steeds klokvormig maar wel beduidend smaller. Twee histogrammen voor 200 steekproefgemiddeldes



	Student	X 1	X 2	X 3	X 4	X 5	Gemiddelde
1	1	3002.6685849	2754.8475711	2915.1166936	3113.2692434	3116.3843194	2980.4572825
2	2	3138.4333315	3040.6978943	2997.7014831	3044.5665303	3003.3665967	3044.9531672
3	3	3083.6358485	2803.3275928	2983.7722675	3023.7965761	3013.465223	2981.5995016
4	4	3012.652982	3182.8646996	3102.903273	2983.8800801	2866.6021016	3029.7806273
5	5	2992.1312803	3122.1013141	3119.9502417	2976.1198246	2979.1901827	3037.8985687
6	6	3016.475296	2874.642051	3045.5307397	2904.1367408	3101.4027804	2988.4375216
7	7	3093.5120276	2930.3191128	2802.936002	3002.8337279	2973.0109294	2960.5223599
8	8	2928.7451567	2733.6422314	2949.5822286	2985.5309564	2996.9011808	2918.8803508
9	9	2972.1380577	3007.9092767	2953.1749961	2839.8548	3134.6064487	2981.5367158
10	10	3047.1200541	3186.2084688	2861.5540236	2808.6086328	2987.1588664	2978.1300091
11	11	3056.0086781	2992.3071469	3042.7047188	3003.665274	3089.75286	3036.8877356
12	12	2804.5323296	3044.8696397	2956.7823256	3013.0488495	2915.1761182	2946.8818525
13	13	3027.4855952	3112.4078005	2874.6057682	2826.7155787	2960.9388176	2960.430712
14	14	3069.4612224	2876.08195	2926.7753377	3049.9048368	2842.5903386	2952.9627371
15	15	2965.7742713	2931.9980201	3155.3385954	2996.6162531	2938.869673	2997.7193626
16	16	3007.4087116	3141.857531	2977.4923338	2789.0545745	2945.3022324	2972.2230767
17	17	3005.689361	2931.5480681	2836.1739186	2899.6186354	3003.3939463	2935.2847859
18	18	2947.6201434	3005.7704412	2945.2295183	2880.3853891	3133.9229707	2982.5856926

Figuur 1.4

Het genereren van nieuwe pseudo-willekeurige waarnemingen in JMP.

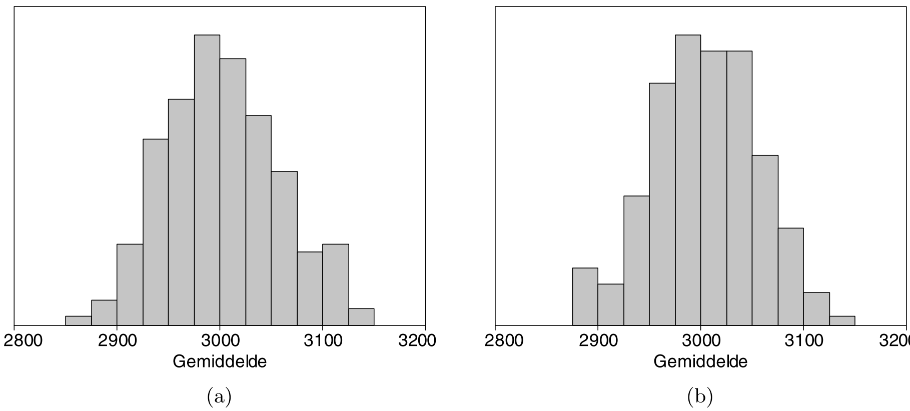


(a) (b)

Figuur 1.5

Twee mogelijke histogrammen van 200 steekproefgemiddeldes voor normaal verdeelde gegevens en steekproeven van 20 waarnemingen.

uit steekproeven met 20 waarnemingen zijn afgebeeld in Figuur 1.5. Dat de histogrammen nog steeds klokvormig zijn, geeft aan dat de steekproefgemiddeldes ook hier normaal verdeeld zijn. Het feit dat de histogrammen smaller zijn, mag geen verwondering baren aangezien de centrale limietstelling aangeeft dat de variantie van het steekproefgemiddelde gelijk is aan σ^2/n. Een steekproef van 20 waarnemingen levert dus een variantie op voor het steekproefgemiddelde die 4 keer kleiner is dan een steekproef van 5 waarnemingen.



Figuur 1.6

Twee mogelijke histogrammen van 200 steekproefgemiddeldes voor uniform verdeelde gegevens en steekproeven van 5 waarnemingen.

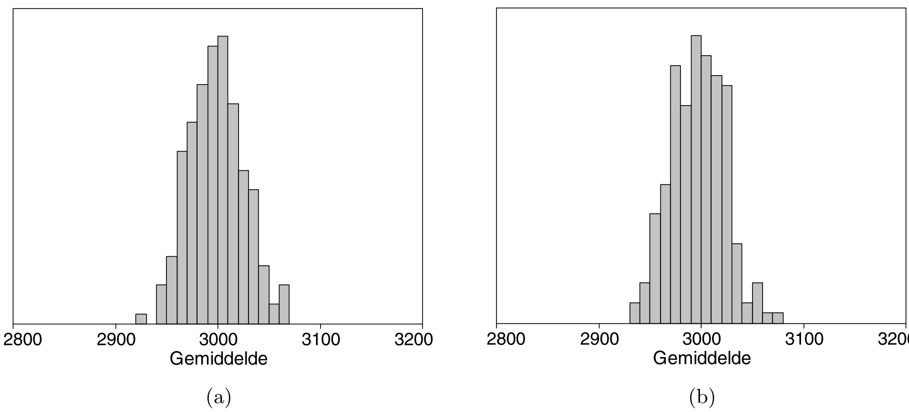
1.5.4.2 Uniform verdeelde *X*

Stel nu dat de waarde van de BEL20-index helemaal niet normaal verdeeld is, maar uniform verdeeld is tussen 2800 en 3200. Stel opnieuw eerst dat elke student een steekproef van 5 waarnemingen neemt. Dan dienen we met behulp van JMP wederom 200 keer 5 waarnemingen te simuleren. Dit komt neer op 5 keer de formule "`2800 + 400 * Random Uniform()`" in te voeren in een gegevenstabel met 200 rijen. Van elke steekproef van 5 waarnemingen berekenen we het gemiddelde, en alle gemiddeldes worden vervolgens in een histogram gezet. Twee mogelijke histogrammen, op die manier verkregen, zijn afgebeeld in Figuur 1.6. Het is opvallend dat de histogrammen opnieuw behoorlijk klokvormig zijn, wat illustreert dat, ook al zijn de oorspronkelijke gegevens uniform verdeeld, de steekproefgemiddeldes toch benaderend normaal verdeeld zijn.

Wanneer de studenten elk 20 in plaats van 5 waarnemingen zouden verrichten, dan zijn de histogrammen weer klokvormig en beduidend smaller. Twee histogrammen voor 200 steekproefgemiddeldes uit steekproeven met 20 waarnemingen zijn afgebeeld in Figuur 1.7.

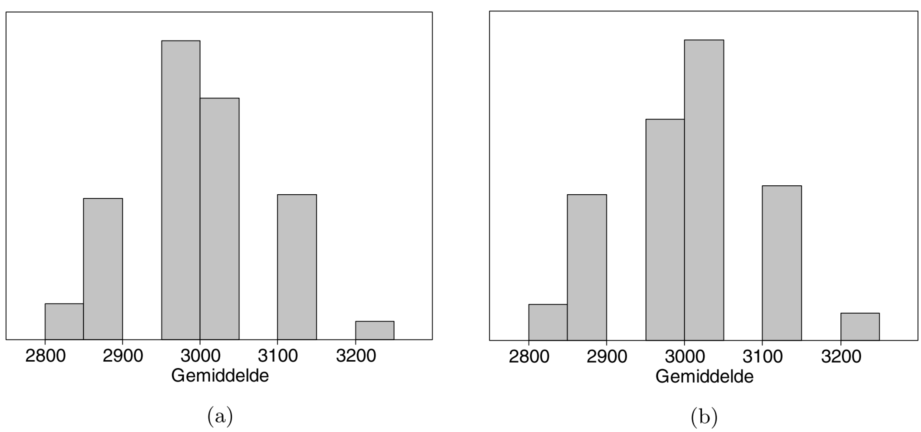
1.5.4.3 Bernoulli-verdeelde *X*

Stel nu dat de waarde van de BEL20-index Bernoulli-verdeeld is met 50% kans dat de waarde 2800 bedraagt en evenveel kans dat de waarde 3200 bedraagt. Stel opnieuw eerst dat elke student een steekproef van 5 waarnemingen neemt. Dan dienen we met behulp van JMP opnieuw 200 keer 5 waarnemingen te simuleren. Dit komt neer op 5 keer de formule "`2800 + 400 * Random Binomial(1, 0.5)`" in te voeren in een gegevenstabel met 200 rijen. Van elke steekproef van 5 waarnemingen berekenen



Figuur 1.7

Twee mogelijke histogrammen van 200 steekproefgemiddeldes voor uniform verdeelde gegevens en steekproeven van 20 waarnemingen.

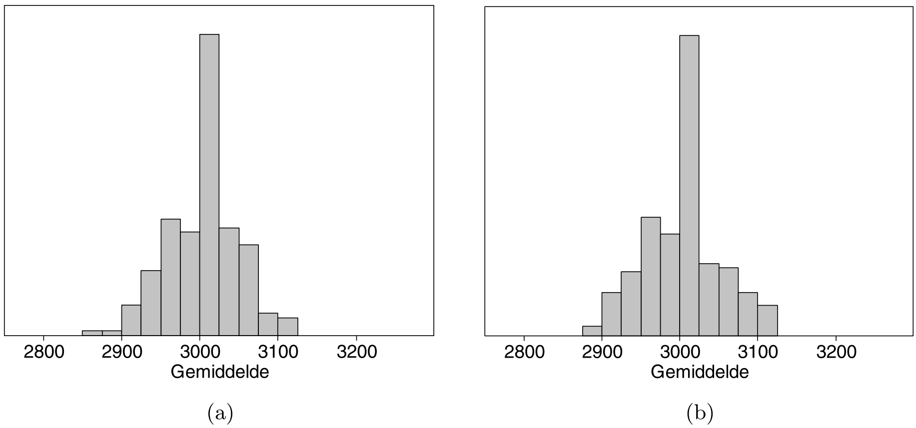


Figuur 1.8

Twee mogelijke histogrammen van 200 steekproefgemiddeldes voor Bernoulli-verdeelde gegevens en steekproeven van 5 waarnemingen.

we het gemiddelde, en alle gemiddeldes worden vervolgens in een histogram gezet. Twee mogelijke histogrammen, op die manier verkregen, zijn afgebeeld in Figuur 1.8. De histogrammen zijn dit keer niet klokvormig. Ze tonen duidelijk dat de originele gegevens afkomstig zijn van een discrete verdeling, namelijk de Bernoulli-verdeling. De centrale limietstelling lijkt voor de Bernoulli-verdeling en een steekproefgrootte van 5 waarnemingen dus niet te werken.

Wanneer de studenten evenwel elk 20 in plaats van 5 waarnemingen zouden verrichten, dan zien de histogrammen er totaal anders uit. Hoewel de histogrammen dan nog geen perfecte klokvorm vertonen, is toch al helemaal niet meer duidelijk dat



Figuur 1.9

Twee mogelijke histogrammen van 200 steekproefgemiddeldes voor Bernoulli-verdeelde gegevens en steekproeven van 20 waarnemingen.

de oorspronkelijke gegevens een discrete kansverdeling hadden. Twee histogrammen voor 200 steekproefgemiddeldes uit steekproeven met 20 waarnemingen zijn afgebeeld in Figuur 1.9. Om de histogrammen nog een betere klokvorm te bezorgen is een nog iets grotere steekproefgrootte vereist.

Dit laatste voorbeeld toont aan dat de centrale limietstelling bijzonder krachtig is. Zelfs kansverdelingen of kansdichtheden die er volstrekt anders uitzien dan de normale dichtheid leiden toch tot steekproefgemiddeldes die benaderend normaal verdeeld zijn. De voorwaarde daartoe is dat de steekproef voldoende waarnemingen bevat.

1.6 De steekproefproportie

Een steekproefproportie is een speciaal geval van het steekproefgemiddelde. Kenmerkend is dat de bestudeerde variabele in de populatie enkel de waarden 0 of 1 kan aannemen. Voorbeelden van dergelijke variabelen zijn geslacht (man/vrouw) en kwaliteit (defect/niet defect). In het algemeen spreekt men van successen of falingen. Voor de onbekende populatieproportie wordt de letter π gebruikt. Deze onbekende populatieproportie wordt geschat met behulp van de steekproefproportie, die niets anders is dan de relatieve frequentie van successen in een steekproef:

Definitie 1.6.1 De steekproefproportie is het aantal successen in een steekproef gedeeld door het aantal waarnemingen.

We kunnen de steekproefproportie beschouwen als een kansvariabele of als een berekend reëel getal. We beschouwen de steekproefproportie als een kansvariabele

Dit boek levert een toegankelijk en diepgaand overzicht van de belangrijkste basistechnieken uit de verklarende statistiek. Het boek legt eerst uit hoe gemiddeldes, proporties en varianties geschat kunnen worden op basis van steekproefgegevens, en hoe de (on)betrouwbaarheid van de schattingen in kaart gebracht kan worden. Daarna wordt aandacht besteed aan het formuleren van onderzoekshypothesen en aan het toetsen of testen van deze hypothesen.

Initieel gaat de aandacht uit naar hypothesetoetsen voor één enkele populatie. Vervolgens behandelt het boek hypothesetoetsen voor de vergelijking van twee populaties, en, ten slotte, hypothesetoetsen voor de vergelijking van meer dan twee populaties.

De klassieke parametrische hypothesetoetsen, die uitgaan van normaal verdeelde populaties, komen uitgebreid aan bod, maar er wordt ook heel wat aandacht besteed aan niet-parametrische alternatieven zoals de tekentoets, de rangtekentoets, de rangsomtoets en de Kruskal-Wallis-toets. Het boek besteedt ook ruime aandacht aan de bepaling van steekproefgroottes, het toetsen van equivalentie en aan hypothesetoetsen voor correlaties.

Doorheen het boek wordt gebruik gemaakt van het gebruiksvriendelijke, interactieve statistische pakket JMP voor het uitvoeren van berekeningen en het creëren van figuren. De benodigde tussenstappen worden in detail beschreven, zodat de lezer niet alleen de aangeleerde basisconcepten begrijpt, maar er ook mee aan de slag kan.

PETER GOOS

is gewoon hoogleraar aan de Faculteit Bio-Ingenieurswetenschappen van de Katholieke Universiteit Leuven en aan de Faculteit Toegepaste Economische Wetenschappen van de Universiteit Antwerpen, waar hij kansrekening en statistiek doceert aan studenten bio-ingenieur, toegepaste economische wetenschappen, sociaal-economische wetenschappen en handelsingenieur. Peter Goos is auteur van de boeken *Kansen en Verwachtingen* (Acco, 2009) en *Beschrijvende statistiek en kansrekenen* (Acco, 2013).

9 789033 495335