Algemeen overzicht 'inleiding kansrekening en statistiek'

Robert Fitzner* Tim Hulshof*

17 Oktober 2012 v.3

1 Voorwoord

Deze tekst geeft een overzicht van de stof die behandeld wordt in de meeste cursussen 'inleiding kansrekening en statistiek' die de wiskundefaculteit als servicevak aanbiedt. Oorspronkelijk is het geschreven als toevoeging op de stof van het vak 'Statistiek voor bouwkunde, 2S410' in 2010, maar het sluit in principe ook goed aan andere cursussen die gebaseerd zijn op hoofdstukken 1 tot en met 9 van Montgomerey & Runger. Het is niet bedoeld als vervanging van het boek of de hoorcolleges, maar als aanvulling op de stof. We beschrijven relevante onderwerpen die tijdens de colleges aan bod zijn gekomen en geven een aantal extra voorbeelden over de toepassing van de theorie.

Om wiskunde goed te beheersen is kennis van de stof belangrijk, maar oefenen met opgaven is minstens net zo belangrijk. Wiskunde leren is net als een taal leren: als je het kan lezen betekent dat nog niet dat je het ook kan schrijven. Schrijven leer je alleen door het te doen.

Het schrijven van wiskunde is ruwweg op te delen in twee verschillende aspecten: de boodschap en de grammatica. De boodschap wil hier zeggen dat de juiste berekening wordt uitgevoerd, de grammatica dat de notatie correct is. Correcte notatie is de enige manier om er zeker van te zijn dat de boodschap overkomt, let hier dus goed op!

Wij hebben zo goed mogelijk geprobeerd om een foutloos document te schrijven, maar in de praktijk blijkt dit bijna onmogelijk te zijn. Daarom waarschuwen we voor klakkeloos overnemen (niet alleen uit dit document maar in het algemeen). Wanneer je een fout tegenkomt in de tekst (of in je eigen berekeningen, wat dat betreft), ga dan na wat de oorzaak is en corrigeer het voor jezelf. Wie zich geroepen voelt kan ook ons aanspreken op eventuele fouten in de tekst, door even langs te komen op ons kantoor (MF 4.086), dan corrigeren wij het in de volgende versie.

Nog een algemene tip: als je op het tentamen in een berekening een antwoord krijgt dat nergens op slaat, zoals bijvoorbeeld een kans van -2 of 1090 (alle kansen moeten immers tussen 0 en 1 liggen), probeer dan het juiste antwoord te vinden. Als dat nou niet lukt, schrijf dan in ieder geval op dat je weet dat het antwoord fout is, dan krijg je op het tentamen misschien nog een paar punten voor het deel van de berekening dat wèl goed ging.

Robert Fitzner en Tim Hulshof

*Afdeling Wiskunde en Informatica, Technische Universiteit Eindhoven, 5600 MB Eindhoven. R.J.Fitzner@tue.nl,w.j.t.hulshof@win.tue.nl

2 Kansrekening

Wat is kansrekening? Kansrekening is de wetenschap van dingen die misschien wel en misschien niet gebeuren.

We drukken een kans (*probability*) altijd uit als een getal tussen 0 en 1. Heeft een gebeurtenis (*event*) kans 0, dan gebeurt het zeker niet, heeft het kans 1 dan zeker wel. Voor alle kansen met een waarde tussen 0 en 1 geeft de waarde van de kans aan hoe vaak we de gebeurtenis waarnemen in proportie met hoe vaak we een experiment doen.

Voorbeeld 1 *Een munt landt met kans 0.5 op kop. Doen we het experiment (de munt opgooien) 100 keer, dan verwachten we dus dat we* $0.5 \times 100 = 50$ *keer kop zien.*

Voorbeeld 2 *Van 240 studenten halen er 50 een 8 op een tentamen. Wat is dan de kans dat een student een 8 heeft? Die kans is* $\frac{50}{240} \approx 0.21$.

Verzamelingen Kansrekening begint altijd met het bepalen van de uitkomstenruimte (*sample space*). Voordat we de kans op een gebeurtenis kunnen berekenen moeten we namelijk wel eerst weten wat de mogelijke uitkomsten zijn. De uitkomstenruimte is een verzameling (*set*). Voor de uitkomstenruimte schrijven we vaak S.

Voorbeeld 3 *De uitkomstenruimte van het gooien met één dobbelsteen is*

$$S = \{1,2,3,4,5,6\}.$$

De uitkomstenruimte van een knikker uit een vaas met 5 kleuren knikkers pakken is

$$S = \{\textit{wit, geel, rood, blauw, zwart}\}.$$

De uitkomstenruimte van waarden op een barometer (in bar) is

$$S = [970, 1100],$$

dat wil dus zeggen, alle getallen (ook met cijfers achter de komma) tussen 970 bar en 1100 bar.

Een gebeurtenis is een verzameling van mogelijke uitkomsten. Een gebeurtenis schrijven we ook vaak met een hoofdletter, bijvoorbeeld E.

Voorbeeld 4 *De gebeurtenis dat er met een dobbelsteen 2,3 of 6 wordt gegooid:*

$$E = \{2,3,6\},$$

of de gebeurtenis dat we een rode of een gele knikker pakken uit de vaas:

$$E = \{\textit{rood, geel}\}.$$

Vaak is het nodig dat we verschillende verzamelingen samenvoegen, of dat we alleen de elementen uit de verzamelingen kiezen die hetzelfde zijn. Als A en B twee verzamelingen zijn, dan schrijven we $A \cup B$ voor de vereniging (*union*), en $A \cap B$ voor de doorsnede (*intersection*). Vaak kan je $\cup$ interpreteren als 'of' en kan je $\cap$ interpreteren als 'en'.

Voorbeeld 5 *Als A en B gegeven worden door*

$$A = \{1,2,3,8,9,10\} \qquad \textit{en} \qquad B = \{1,2,3,4,5\}$$

dan hebben we

$$A \cup B = \{1,2,3,4,5,8,9,10\} \qquad \textit{en} \qquad A \cap B = \{1,2,3\}.$$

Als A en B gegeven worden door

$$A = [0,10] \qquad \textit{en} \qquad B = [7,11]$$

dan hebben we

$$A \cup B = [0,11] \qquad \textit{en} \qquad A \cap B = [7,10].$$

Als we een gebeurtenis A hebben met een uitkomstenruimte S, dan is het complement van A alles in S, dat niet in A zit. We schrijven A^C voor het complement van A. Soms hebben we de lege verzameling (*empty set*) nodig. Dat is de verzameling zonder elementen. We schrijven $\emptyset$ voor de lege verzameling. Als je A en A^C verenigd, krijg je dus de hele uitkomstenruimte S, maar de doorsnede van A en A^C is leeg:

$$A \cup A^C = S \qquad \text{en} \qquad A \cap A^C = \emptyset. \tag{1}$$

Kansen Wat we uiteindelijk willen bepalen is de kans dat we een bepaalde gebeurtenis waarnemen. Als E die gebeurtenis is, dan schrijven we $\mathbb{P}(E)$ voor de kans op die gebeurtenis. De gebeurtenis E is een verzameling, maar $\mathbb{P}(E)$ is een getal tussen 0 en 1! Over het algemeen is er veel werk voor nodig om die kans te berekenen. Soms is het makkelijk, zoals bij dobbelstenen of munten, waar de kans op elke uitkomst gelijk is. Ook is de kans op S altijd 1, en de kans op $\emptyset$ altijd 0. Een paar handige vergelijkingen zijn:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B); \tag{2}$$

$$\mathbb{P}(A \cup A^C) = \mathbb{P}(S) = 1; \tag{3}$$

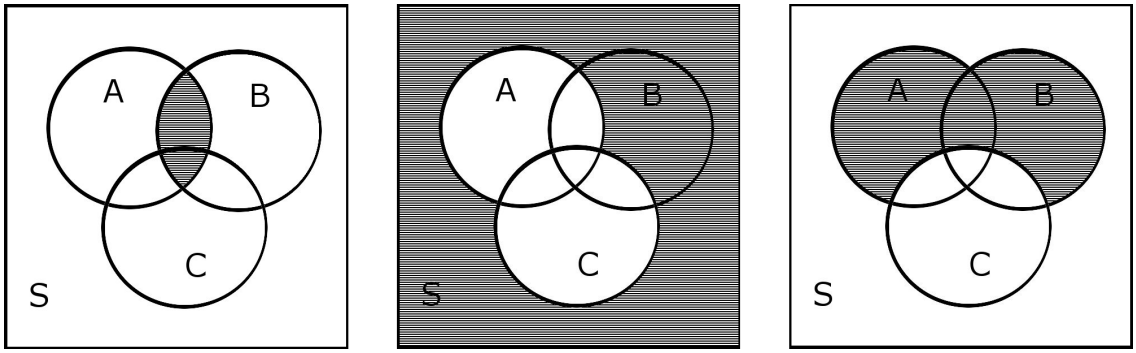
$$\mathbb{P}(A \cap A^C) = \mathbb{P}(\emptyset) = 0. \tag{4}$$

Twee gebeurtenissen A en B die elkaar uitsluiten (zoals bijvoorbeeld in één worp zowel kop als munt krijgen) hebben de eigenschap dat

$$A \cap B = \emptyset, \qquad \text{dus ook} \qquad \mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0 \tag{5}$$

$$\text{en} \qquad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B). \tag{6}$$

De relaties tussen verzameling kunnen we zichtbaar maken met een Venn diagram. In een Venn diagram worden gebeurtenissen die elkaar uitsluiten altijd weergegeven als cirkels die elkaar niet snijden (zie Figuur 1).



Figuur 1: De donkere delen in het Venn diagram komen overeen met links: $A \cap B$, midden: $(A \cup C)^C$, rechts: $(A \cup B) \cap C^C$.

Conditionele kansen Soms is het handig om te weten wat de kans is dat een gebeurtenis A gezien wordt als we weten dat een andere gebeurtenis, B, al gezien is. Stel dat A de gebeurtenis is dat er een tsunami op komst is. De kans daarop is (hopelijk) erg klein, en niet echt interessant. Maar als B de gebeurtenis is dat er een aardbeving heeft plaatsgevonden, dan is de kans van A geconditioneerd op B misschien veel groter, en wat relevanter. Voor de kans op A geconditioneerd op B schrijven we $\mathbb{P}(A|B)$. Een belangrijke formule bij het berekenen van conditionele kansen is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{7}$$

Omgekeerd geldt ook dat

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A). \tag{8}$$

Onafhankelijkheid Als twee gebeurtenissen, A en B, onafhankelijk zijn dan heeft het wel of niet gebeuren van A dus van geen enkele invloed op het wel of niet gebeuren van B. Dat betekent dus ook dat het conditioneren op B geen invloed heeft op de kans van A en vice versa:

$$\mathbb{P}(A|B) = \mathbb{P}(A) \qquad \text{en} \qquad \mathbb{P}(B|A) = \mathbb{P}(B). \tag{9}$$

Let wel op! Dit is niet waar als A en B niet onafhankelijk zijn.

Voorbeeld 6 *Simpele voorbeelden van onafhankelijke gebeurtenissen zijn munt- en dobbelsteenworpen. Stel dat $A = \{$de 101e worp met een munt is kop$\}$ en $B = \{$worpen 1 t/m 100 zijn allemaal kop$\}$. Deze gebeurtenissen zijn onafhankelijk want een munt heeft geen geheugen, dus*

$$\mathbb{P}(A|B) = \mathbb{P}(A) = \frac{1}{2}$$

terwijl

$$\mathbb{P}(B) = \left(\frac{1}{2}\right)^{100} \approx 0.000000000000000000000000000789.$$

Een andere eigenschap van onafhankelijke gebeurtenissen is dat $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$ (voor algemene gebeurtenissen is dit dus niet waar!).

Totale kansregel en Bayes' regel De kans dat gebeurtenis A gebeurt is gelijk aan de kans dat zowel A als B gebeuren plus de kans dat A gebeurt en B niet gebeurt (i.e. B^C):

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^C). \tag{10}$$

Als we formule (7) toepassen krijgen we de totale kansregel:

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^C)\mathbb{P}(B^C). \tag{11}$$

Voorbeeld 7 *Stel (1) dat* 90% *van alle Nederlandse kabouters een tuinkabouter is, en (2) dat van de kabouters die niet uit Nederland komen, slechts* 15% *een tuinkabouter is. Stel verder (3) dat* 5% *van alle kabouters uit Nederland komen. Hoe groot is dan de kans dat de eerste kabouter die je op de internationale kabouterconferentie tegenkomt een tuinkabouter is?*

Om die vraag op te lossen passen we de totale kansregel toe: Noem A de gebeurtenis dat een kabouter uit Nederland komt en B de gebeurtenis dat een kabouter een tuinkabouter is. Dan volgt uit (1), (2) en (3) dat

$$\mathbb{P}(B|A) = 0.9, \qquad \mathbb{P}(B|A^C) = 0.15 \qquad \textit{en} \qquad \mathbb{P}(A) = 0.05.$$

Deze passen we toe:

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^C)\mathbb{P}(A^C) = 0.9 \times 0.05 + 0.15 \times 0.95 = 0.1875.$$

Soms willen we $\mathbb{P}(A|B)$ weten, maar weten we alleen $\mathbb{P}(A)$, $\mathbb{P}(B)$ en $\mathbb{P}(B|A)$. Dan kunnen we Bayes' regel gebruiken om uit wat we wel weten, de kans van A geconditioneerd op B te berekenen. Bayes' regel:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{12}$$

Voorbeeld 8 *Als we de kansen uit Voorbeeld 7 gebruiken, kunnen we ook de kans berekenen dat een kabouter uit Nederland komt als we conditioneren op de gebeurtenis dat het een tuinkabouter is, oftewel, de kans $\mathbb{P}(A|B)$. We gebruiken Bayes' regel:*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{0.9 \times 0.05}{0.1875} = 0.24.$$

Merk op dat we om deze kans te berekenen uit de gegevens we eerst de totale kansregel moeten toepassen om $\mathbb{P}(B)$ te berekenen, en pas daarna Bayes' regel kunnen gebruiken.

Stochasten Als we een experiment uitvoeren is de uitkomst (in ieder geval in de situaties die wij meestal bestuderen) een getal. Bijvoorbeeld het aantal stippen op de bovenste zijde van een dobbelsteen die we geworpen hebben. We noemen de uitkomst van zo'n experiment een stochast (*random variable*). Een stochast is willekeurig/toevallig (*random*) in de zin dat elke herhaling van het experiment een ander resultaat op kan leveren. We schrijven stochasten steeds met een hoofdletter (vaak gebruiken we X). De kans dan dat X, de uitkomst van ons experiment, gelijk is aan een getal k schrijven we als $\mathbb{P}(X = k)$. Dit is dus de kans op de gebeurtenis $\{X = k\}$.

2.1 Discrete kansverdeling

We gebruiken de discrete kansverdeling (*discrete probability distribution*) als de uitkomstenruimte aftelbaar is. Dit is bijvoorbeeld het geval als het experiment maar een eindig aantal verschillende uitkomsten kan hebben (bijvoorbeeld een dobbelsteen, waar $S = \{1,2,\ldots,6\}$). Ook de oneindige uitkomstenruimte van de natuurlijke getallen, dat wil zeggen, $S = \{1,2,3,\ldots\}$ is aftelbaar. Het woord 'discreet' betekent in onze context 'gescheiden' of 'niet continu'.

Voorbeeld 9 *Een aantal voorbeelden van discrete uitkomstenruimten:*

- *Het aantal mensen in een bus:* $S = \{0,1,2,\ldots 100\}$*;*
- *Het aantal ingrediënten in een soep:* $S = \{2,3,4,\ldots\}$*;*
- *De temperatuur afgerond naar hele graden (Celsius):* $S = \{-273,-272,-271,\ldots\}$*;*
- *De tijd afgerond naar tienden van seconden:* $S = \{0.0,0.1,0.2,\ldots\}$.

Eén van de belangrijkste eigenschappen van stochasten is de verwachte waarde (*expectation*). Voor de verwachte waarde van een stochast X schrijven we $\mathbb{E}[X]$. Als we de kansen $\mathbb{P}(X = k)$ weten voor elke k in de uitkomstenruimte, dan wordt $\mathbb{E}[X]$ voor een discreet verdeelde stochast X gegeven door

$$\mathbb{E}[X] = \sum_k k\mathbb{P}(X = k). \tag{13}$$

waar we voor k altijd alle elementen in de uitkomstenruimte S invullen. De verwachting kun je zien als een soort 'gemiddelde uitkomst' van het experiment. Let wel op, als X bijvoorbeeld de stochast van een worp met een dobbelsteen is, dan krijgen we $\mathbb{E}[X] = 3\frac{1}{2}$ (zie Voorbeeld 10), hoewel er geen halve stippen op de dobbelsteen staan. De verwachting van X hoeft dus niet per sé een mogelijke uitkomst te zijn. Vaak schrijven we voor de verwachting van een stochast de Griekse letter μ.

Naast de verwachting willen we ook weten hoe ver een uitkomst in een typisch geval naast de verwachting zit. Deze afwijking kunnen we uitdrukken met behulp van een andere belangrijke eigenschap van stochasten: de variantie (*variance*). Voor de variantie zijn een aantal verschillende (maar equivalente) formules te bedenken. Helaas zit er niet een erg makkelijke tussen. Een aantal veelgebruikte formules zijn:

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 && (14)\\ &= \mathbb{E}[X^2] - \mu^2 && (15)\\ &= \sum_k k^2\mathbb{P}(X = k) - (\mathbb{E}[X])^2. && (16) \end{aligned}$$

Omdat de variantie uitgedrukt wordt als een som van kwadraten zijn de bijbehorende eenheden voor de variantie ook anders dan de eenheden voor de stochast. Dus als X wordt uitgedrukt in meters, dan wordt $\mathrm{Var}(X)$ uitgedrukt in meters^2 (vierkante meters). We schrijven dan ook vaak voor de variantie σ^2. Een natuurlijker getal om mee te werken is dus de wortel van de variantie. We noemen de wortel van de variantie ook wel de standaarddeviatie (*standard deviation*). We schrijven hiervoor σ:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathrm{Var}(X)}. \tag{17}$$

Als de standaarddeviatie σ klein is ten opzichte van de verwachting μ, dan wijkt een typische meting (X) niet veel af van het verwachte resultaat ($\mathbb{E}[X]$) en het omgekeerde geldt natuurlijk ook.

Voorbeeld 10 *Een makkelijk voorbeeld om te beginnen. De stochast X is de uitkomst van een worp met een dobbelsteen. De mogelijke uitkomsten zijn* $\{1,2,3,4,5,6\}$. *Verder hebben alle uitkomsten dezelfde kans, namelijk* $\mathbb{P}(X=k)=\frac{1}{6}$ *voor* $k=1,2,\ldots,6$. *We berekenen de verwachting, de variantie en de standaarddeviatie van X:*

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=1}^{6} k\,\mathbb{P}(X=k) = \sum_{k=1}^{6} \frac{k}{6} = \frac{1}{6}(1+2+3+4+5+6) = 3.5;\\
\mathrm{Var}(X) &= \sum_{k=1}^{6} k^2\mathbb{P}(X=k) - (\mathbb{E}[X])^2 = \sum_{k=1}^{6} \frac{k^2}{6} - (3.5)^2\\
&= \frac{1}{6}\left(1+2^2+3^2+4^2+5^2+6^2\right) - 12.25 = \frac{91}{6} - 12.25 = 2.92;\\
\sigma &= \sqrt{\mathrm{Var}(X)} = 1.71.
\end{aligned}$$

Voorbeeld 11 *Stel dat X de stochast is van het totaal aantal fietsen dat tijdens je studie van je gestolen wordt. Laten we stellen dat de kansen $\mathbb{P}(X=k)$ als volgt zijn:*

Aantal fietsen gestolen: k	*0*	*1*	*2*	*3*	*4*	*5*
Kans op dit aantal: $\mathbb{P}(X=k)$	*0.05*	*0.1*	*0.2*	*0.3*	*0.3*	*0.05*

De verwachting en de variantie zijn dan

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{5} k\,\mathbb{P}(X=k)\\
&= 0\times 0.05 + 1\times 0.1 + 2\times 0.2 + 3\times 0.3 + 4\times 0.3 + 5\times 0.05 = 2.85\\
\mathrm{Var}(X) &= \sum_{k=0}^{5} k^2\mathbb{P}(X=k) - (\mathbb{E}[X])^2\\
&= 0^2\times 0.05 + 1^2\times 0.1 + 2^2\times 0.2 + 3^2\times 0.3 + 4^2\times 0.3 + 5^2\times 0.05 - (2.85)^2\\
&= 9.65 - 8.1225 = 1.5275.
\end{aligned}$$

en de standaarddeviatie is $\sigma = \sqrt{1.5275} = 1.23592$.

De kansverdelingsfunctie We kunnen de kans $\mathbb{P}(X=k)$ ook zien als een functie van k. We schrijven $f(k)=\mathbb{P}(X=k)$. Deze functie $f(k)$ noemen we de kansverdelingsfunctie (*probability mass function*). Kansverdelingsfuncties hebben een aantal eigenschappen: stel X is een stochast met mogelijke uitkomsten $x_1,x_2,\ldots,x_n$, dan geldt dat

1. $f(x_i)\geq 0$ voor $i=1,2,\ldots,n$;
2. $\sum_{i=1}^{n} f(x_i)=1$;
3. $f(x_i)=\mathbb{P}(X=x_i)$.

Binomiaalverdeling Als we een experiment n keer uitvoeren, en de kans op succes van één experiment is p, wat is dan de kans dat we x keer een succes zien? De verdeling die bij dit soort vragen hoort is de binomiaalverdeling (*binomial distribution*), en de kansverdelingsfunctie wordt gegeven door

$$f(x) = \mathbb{P}(x \text{ keer succes in } n \text{ experimenten}) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x=0,1,\ldots,n. \tag{18}$$

Hier is $\binom{n}{x}$ de binomiaalcoëfficiënt van n en x, en deze wordt gegeven door

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}, \qquad \text{met} \qquad m! = m\times(m-1)\times\cdots\times 2\times 1. \tag{19}$$

Voorbeeld 12 *We gooien zeven keer met een dobbelsteen, en we willen de kans berekenen dat we drie keer één gooien. Dus $n = 7$, $x = 3$ en $p = \frac{1}{6}$. We vullen dit in in formule* (18)*:*

$$\begin{aligned} f(3) &= \binom{7}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^4 = \frac{7\times 6\times\cdots\times 2\times 1}{(3\times 2\times 1)\times(4\times 3\times 2\times 1)}\left(\frac{1}{216}\right)\left(\frac{625}{1296}\right) \\ &= 35\times\frac{625}{279936} \approx 0.078. \end{aligned}$$

2.2 Continue kansverdelingen

Als de uitkomst van een experiment (X dus) alle waarden in een interval kan aannemen (bijvoorbeeld alle kommagetallen tussen 0 en 1) gebruiken we een continue kansverdeling (*continuous probability distribution*). Om deze verdeling te beschrijven gebruiken we de bijbehorende kansdichtheidsfunctie (*probability density function*). We schrijven hiervoor $f(x)$. Een kansdichtheisfunctie heeft drie eigenschappen die altijd moeten gelden, namelijk:

1. $f(x) \geq 0$;
2. $\int_{-\infty}^{+\infty} f(x)\,dx = 1$;
3. $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx =$ het oppervlak onder $f(x)$ van a naar b.

Als we dus eigenschap (3) toepassen, kunnen we deze functie beschrijven om de kans uit te rekenen van gebeurtenissen van de vorm $\{a \leq X \leq b\}$, dat wil zeggen, de kans dat de uitkomst van het experiment in het interval $[a,b]$ ligt.

De kans dat een continue stochast één precieze waarde aanneemt is 0:

$$\mathbb{P}(X = a) = \mathbb{P}(a \leq X \leq a) = \int_a^a f(x)\,dx = F(a) - F(a) = 0, \tag{20}$$

waar $F(x)$ de primitieve van $f(x)$ is. Dit betekent ook dat het voor continue stochasten niet uitmaakt of we kijken naar 'kleiner dan en gelijk' ($\leq$) of naar 'strikt kleiner dan' ($<$):

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b). \tag{21}$$

De verwachting kunnen we nu als volgt berekenen:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,dx \tag{22}$$

en de variantie kunnen we dan ook berekenen:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - (\mathbb{E}[X])^2. \tag{23}$$

Hier integreren we in beide gevallen over het hele domein van $-\infty$ naar $+\infty$. Vaak zijn kansdichtheidsfuncties over een groot deel van dat domein gelijk aan 0, en alleen in een interval $[a,b]$ positief. We kunnen dan het deel van het domein waar $f(x) = 0$ weglaten, en alleen integreren van a naar b.

Integreren Voor we verder gaan met de kansdichtheidsfunctie gaan we kort in op integreren. Integralen zijn lineair, wat inhoudt dat voor elke twee functies $f(x)$ en $g(x)$ en ieder getal α geldt dat

$$\int_a^b \alpha f(x)\,dx = \alpha\int_a^b f(x)\,dx; \tag{24}$$

$$\int_a^b (f(x) + g(x))\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx. \tag{25}$$

Verder geldt voor het integreren van de functie $f(x) = x^n$ dat

$$\int_a^b x^n dx = \frac{1}{n+1} x^{n+1}\Big|_a^b = \frac{1}{n+1} b^{n+1} - \frac{1}{n+1} a^{n+1}. \tag{26}$$

Deze regel geldt voor alle waarden van n behalve $n = -1$. Als $n = -1$ (dus als $f(x) = \frac{1}{x}$) dan hebben we een speciale regel:

$$\int_a^b \frac{1}{x} dx = \ln(|x|)\Big|_a^b = \ln(|b|) - \ln(|a|). \tag{27}$$

Hier is $|x|$ de absolute waarde van x.

Als we een exponentiële functie willen integreren hebben we een andere regel nodig:

$$\int_a^b e^{\alpha x} dx = \frac{1}{\alpha} e^{\alpha x}\Big|_a^b = \frac{e^{\alpha b}}{\alpha} - \frac{e^{\alpha a}}{\alpha} \tag{28}$$

waar α een getal is.

Voorbeeld 13 *Een paar integralen:*

$$\begin{aligned}
\int_1^2 x^5 dx &= \frac{1}{6} x^6\Big|_1^2 = \frac{2^6}{6} - \frac{1^6}{6} = \frac{63}{6} = 10.5; \\
\int_{-1}^3 e^{2x-6} dx &= e^{-6} \int_{-1}^3 e^{2x} dx = \frac{e^{-6}}{2} e^{2x}\Big|_{-1}^3 \\
&= \frac{e^{-6}}{4} e^{2\times 3} - \frac{e^{-6}}{4} e^{2\times(-1)} = \frac{1}{4} - \frac{e^{-8}}{4} \approx 0.25.
\end{aligned}$$

Voorbeeld 14 *Stel dat we een stochast X hebben met een uniforme verdeling op het interval $[0,10]$. Deze verdeling houdt in dat X gewoon een waarde tussen 0 en 10 kiest. De kansdichtheidsfunctie die hier bij hoort is*

$$f(x) = \begin{cases} \frac{1}{10} & \text{als } 0 \le x \le 10; \\ 0 & \text{anders,} \end{cases}$$

dus deze functie is 1/10 tussen 0 en 10, en 0 daarbuiten. Stel dat a en b allebei tussen 0 en 10 liggen. De kans dat X ergens tussen a en b ligt, $\mathbb{P}(a \le X \le b)$, kunnen we dan uitrekenen met behulp van een integraal:

$$\mathbb{P}(a \le X \le b) = \int_a^b f(x) dx = \int_a^b \frac{1}{10} dx = \frac{1}{10} x\Big|_a^b = \frac{1}{10} b - \frac{1}{10} a = \frac{b-a}{10}.$$

We kunnen ook de verwachting en de variantie uitrekenen:

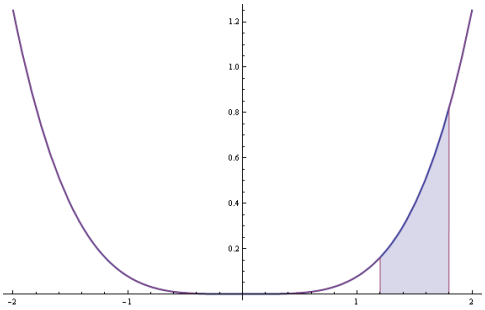
$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^0 0 dx + \int_0^{10} \frac{1}{10} x dx + \int_{10}^{\infty} 0 dx \\
&= \int_0^{10} \frac{1}{10} x dx = \frac{1}{10} \times \frac{1}{2} x^2\Big|_0^{10} = \frac{1}{20} 100 - \frac{1}{20} 0 = 5; \\
\mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \int_0^{10} \frac{1}{10} x^2 dx - (5)^2 = \frac{1}{10} \times \frac{1}{3} x^3\Big|_0^{10} - (5)^2 \\
&= \frac{1}{30} 1000 - 25 = 33\frac{1}{3} - 25 = 8\frac{1}{3};
\end{aligned}$$

en de standaarddeviatie is $\sigma = \sqrt{8\frac{1}{3}} = 2.88$.

Voorbeeld 15 *Een wat moeilijker voorbeeld. Stel X heeft kansdichtheidsfunctie $f(x) = \frac{5}{64}x^4$ op $[-2,2]$ en 0 daarbuiten. De kans $\mathbb{P}(a \leq X \leq b)$ als a en b beide tussen -2 en $+2$ liggen:*

$$\mathbb{P}(a \leq X \leq b) \quad = \quad \int_a^b f(x)dx = \int_a^b \frac{5}{64}x^4 dx = \frac{5}{320}x^5\Big|_a^b = \frac{5}{320}b^5 - \frac{5}{320}a^5$$

De grootte van het oppervlak onder $f(x)$ tussen a en b komt overeen met de kans $\mathbb{P}(a \leq X \leq b)$. In figuur 2 zie je de functie en het oppervlak onder $f(x)$ voor $a = 1.2$ en $b = 1.8$.



Figuur 2: Een grafische weergave van $\mathbb{P}(1.2 \leq X \leq 1.8)$ met $f(x) = \frac{5}{64}x^4$.

De verwachting en de variantie kunnen als volgt berekend worden:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x)dx = \int_{-2}^{2} x\frac{5}{64}x^4 dx \\
&= \int_{-2}^{2} \frac{5}{64}x^5 dx = \frac{5}{384}x^6\Big|_{-2}^{2} = \frac{320}{384} - \frac{320}{384} = 0; \\
\mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \int_{-2}^{2} x^2\frac{5}{64}x^4 dx - (0)^2 \\
&= \int_{-2}^{2} \frac{5}{64}x^6 dx = \frac{5}{7\times 64}x^7\Big|_{-2}^{2} = \frac{5\times 128}{488} - \frac{5\times(-128)}{488} = \frac{1280}{488} = 2.62;
\end{aligned}$$

en de standaarddeviatie is $\sigma = \sqrt{2.62} = 1.62$.

2.3 De cumulatieve distributiefunctie

De cumulatieve distributiefunctie (*cumulative distribution function*) of CDF, is een handig middel bij het berekenen van de kans dat een stochast in een interval $[a,b]$ ligt. Voor zowel discreet als continu verdeelde stochasten kunnen we een CDF vinden. De CDF drukt de kans uit dat X kleiner of gelijk is dan een gegeven getal x. We schrijven $F(x) = \mathbb{P}(X \leq x)$. Voor discrete verdelingen hebben we de volgende formule:

$$F(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f(x_i), \tag{29}$$

waar we sommeren over alle mogelijke uitkomsten x_i die kleiner dan of gelijk zijn aan x. Voor continue verdelingen moeten we integreren:

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(s)ds, \tag{30}$$

we integreren nu over variabele s omdat we x nu de bovengrens van het integratiedomein is, maar dit is alleen voor de duidelijkheid. De CDF van de kansdichtheidsfunctie $f(x)$ is dus niets anders dan de primitieve van $f(x)$. De eigenschappen van $f(x)$ vertalen zich dan ook naar overeenkomstige eigenschappen voor $F(x)$: $F(-\infty) = 0$, $F(+\infty) = 1$ en $F(x)$ is een functie die niet kan dalen, dat wil zeggen, als $x_1 < x_2$, dan geldt dat $F(x_1) \leq F(x_2)$.

Het is belangrijk om op te merken dat bij de discreet verdeelde stochasten er wel degelijk een verschil is tussen $\mathbb{P}(X \leq x)$ en $\mathbb{P}(X < x)$:

$$\mathbb{P}(X \leq x) = F(x), \qquad \text{terwijl} \qquad \mathbb{P}(X < x) = F(x_i),$$

zodanig dat x_i de grootste mogelijke uitkomst is die toch nog kleiner is dan x (bijvoorbeeld, als uitkomsten 1,2,3 en 4 mogelijk zijn, en we willen $\mathbb{P}(X \leq 3)$ weten, dan nemen we $F(3)$, maar als we $\mathbb{P}(X < 3)$ willen weten, dan nemen we $F(2)$).

We kunnen vanuit de CDF ook de kansverdelingsfunctie of de kansdichtheidsfunctie weer reconstueren. Voor discreet verdeelde stochasten is de kansverdelingsfunctie $f(x_i) = \mathbb{P}(X = x_i)$, dus

$$f(x_i) = \mathbb{P}(X \leq x_i) - \mathbb{P}(X \leq x_{i-1}) = F(x_i) - F(x_{i-1}). \tag{31}$$

Voor continu verdeelde stochasten kunnen we de kansdichtheidsfunctie vinden door de CDF te differentiëren (dat is immers het omgekeerde van integreren):

$$f(x) = \frac{dF(x)}{dx}. \tag{32}$$

Voorbeeld 16 *Stel we hebben een discreet verdeelde stochast X met CDF*

$$F(x) \quad = \quad \begin{cases} 0 & \text{voor } x < 1; \\ \frac{1}{7} & \text{voor } 1 \leq x < 2; \\ \frac{4}{7} & \text{voor } 2 \leq x < 3; \\ \frac{5}{7} & \text{voor } 3 \leq x < 5; \\ 1 & \text{voor } 5 \leq . \end{cases}$$

We rekenen een paar kansen uit:

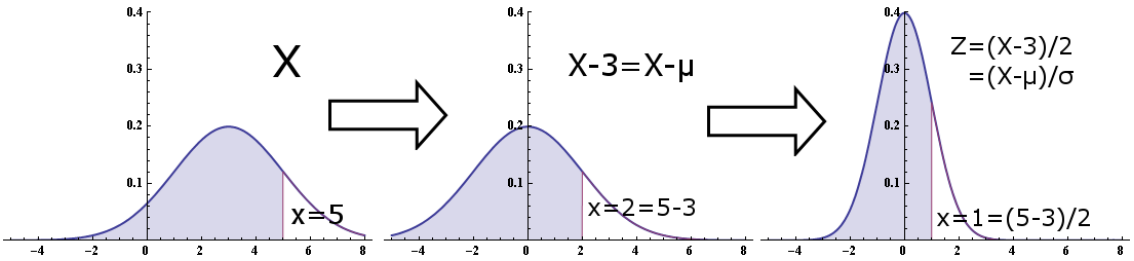
$$\begin{aligned} \mathbb{P}(X \leq 3) &= \frac{4}{7}; \\ \mathbb{P}(X \leq 3) &= F(3) = \frac{5}{7}; \\ \mathbb{P}(X > 4.2) &= 1 - \mathbb{P}(X \leq 4.2) = 1 - \frac{5}{7} = \frac{2}{7}; \\ \mathbb{P}(1 \leq X \leq 3) &= \mathbb{P}(X \leq 3) - \mathbb{P}(X \leq 1) = F(3) - F(1) = \frac{5}{7} - \frac{1}{7} = \frac{4}{7}. \end{aligned}$$

Met $f(x_i) = F(x_i) - F(x_{i-1})$ kunnen we ook de kansverdelingsfunctie uitrekenen:

$$\begin{aligned} F(1) &= f(1) = \frac{1}{7}; \\ F(2) - F(1) &= f(2) = \frac{4}{7} - \frac{1}{7} = \frac{3}{7}; \\ F(3) - F(2) &= f(3) = \frac{5}{7} - \frac{4}{7} = \frac{1}{7}; \\ F(5) - F(3) &= f(5) = 1 - \frac{5}{7} = \frac{2}{7}. \end{aligned}$$

Voorbeeld 17 *Stel we hebben een continu verdeelde stochast met CDF*

$$F(x) \quad = \quad \begin{cases} 0 & \text{voor } x \leq 0; \\ \frac{x^2}{4} & \text{voor } 0 \leq x \leq 2; \\ 1 & \text{voor } 2 \leq x. \end{cases}$$



Figuur 3: De standaardisering van een normaalverdeling met $\mu = 3$ en $\sigma^2 = 4$. Ook is het gebied dat overeenkomt met $\mathbb{P}(X \leq x)$ weergegeven voor $x = 5$. Het gekleurde gebied onder de grafiek heeft bij alle drie de verdelingen hetzelfde oppervlak.

We rekenen een paar kansen uit:

$$\begin{aligned}
\mathbb{P}(X \leq 0.3) &= F(0.3) = \frac{0.3^2}{4} = 0.0225; \\
\mathbb{P}(X \geq 1) &= 1 - \mathbb{P}(X \leq 1) = 1 - \frac{1^2}{4} = \frac{3}{4}; \\
\mathbb{P}(0.6 \leq X \leq 1.6) &= \mathbb{P}(X \leq 1.6) - \mathbb{P}(X \leq 0.6) = F(1.6) - F(0.6) \\
&= \frac{1.6^2}{4} - \frac{0.6^2}{4} = 0.64 - 0.09 = 0.55.
\end{aligned}$$

We kunnen ook de kansdichtheidsfunctie berekenen voor een punt x in het interval $[0,2]$:

$$f(x) = \frac{dF(x)}{dx} = \begin{cases} \frac{d}{dx}0 & \text{voor } x \leq 0; \\ \frac{1}{4}\frac{d}{dx}(x^2) & \text{voor } 0 \leq x \leq 2; \\ \frac{d}{dx}1 & \text{voor } 2 \leq x; \end{cases} = \begin{cases} 0 & \text{voor } x \leq 0; \\ \frac{x}{2} & \text{voor } 0 \leq x < 2; \\ 0 & \text{voor } 2 \leq x. \end{cases}$$

2.4 De normaalverdeling

Verreweg de belangrijkste verdeling is de normaalverdeling (*normal distribution*). Dit komt door het opmerkelijke feit dat de som van verschillende stochasten met allemaal dezelfde verdeling ook een stochast is, en dat deze nieuwe stochast een eigen verdeling heeft die (als we maar genoeg stochasten bij elkaar optellen) verdeeld is volgens de normaalverdeling. Dit is één van de belangrijkste resultaten uit de kansrekening: de Centrale Limietstelling (*Central Limit Theorem*). Let wel op! Een som van stochasten is niet altijd normaal verdeeld, en omgekeerd kan een stochast ook normaal verdeeld zijn als het geen som is van stochasten.

Voor de normaalverdeling schrijven we $\mathcal{N}(\mu, \sigma^2)$. Er zijn twee getallen die de normaalverdeling volledig beschrijven: μ is de verwachte waarde van een normaal verdeelde stochast, en σ^2 is de variantie. De kansdichtheidsfunctie van de normaalverdeling wordt gegeven door

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \text{voor alle } -\infty < x < +\infty.$$

Dit is een behoorlijk moeilijke functie om bijvoorbeeld te integreren, maar dat is gelukkig niet nodig. Als we een grafiek maken bij de functie is het wel goed te zien wat de betekenis van μ en σ is. De piek van $f(x)$ ligt namelijk precies waar $x = \mu$. Iets lastiger te zien is dat σ bepaalt hoe 'breed' de verdeling is. Als σ groot is, dan is de verdeling vrij breed, terwijl als σ heel klein is, de verdeling een scherpe piek vormt rond μ. Dit is ook logisch: als σ heel klein is, dan verwachten we ook dat er niet zo'n groot verschil is tussen verschillende stochasten met dezelfde verdeling, terwijl we bij een heel grote σ dat juist wel zouden verwachten.

Standaard normaalverdeling De normaalverdeling met $\mu = 0$ en $\sigma^2 = 1$ noemen we de standaard normaalverdeling (*standard normal distribution*). In de afgelopen eeuwen hebben overijverige wiskundigen voor een gigantisch aantal verschillende waarden van x de waarde van de cumulatieve distributiefunctie $F(x)$ van $\mathcal{N}(0,1)$ berekend en getabuleerd. Daar hebben wij wat aan omdat ongeacht de precieze waardes van μ en σ de normaalverdeling wel altijd dezelfde 'vorm' heeft. Als we de normaalverdeling met gegeven μ en σ 'verschuiven' over een afstand $-\mu$ en 'uitrekken' of 'samenpersen' met een factor $1/\sigma$, dan krijgen we de standaardnormaalverdeling. Dat kunnen we iets preciezer zeggen: als X een stochast is die normaal verdeeld is met μ en σ, dan is

$$Z = \frac{X-\mu}{\sigma} \tag{33}$$

ook een normaal verdeelde stochast, maar nu met $\mathbb{E}[Z] = 0$ en $\mathrm{Var}(Z) = 1$, dus Z is standaard normaalverdeeld (we noemen Z dan ook een standaard normale stochast). In Figuur 3 is schematisch weergegeven hoe de standaardisering van $\mathcal{N}(3,4)$ naar $\mathcal{N}(0,1)$ gaat.

Een opmerkelijk resultaat is dus dat de voor een stochast X die verdeeld is volgens $\mathcal{N}(\mu,\sigma^2)$ de kans op de gebeurtenis $\{X \le x\}$ precies hetzelfde is als de kans op $\{Z \le \frac{x-\mu}{\sigma}\}$, waar Z verdeeld is volgens $\mathcal{N}(0,1)$:

$$\mathbb{P}(X \le x) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \mathbb{P}(Z \le z). \tag{34}$$

In het statistisch compendium staat een tabel voor kansen van de vorm $\mathbb{P}(Z \le z)$.

Voorbeeld 18 *In de tabel voor $\mathbb{P}(Z \le z)$ staan de eerste twee decimalen van z in de kolom aan de linkerkant, en de laatste decimaal in de bovenste rij. Als we dus bijvoorbeeld de kans $\mathbb{P}(Z \le 1.64)$ willen berekenen kijken we in de kolom bij 1.6 en in de rij bij 0.04. Daar waar deze rij en kolom kruisen staat de kans. In dit geval is dat 0.949497. Omgekeerd kunnen we op die manier ook bij een gegeven kans de bijbehorende z vinden. Stel dat we willen weten voor welke waarde van z de kans $\mathbb{P}(Z \le z) = 0.35$ is. We zoeken dan in de tabel eerst het getal dat het dichtst bij 0.35 ligt. In dit geval is dat 0.348268. Bij deze kans kunnen we de kolom en de rij volgen. Dan zien we in de kolom -0.3 en in de rij -0.09. Tellen we deze getallen bij elkaar op, dan krijgen we $z = -0.39$.*

Voorbeeld 19 *De IQ-test is zo opgezet dat de uitkomsten voor een populatie normaal verdeeld zijn met $\mu = 100$ en $\sigma = 15$. Als X het IQ is van een willekeurige persoon (X is dus de stochast), dan kunnen we de kans berekenen dat die hoger is dan x (x is dus de parameter). Wat is bijvoorbeeld de kans dat iemand een IQ heeft dat hoger is dan 125?*

$$\begin{aligned}\mathbb{P}(X > 125) &= 1 - \mathbb{P}(X \le 125) = 1 - \mathbb{P}\left(\frac{X-100}{15} \le \frac{125-100}{15}\right)\\ &= 1 - \mathbb{P}(Z \le 1.67) = 1 - 0.952540 = 0.047460.\end{aligned}$$

Dat wil zeggen, minder dan 5% van de hele populatie heeft een IQ van boven de 125.

Stel nu dat je net een IQ-test hebt gedaan, en dat blijkt dat je die beter hebt gedaan dan 99.97% van de populatie. Wat is dan je IQ? We zoeken in de tabel de waarde van z die hier bij hoort. In dit geval is dat $z = 3.43$. Omdat $z = \frac{x-\mu}{\sigma}$ volgt dat $x = \sigma z + \mu$. In ons geval geeft dat $15 \times 3.43 + 100 = 151,45$. Je hebt dus een IQ van 151!

2.5 Lineaire combinaties van stochasten

Soms is een stochast uit te drukken in termen van andere stochasten. Veel voorkomend is dat een stochast een lineaire combinatie (*linear combination*) van andere stochasten is. Een lineaire combinatie houdt eigenlijk in dat we de stochasten mogen optellen, of dat we ze met een constante (een getal dus) mogen vermenigvuldigen. Verder mogen we niets.

Voorbeeld 20 *Stel dat X en Y stochasten zijn. Dan is $2X-3Y$ een lineaire combinatie, en $14X$ is in zekere zin ook een lineaire combinatie, maar $3XY$ is geen lineaire combinatie, en $Y + X - 4$ ook niet.*

Als we het voorbeeld wat generaliseren krijgen we een algemene formule voor de mogelijke lineaire combinaties Y van stochasten $X_1, X_2, \ldots, X_p$:

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_p X_p, \tag{35}$$

waar $c_1, c_2, \ldots, c_p$ constantes zijn (dus getallen, mogelijk negatief of 0).

Onafhankelijke stochasten Net als gebeurtenissen kunnen stochasten onafhankelijk zijn. In het geval van stochasten wil dat simpelweg zeggen dat wat we observeren in een experiment niet afhangt van wat we al eerder of ergens anders gezien hebben. De uitkomst van een opgeworpen munt is bijvoorbeeld een onafhankelijke stochast, maar de stand van een voetbalwedstrijd na n minuten, X_n, is geen onafhankelijke stochast, want die wordt beïnvloed door de stand in de voorgaande minuten (Als $X_{15} = (2-0)$, dan kan het onmogelijk zo zijn dat $X_{30} = (1-0)$.)

In veruit de meeste problemen die wij zullen tegenkomen zijn de stochasten onafhankelijk. Daarom gaan we altijd uit van onderlinge onafhankelijkheid van de stochasten, tenzij het anders wordt vermeld.

Verwachting en variantie De verwachting en variantie van lineaire combinaties van onafhankelijke stochasten zijn makkelijk te berekenen uit de verwachtingen en varianties van de onderliggende stochasten. Voor de verwachting van $Y = aX_1 + bX_2$ hebben we de simpele formule

$$\mathbb{E}[Y] = \mathbb{E}[aX_1 + bX_2] = a\mathbb{E}[X_1] + b\mathbb{E}[X_2]. \tag{36}$$

We mogen dus gewoon de verwachtingen van X_1 en X_2 nemen, ze vermenigvuldigen met a en b, respectievelijk, en optellen. Voor de variantie van Y moeten we iets beter opletten:

$$\mathrm{Var}(Y) = \mathrm{Var}(aX_1 + bX_2) = a^2\mathrm{Var}(X_1) + b^2\mathrm{Var}(X_2), \tag{37}$$

hier moeten we dus de varianties van X_1 en X_2 vermenigvuldigen met a^2 en b^2 voor we ze optellen. We kunnen de formules voor verwachting en variantie generaliseren naar lineaire combinaties van meer dan twee stochasten:

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[c_1 X_1 + c_2 X_2 + \cdots + c_p X_p] = c_1\mathbb{E}[X_1] + c_2\mathbb{E}[X_2] + \cdots + c_p\mathbb{E}[X_p]; && (38)\\
\mathrm{Var}(Y) &= \mathrm{Var}(c_1 X_1 + c_2 X_2 + \cdots + c_p X_p) && (39)\\
&= c_1^2\mathrm{Var}(X_1) + c_2^2\mathrm{Var}(X_2) + \cdots + c_p^2\mathrm{Var}(X_p). && (40)
\end{aligned}$$

Voorbeeld 21 *Stel we hebben de volgende lineaire combinaties van onafhankelijke stochasten:*

$$\begin{aligned}
Y_1 &= 2X_1 + 4X_2 - 8X_3;\\
Y_2 &= \frac{1}{4}\sum_{i=1}^{15} X_i;\\
Y_3 &= X_1 - X_2 - X_3 + X_4.
\end{aligned}$$

Stel dat $\mathbb{E}[X_i] = \mu$ *en* $\mathrm{Var}(X_i) = \sigma^2$ *voor alle* $i = 1, 2, \ldots, 15$. *Dan zijn de verwachtingen van* Y_1, Y_2 *en* Y_3*:*

$$\begin{aligned}
\mathbb{E}[Y_1] &= 2\mathbb{E}[X_1] + 4\mathbb{E}[X_2] - 8\mathbb{E}[X_3] = 2\mu + 4\mu - 8\mu = -2\mu;\\
\mathbb{E}[Y_2] &= \frac{1}{4}\sum_{i=1}^{15}\mathbb{E}[X_i] = \frac{15}{4}\mu;\\
\mathbb{E}[Y_3] &= \mathbb{E}[X_1] - \mathbb{E}[X_2] - \mathbb{E}[X_3] + \mathbb{E}[X_4] = \mu - \mu - \mu + \mu = 0.
\end{aligned}$$

De varianties zijn

$$\begin{aligned}
\mathrm{Var}(Y_1) &= 2^2\mathrm{Var}(X_1) + 4^2\mathrm{Var}(X_2) + (-8)^2\mathrm{Var}(X_3) = 4\sigma^2 + 16\sigma^2 + 64\sigma^2 = 84\sigma^2;\\
\mathrm{Var}(Y_2) &= \left(\frac{1}{4}\right)^2\sum_{i=1}^{15}\mathrm{Var}(X_i) = \frac{15}{16}\sigma^2;\\
\mathrm{Var}(Y_3) &= 1^2\mathrm{Var}(X_1) + (-1)^2\mathrm{Var}(X_2) + (-1)^2\mathrm{Var}(X_3) + 1^2\mathrm{Var}(X_4) = 4\sigma^2.
\end{aligned}$$

3 Statistiek

Bij kansrekening is het uitgangspunt dat stochasten op een bepaalde manier verdeeld zijn. Vanuit die gedachte berekenen we de kansen op bepaalde gebeurtenissen. Dat is leuk in theorie, maar als we een echt experiment uitvoeren is er niemand in de buurt om ons in te fluisteren wat die verdeling dan moge zijn. Wat we nodig hebben is een theorie voor het beschrijven van een gebeurtenis aan de hand van wat we observeren (namelijk de uitkomsten van een (herhaald) experiment). De tak van de wiskunde die zich hier mee bezig houdt is de statistiek. Om een statistiek te maken moeten we een aantal vragen beantwoorden:

1. Wat willen we meten (dus wat is onze stochast X)?
2. Wat zou de verdeling van deze stochast kunnen zijn (bijvoorbeeld: binomiaal, normaal)?
3. Van welke grootheden willen we een schatting maken (bijvoorbeeld: verwachting, variantie)?
4. Hoe goed is die schatting?
5. Wat is de conclusie?

Soms is de eerste vraag beantwoorden het moeilijkst. Bijvoorbeeld omdat we niet altijd in staat zijn precies dat te meten waar we werkelijk in geïnteresseerd zijn. Een andere reden kan zijn dat het systeem waar we aan willen meten zo complex is dat we niet weten wat de oorzaak is van het stochastische gedrag van onze metingen. We zullen hier niet verder op ingaan en in plaats daarvan aannemen dat we altijd weten wat onze stochasten zijn. Ook de tweede en derde vraag laten we hier onbeantwoord. (Op het tentamen moeten deze uit de vraagstelling wel duidelijk worden.) In de rest van deze samenvatting zullen we ons bezig houden met het antwoord op vragen (4) en (5).

We beginnen met schatters, dan betrouwbaarheidsintervallen (*confidence intervals*) waarvan we laten zien hoe we deze kunnen gebruiken om vraag (4) te beantwoorden. Daarna geven we aan hoe een hypothese (*hypothesis*) voor een experiment opgesteld en getoetst dient te worden.

3.1 Schatters

Statistiek laat zich beter uitleggen aan de hand van voorbeelden. Daarom nemen we hier even een lekker afgezaagd voorbeeldje door: Stel dat we een bepaalde grootheid willen weten, zoals bijvoorbeeld het gemiddelde aantal fietsen (noem het θ) die een persoon bezit tijdens zijn/haar studie in Eindhoven. Aan iedereen gaan vragen hoeveel dat er zijn is veel te veel werk voor zo'n trivialiteit, dus zijn we ook wel tevreden met een goede schatting van het gemiddelde, als dat minder werk is. Laten we voor het gemak eens beginnen met een steekproef onder 10 studenten. Hun antwoorden zijn $X_1, X_2, \ldots, X_{10}$. Nu nemen we het steekproefgemiddelde (*sample mean*):

$$\Theta_1 = \frac{1}{10}\sum_{i=1}^{10} X_i.$$

Dit is een schatter van het echte gemiddelde. Echter, het viel ons op dat de eerste en tweede persoon wel erg betrouwbaar overkwamen, terwijl de tiende persoon een nogal verwarde indruk wekte. Daarom tellen we X_1 en X_2 drie keer, en laten we het antwoord van persoon 10 weg. Dan kunnen we ook een schatter schrijven als volgt:

$$\Theta_2 = \frac{1}{13}\left(3X_1 + 3X_2 + \sum_{i=3}^{9} X_i\right).$$

Eigenlijk is elke combinatie van X_1 tot en met X_{10} wel mogelijk. Zo kunnen we bijvoorbeeld ook de schatter

$$\Theta_3 = \frac{1}{3}(3X_1 + 2X_2 + X_3)$$

bedenken.

Nu is het alleen nog de vraag welke schatter de juiste keuze is.

Zuivere en onzuivere schatters Een goede schatter moet om te beginnen de juiste waarde van de grootheid meten. Dat betekent dat de verwachting van de schatter de waarde moet zijn die we zoeken. Om dit soort precisie uit te drukken spreken we van de onzuiverheid (*bias*) van een schatter. De onzuiverheid is het verschil tussen de verwachte waarde van de schatter en de verwachte waarde van de te schatten grootheid:

$$\text{bias}(\Theta) \quad = \quad \mathbb{E}[\Theta] - \theta \tag{41}$$

Als de onzuiverheid 0 is, noemen we de schatter zuiver (*unbiased*).

Kiezen tussen schatters (MSE) Een kleine onzuiverheid is niet het enige criterion waar we onze schatter op moeten kiezen. We willen ook dat de variantie van de schatter klein is. Om een getalswaarde te geven aan de combinatie van zuiverheid en variantie gebruiken we de 'Mean Square Error' (MSE):

$$\begin{aligned} \text{MSE}(\Theta) &= \mathbb{E}\left[(\Theta-\theta)^2\right] && (42)\\ &= \text{Var}(\Theta) + (\text{bias}(\Theta))^2. && (43) \end{aligned}$$

Bij het vergelijken van schatters geven we altijd de voorkeur aan de schatter met de kleinste MSE.

Voorbeeld 22 *We vergelijken de drie schatters voor het gemiddelde aantal fietsen. We beginnen met na te gaan of ze zuiver zijn of niet:*

$$\begin{aligned} \mathbb{E}[\Theta_1] &= \mathbb{E}\left[\frac{1}{10}\sum_{i=1}^{10} X_i\right] = \frac{1}{10}\sum_{i=1}^{10}\mathbb{E}[X_i] = \frac{1}{10}\sum_{i=1}^{10}\theta = \theta;\\ \mathbb{E}[\Theta_2] &= \mathbb{E}\left[\frac{1}{13}\left(3X_1+3X_2+\sum_{i=3}^{9} X_i\right)\right] = \frac{1}{13}\left(3\mathbb{E}[X_1]+3\mathbb{E}[X_2]+\sum_{i=3}^{9}\mathbb{E}[X_i]\right)\\ &= \frac{1}{13}\left(6\theta+\sum_{i=3}^{9}\theta\right) = \theta;\\ \mathbb{E}[\Theta_3] &= \mathbb{E}\left[\frac{1}{3}(3X_1+2X_2+X_3)\right] = \frac{1}{3}(3\theta+2\theta+\theta) = 2\theta. \end{aligned}$$

Dus Θ_1 en Θ_2 zijn zuiver want hun verwachting is de θ die we zoeken, maar Θ_3 is onzuiver. De onzuiverheid wordt gegeven door $\mathbb{E}[\Theta_3]-\theta = \theta$.

Om te kiezen welke schatter het beste is moeten we de variantie ook berekenen. We nemen aan dat de antwoorden X_i onafhankelijk van elkaar zijn, en dat ze allemaal dezelfde variantie σ^2 hebben. Dan kunnen we de regels voor de variantie van een combinatie van onafhankelijke stochasten gebruiken:

$$\begin{aligned} \text{Var}(\Theta_1) &= \left(\frac{1}{10}\right)^2 \text{Var}\left(\sum_{i=1}^{10} X_i\right) = \frac{1}{100}\sum_{i=1}^{10}\text{Var}(X_i) = 0.1\sigma^2;\\ \text{Var}(\Theta_2) &= \left(\frac{1}{13}\right)^2 \text{Var}\left(3X_1+3X_2+\sum_{i=3}^{9} X_i\right)\\ &= \frac{1}{169}\left(3^2\text{Var}(X_1)+3^2\text{Var}(X_2)+\sum_{i=3}^{9}\text{Var}(X_i)\right) = \frac{25}{169}\sigma^2 = 0.148\sigma^2;\\ \text{Var}(\Theta_3) &= \left(\frac{1}{3}\right)^2 \text{Var}\left(3X_1+2X_2+X_1\right)\\ &= \frac{1}{9}\left(3^2\text{Var}(X_1)+2^2\text{Var}(X_2)+\text{Var}(X_1)\right) = \frac{14}{12}\sigma^2 = 1.17\sigma^2. \end{aligned}$$

Dus Θ_1 heeft de kleinste variantie, en zoals we zullen zien, ook de kleinste MSE:

$$\begin{aligned} \text{MSE}(\Theta_1) &= \text{Var}(\Theta_1) + (\text{bias}(\Theta_1))^2 = 0.1\sigma^2;\\ \text{MSE}(\Theta_2) &= \text{Var}(\Theta_2) + (\text{bias}(\Theta_2))^2 = 0.148\sigma^2;\\ \text{MSE}(\Theta_3) &= \text{Var}(\Theta_3) + (\text{bias}(\Theta_3))^2 = 1.17\sigma^2+\theta^2. \end{aligned}$$

In dit geval is Θ_1 *dus altijd nog de beste schatter (de MSE van* Θ_3 *hangt nog af van* θ*, maar zelfs in het beste geval, als* $\theta = 0$ *is deze nog een stuk groter dan de andere twee MSE's).*

In de meeste gevallen is de beste schatter gewoon het gemiddelde over alle gemeten waarden (in ieder geval als de variantie altijd hetzelfde is). Dit gemiddelde hebben wij ook al gebruikt voor Θ_1. De standaard notatie voor het steekproefgemiddelde is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{44}$$

Als een schatter voor de variantie is vaak de steekproefvariantie de beste keuze:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2. \tag{45}$$

3.2 Het betrouwbaarheidsinterval

Het kan zijn dat we het ware gemiddelde μ voor een experiment willen bepalen, maar dat we alleen beschikken over een steekproefgemiddelde $\bar{X}$. Stel dat de X_i onafhankelijk en normaal verdeeld zijn met een onbekende μ en een bekende variantie σ^2. Dan kunnen we een interval bepalen zodat μ met een kans van $100(1-\alpha)\%$ in dit interval ligt. Zo'n interval noemen we een $100(1-\alpha)\%$-betrouwbaarheidsinterval (*-confidence interval*).

Bekende variantie

Voordat we het betrouwbaarheidsinterval kunnen bepalen moeten we eerst de betrouwbaarheidscoëfficiënt $1-\alpha$ weten. Deze geeft aan met welke kans je μ in het (nog te bepalen) interval wil vinden. Verder moeten we ook de uitkomst van het experiment, $\bar{x}$, kennen (hier schrijven we kleine $\bar{x}$, omdat het nu de uitkomst is van een experiment, en geen schatter). Verder is de variantie σ^2 al bekend. Het betrouwbaarheidsinterval wordt dan gegeven door

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{46}$$

In deze formule kunnen we $z_{\alpha/2}$ in een tabel vinden met de methode uit Voorbeeld 18.

Voorbeeld 23 *We willen berekenen wat de gemiddelde kaasconsumptie van de inwoners van Tilburg is. We nemen aan dat de jaarlijkse kaasconsumptie normaal verdeeld is met een variantie van* 3 $kilogram^2$*. We volgen de eetgewoonten van 40 inwoners een jaar lang. Daaruit volgt een steekproefgemiddelde van 14 kilo kaas per jaar.*

Geef het 95%*-betrouwbaarheidsinterval voor de gemiddelde kaasconsumptie.*

Gegeven zijn $\bar{x} = 14$*,* $\sigma^2 = 3$ *en* $\alpha = 0.05$*. In de tabel zien we dat* $z_{\alpha/2} = z_{0.025} = 1.96$*. We stoppen deze waarden in de formule voor het betrouwbaarheidsinterval:*

$$\left[14 - 1.96\frac{\sqrt{3}}{\sqrt{40}}, 14 + 1.96\frac{\sqrt{3}}{\sqrt{40}}\right] = [13.46, 14.54]$$

Dus ligt met een zekerheid van 95% *de gemiddelde kaasconsumptie van inwoners van Tilburg tussen* 13.46 *en* 14.54 *per jaar.*

In het vorige voorbeeld gebruikten we een tweezijdig betrouwbaarheidsinterval. Soms willen we echter alleen een boven- of benedengrens op μ weten. In zo'n geval gebruiken we een eenzijdig betrouwbaarheidsinterval:

$$\mu \leq \bar{x} + z_\alpha\frac{\sigma}{\sqrt{n}}; \tag{47}$$

$$\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}} \leq \mu. \tag{48}$$

Voorbeeld 24 *Voor de straat waar je aan woont is bij wet bepaald dat er een snelheidslimiet van* 30 *km/h opgelegd kan worden als er gemiddeld meer dan honderd voertuigen per uur passeren. We meten gedurende* 25 *uren op verschillende dagen en tijden hoeveel voertuigen er langs komen. Het verkeer lijkt normaal verdeeld te zijn met steekproefgemiddelde van* 105 *voertuigen per uur, en standaardeviatie* $\sigma = 20$. *Kunnen we met de voorgeschreven* 98% *zekerheid zeggen dat er gemiddeld meer dan 100 voertuigen per uur passeren?*

Gegeven zijn $\mu = 100$, $\bar{x} = 105$, $\sigma = 20$ *en* $\alpha = 0.02$. *In de tabel vinden we dat* $z_\alpha = z_{0.02} = 2.06$. *Daarmee berekenen we het van beneden begrensde betrouwbaarheidsinterval:*

$$105 - 2.33\frac{20}{\sqrt{25}} = 95.68 \quad \leq \quad \mu.$$

We kunnen er dus niet zeker van zijn dat het gemiddelde aantal voertuigen meer dan 100 per uur is.

Steekproefgrootte In het vorige voorbeeld was de standaarddeviatie te groot om onze bewering te bewijzen. Als we een dergelijke situatie tegenkomen moeten we de precisie verhogen door meer metingen doen. De steekproefvariantie σ^2/n wordt namelijk kleiner als het aantal metingen n groter wordt. Als we een tweezijdig betrouwbaarheidsinterval willen hebben met een breedte van hooguit $2E$, zodat $\bar{x}$ op een afstand van niet meer dan E van μ ligt, moeten we een steekproefgrootte nemen van tenminste

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2. \tag{49}$$

(In het Engels wordt E ook wel *error* genoemd.) Als we een eenzijdig betrouwbaarheidsinterval willen bepalen zodat $\bar{x}$ niet meer dan E boven of onder μ ligt (één van de twee in dit geval), dan moeten we een steekproefgrootte kiezen van ten minste

$$n = \left(\frac{z_\alpha\sigma}{E}\right)^2. \tag{50}$$

Voorbeeld 25 *We gaan verder met Voorbeeld 24. We willen namelijk weten hoeveel uur we langs de weg moeten meten om er* 98% *zeker van te zijn dat er meer dan 100 voertuigen per uur passeren. We weten dat* $\sigma = 20$ *en dat* $\alpha = 0.02$ *(dus ook nog steeds* $z_\alpha = 2.06$*) en we hebben tot nu toe een steekproefgemiddelde van* 105 *voertuigen per uur. Onze fout E mag dus hooguit 5 zijn. We hebben ook nog steeds te maken met een eenzijdig betrouwbaarheidsinterval. We vullen deze getallen in in formule* (50)*:*

$$n = \left(\frac{z_\alpha\sigma}{E}\right)^2 = \left(\frac{2.06 \times 20}{5}\right)^2 = 67.90. \tag{51}$$

Dus we moeten tenminste 68 uur langs de weg meten om met 98% *zekerheid te bepalen of er meer dan 100 voertuigen per uur passeren of niet.*

Onbekende variantie

In realistische situaties is het aannemelijk dat we de variantie niet kennen. In dat soort gevallen moeten we de schatter voor de variantie S^2 gebruiken (zoals gegeven in formule (45)). Deze schatter voor de variantie is ook afhankelijk van het aantal metingen dat er is uitgevoerd. De normaalverdeling biedt geen mogelijkheid om deze extra informatie te gebruiken, dus moeten we een andere verdeling gebruiken. De verdeling die daarvoor geschikt blijkt is de t-verdeling (*t-distribution*) te zijn.

Deze verdeling lijkt veel op de normaalverdeling, maar heeft een dikkere 'staart', wat wil zeggen dat de kans op een grote afwijking groter is dan bij de normaalverdeling. De staart wordt kleiner naarmate we meer meetpunten hebben. Ook is de vorm van de t-verdeling afhankelijk van het aantal metingen n dat is gedaan (eigenlijk van het aantal $n-1$, dat ook wel het aantal 'vrijheidsgraden' of *degrees of freedom* wordt genoemd). Als het aantal vrijheidsgraden van de t-distributie een grote waarde nadert, begint de t-verdeling steeds meer te lijken op een normaalverdeling.

De formules voor het betrouwbaarheidsinterval lijken ook erg veel op die van het betrouwbaarheidsinterval van een normaal verdeelde stochast. Nu zoeken we alleen in een tabel $t_{\alpha,n-1}$ op in plaats van z_α (dit is dus een andere tabel!). In deze tabel staat bovenaan de waarde van α, en in de meest linker kolom de waarde van $n-1$.

Het tweezijdige betrouwbaarheidsinterval voor een t-verdeling wordt gegeven door

$$\left[\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right], \tag{52}$$

en voor het eenzijdige betrouwbaarheidsinterval hebben we

$$\mu \leq \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}; \tag{53}$$

$$\bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} \leq \mu. \tag{54}$$

In deze formules is s^2 de steekproefvariantie.

Voorbeeld 26 *We willen weten hoeveel een eekhoorntje gemiddeld per dag eet. Het probleem is echter dat we maar weinig afweten van eekhoorntjes, dus we hebben eigenlijk geen enkel idee wat de variantie is. Gedurende 5 dagen volgen we 5 eekhoorntjes in het bos, en zo bepalen we dat het steekproefgemiddelde* 90 *gram per dag is met een steekproefvariantie* s^2 *van 100 gram*^{2}*. We willen het* 95% *betrouwbaarheidsinterval bepalen voor de werkelijke gemiddelde consumptie van een eekhoorntje per dag.*

We hebben in totaal 25 metingen gedaan, dus $n = 25$*, en we hebben* $\bar{x} = 90$ *en* $s = 10$*. Verder hebben we* $\alpha = 0.05$*, dus zoeken we in de tabel de waarde op van* $t_{\alpha/2,n-1} = t_{0.025,24}$*. Dat blijkt 2.064 te zijn. Deze waarden stoppen we in formule* (52)*:*

$$\left[90 - 2.064\frac{10}{\sqrt{25}}, 90 + 2.064\frac{10}{\sqrt{25}}\right] = [85.872, 94.128].$$

Proportie van een populatie

In veel toepassingen willen we weten hoeveel procent van een populatie een bepaalde eigenschap heeft. In deze situaties is de stochast die we willen bestuderen binomiaal verdeeld, want elk lid van de populatie heeft deze eigenschap óf wel, óf niet. We zeggen dat een stochast X de eigenschap heeft met kans p (en de eigenschap dus niet heeft met kans $1-p$). We willen een schatting hebben voor deze p. Als n groot genoeg is (in ons geval is hiervoor de regel dat zowel np als $n(1-p)$ groter zijn dan 5) kunnen we de binomiaalverdeling goed benaderen met een normaalverdeling.

Stel dat we voor een experiment n leden van een populatie testen op een eigenschap, en dat van die n er x zijn met die eigenschap. Dan is onze schatter voor p gegeven door $\hat{p} = x/n$. De steekproef standaarddeviatie is dan $\sqrt{\hat{p}(1-\hat{p})}$. (Hint: als er in een tentamenopgave geen standaarddeviatie of variantie wordt gemeld, lees dan de opgave nog eens goed door. Misschien moet er worden nagegaan of de proportie p en n groot genoeg zijn!)

Voor proporties wordt het tweezijdige betrouwbaarheidsinterval gegeven door

$$\left[\hat{p} - z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right]. \tag{55}$$

Als we alleen willen weten of de proportie groter of kleiner is dan een bepaalde waarde gebruiken we de eenzijdige betrouwbaarheidsintervallen

$$p \leq \hat{p} + z_\alpha\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}; \tag{56}$$

$$\hat{p} - z_\alpha\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq p. \tag{57}$$

Voorbeeld 27 *We willen weten welk percentage van de Nederlandse mannen linkshandig is. Dus we vragen het aan 216 Nederlandse mannen en 18 van hen blijken linkshandig te zijn. Geef het* 95%*-betrouwbaarheidsinterval voor* p*.*

Gegeven zijn $n = 216$*,* $x = 18$ *en* $\alpha = 0.05$*. Dus* $\hat{p} = \frac{18}{216} = \frac{1}{12}$ *en*

$$s = \sqrt{\hat{p}(1-\hat{p})} = \sqrt{\frac{1}{12} \times \frac{11}{12}} = \frac{\sqrt{11}}{12} = 0.276.$$

Verder geldt dat $n\hat{p} = 18$ *en* $n(1-\hat{p}) = 198$*, beide ruim groter dan 5, dus we mogen benaderen met een normaalverdeling. We zoeken in de tabel op dat* $z_{\alpha/2} = z_{0.025} = 1.96$*. Het betrouwbaarheidsinterval wordt dus gegeven door*

$$\left[\frac{1}{12} - 1.96\frac{0.276}{\sqrt{216}}, \frac{1}{12} + 1.96\frac{\sqrt{0.276}}{\sqrt{216}} \right] = [0.046, 0.12].$$

4 Toetsen van hypotheses

Als we willen aantonen dat het gemiddelde van een stochast μ de waarde heeft die we denken dat het heeft (noem die waarde μ_0), dan is de wetenschappelijke aanpak om dat te doen via het formuleren en toetsen van hypotheses. Het is conventie dat onze eerste hypothese, H_0, zal zijn dat het ware gemiddelde hetzelfde is als wat we denken dat het is, dus $\mu = \mu_0$. Voor de alternatieve hypothese, H_1, hebben we dan drie keuzes: We kiezen

1. $H_1 : \mu \neq \mu_0$ wanneer we willen aantonen dat μ exact de waarde heeft die we denken dat hij heeft;
2. $H_1 : \mu \leq \mu_0$ wanneer we willen aantonen dat μ tenminste μ_0 is;
3. $H_1 : \mu \geq \mu_0$ wanneer we willen aantonen dat μ hoogstens μ_0 is.

Als nu het steekproefgemiddelde van onze meting, $\bar{X}$, in het acceptatiegebied (*acceptance region*) ligt, accepteren we de hypothese H_0. (Let op! In hypothese H_0 staat $\mu = \mu_0$, maar alle waarden van μ die in het acceptatiegebied liggen leiden tot acceptatie van H_0.) De keuze van het acceptatiegebied laten we afhangen van onze keuze voor de alternatieve hypothese H_1:

1. als $H_1 : \mu \neq \mu_0$ is het acceptatiegebied het interval $[a, b]$ en accepteren we H_0 als $\bar{X}$ in dit interval ligt;
2. als $H_1 : \mu < \mu_0$ is het acceptatiegebied het interval $[a, +\infty]$ en accepteren we H_0 als $\bar{X}$ groter is dan a.
3. als $H_1 : \mu > \mu_0$ is het acceptatiegebied het interval $[-\infty, b]$ en accepteren we H_0 als $\bar{X}$ kleiner is dan b.

(De waarden van a en b moeten we nog bepalen, maar dat komt later aan de orde.) Het algemene idee van de twee hypotheses is dat wanneer je H_0 afwijst, je dit met een grote zekerheid wil doen. Dus wat we willen vermijden is de zogenoemde type I fout (*type I error*). Deze houdt in dat je de nulhypothese afwijst als die toch correct is. Een fout die minder ernstig is, de zogenoemde type II fout, is dat we de nulhypothese accepteren terwijl die eigenlijk fout is. De kans op een type I fout noemen we ook wel α en de kans op een type II fout β:

$$\alpha = \mathbb{P}(\text{wijs } H_0 \text{ af als het ware gemiddelde } \mu_0 \text{ is}); \tag{58}$$

$$\beta = \mathbb{P}(\text{accepteer } H_0 \text{ als het ware gemiddelde niet } \mu_0 \text{ is}). \tag{59}$$

De kans op een type I fout, α, is makkelijk te berekenen. Om de kans op een type II fout, β, te berekenen moeten we altijd aannemen dat het gemiddelde een andere waarde heeft, noem die θ. Dit betekent dus dat als we een experiment opzetten, de type I fout makkelijk onder controle te houden valt, terwijl dat bij de type II fout moeilijker is (omdat we hiervoor informatie over θ nodig hebben).

Een andere belangrijke grootheid is het onderscheidingsvermogen (*power*) van de toets. Het onderscheidingsvermogen is de kans dat we H_0 afwijzen als H_0 niet waar is. Het onderscheidingsvermogen van een toets geeft dus aan hoe goed de toets is in het afwijzen van een foute H_0. Het onderscheidingsvermogen is dus per definitie $1 - \beta$ (en het is dus ook moeilijk om hier in een experiment controle over te krijgen).

Het berekenen van α Stel dat voor onze toets de X_i normaal verdeeld zijn met een gemiddelde μ_0 en een bekende variantie σ^2. Als we een tweezijdige toets maken, en H_0 alleen accepteren als $\bar{X}$ in het acceptatiegebied $[a, b]$ ligt, dan kunnen we α als volgt berekenen:

$$\begin{aligned} \alpha &= \mathbb{P}(\text{wijs } H_0 \text{ af als het ware gemiddelde } \mu_0 \text{ is}) \\ &= \mathbb{P}(\bar{X} < a) + \mathbb{P}(\bar{X} > b) = \mathbb{P}\left(Z < \frac{a - \mu_0}{\sigma/\sqrt{n}}\right) + \mathbb{P}\left(Z > \frac{b - \mu_0}{\sigma/\sqrt{n}}\right). \end{aligned} \tag{60}$$

Als we een eenzijdige toets uitvoeren en H_0 alleen accepteren als $\bar{X}$ in het acceptatiegebied $[a,+\infty]$ (respectievelijk $[-\infty,b]$) ligt, dan berekenen we α als volgt:

$$\alpha = \mathbb{P}(\bar{X} < a) = \mathbb{P}\left(Z < \frac{a-\mu_0}{\sigma/\sqrt{n}}\right) \tag{61}$$

$$\text{resp.} \quad \alpha = \mathbb{P}(\bar{X} > b) = \mathbb{P}\left(Z > \frac{b-\mu_0}{\sigma/\sqrt{n}}\right). \tag{62}$$

Voorbeeld 28 *Een bank is regelmatig in de gelegenheid goudstaven in bulk te kopen van een investeringsmaatschappij. De goudstaven hebben vaste afmetingen, en horen ook een vast gewicht te hebben van precies* 12400 *gram. Als er echter onzuiverheden in het goud zitten, zal de staaf lichter of zwaarder zijn. We weten dat het gewicht van bona fide staven normaal verdeeld is met* $\sigma^2 = 6.25$ *gram*^{2}*. De bank wil de staven alleen kopen voor de vaste goudprijs als een steekproef van* 16 *staven een gemiddelde heeft van tussen de* 12389 *en* 12411 *gram. Bepaal* α *en concludeer of dit een verstandig beleid is.*

We passen formule (60) *toe met* $\mu_0 = 12400$, $\sigma = 2.5$, $n = 16$ *en* $[a,b] = [12389, 12411]$*:*

$$\begin{aligned}
\alpha &= \mathbb{P}\left(Z < \frac{a-\mu_0}{\sigma/\sqrt{n}}\right) + \mathbb{P}\left(Z > \frac{b-\mu_0}{\sigma/\sqrt{n}}\right) \\
&= \mathbb{P}\left(Z \le \frac{12389-12400}{2.5/4}\right) + \mathbb{P}\left(Z \ge \frac{12411-12400}{2.5/4}\right) \\
&= \mathbb{P}(Z \le -17.6) + \mathbb{P}(Z \ge 17.6) = 0.
\end{aligned}$$

Het interval $[a,b]$ *is zo ruim dat* $\alpha = 0$*, en het zal dus in de praktijk nooit gebeuren dat de hypothese wordt afgewezen. Een slecht beleid dus.*

Het berekenen van β Om β en het onderscheidingsvermogen $1-\beta$ uit te rekenen moeten we de aanname maken dat het gemiddelde van de X_i eigenlijk θ is en dat $\theta \neq \mu_0$. Daarmee kunnen we voor de verschillende scenarios β als volgt berekenen:

$$\beta = \mathbb{P}(\text{accepteer } H_0 \text{ als het ware gemiddelde } \theta \neq \mu_0 \text{ is})$$

$$\text{tweezijdig: } \beta = \mathbb{P}(a < \bar{X} < b) = \mathbb{P}\left(Z < \frac{b-\theta}{\sigma/\sqrt{n}}\right) - \mathbb{P}\left(Z < \frac{a-\theta}{\sigma/\sqrt{n}}\right); \tag{63}$$

$$\text{eenzijdig: } \beta = \mathbb{P}(a > \bar{X}) = \mathbb{P}\left(Z > \frac{a-\theta}{\sigma/\sqrt{n}}\right) \tag{64}$$

$$\text{resp. } \beta = \mathbb{P}(\bar{X} < b) = \mathbb{P}\left(Z < \frac{b-\theta}{\sigma/\sqrt{n}}\right). \tag{65}$$

Voorbeeld 29 *Ernstig geschrokken door onze bevindingen heeft de bank (uit het vorige voorbeeld) besloten tot een aanscherping van de grenzen op het acceptatiegebied. Nu accepteren ze de lading alleen als het steekproefgemiddelde tussen* 12398.2 *gram en* 12401.8 *gram ligt. Ze maken zich echter zorgen dat de investeringsmaatschappij in de afgelopen jaren het goud heeft afgewaterd met een goedkoper metaal. (Eén goudstaaf kosts al gauw* 320,000 *Euro, dus zwendel is lucratief...) Kan de bank met het nieuwe acceptatiegebied voorkomen dat staven met een gemiddeld gewicht van* 12397 *gram voor zuiver worden aangezien? Dat wil zeggen, accepteren ze* H_0 *als* $\mu_0 \neq \theta = 12397$*?*

Er wordt gevraagd wat β *is als* $\theta = 12397$ *gram. Het acceptatiegebied* $[a,b] = [12398.2, 12401.8]$ *en de andere gegevens vullen we in formule* (63) *in:*

$$\begin{aligned}
\beta &= \mathbb{P}\left(Z < \frac{b-\theta}{\sigma/\sqrt{n}}\right) - \mathbb{P}\left(Z < \frac{a-\theta}{\sigma/\sqrt{n}}\right) \\
&= \mathbb{P}\left(Z < \frac{12401.8-12397}{2.5/4}\right) - \mathbb{P}\left(Z < \frac{12398.2-12397}{2.5/4}\right) \\
&= \mathbb{P}(Z < 7.38) - \mathbb{P}(Z < 1.85) = 1 - 0.967 = 0.033.
\end{aligned}$$

Dus de kans dat de investeringsmaatschappij onzuiver goud als zuiver kan verkopen is dus nog steeds 97%.

De P-waarde Eén manier waarop we de resultaten van onze toets kunnen rapporteren, is door te zeggen dat de nulhypothese wel of niet werd afgewezen bij een gegeven waarde van α. Deze arbitraire keuze van α is soms problematisch. Het verschaft ons geen inzicht in hoe ver de gemeten waarde buiten het acceptatiegebied ligt. Ook reduceert het een vraag over de kans tot een vraag met een ja/nee antwoord, namelijk, "wordt de hypothese afgewezen?"

Een manier om dit te omzeilen, is door de beslissing over de hypothese open te laten en in plaats daarvan de P-waarde (*P-value*) te noemen.

Het significantieniveau van een toets is gegeven door de kans dat een meting buiten het acceptatiegebied valt als H_0 klopt. Dus het significantieniveau van een steekproef is simpelweg α.

De P-waarde is het kleinste significantieniveau dat zou leiden tot het afwijzen van de nulhypothese, gegeven een steekproef. Het significantieniveau is dus de α van de kleinste afwijking van μ die we mogen hebben in ons acceptatiegebied, zodanig dat onze steekproef de nulhypothese bevestigd:

$$P = \begin{cases} 2[1-\mathbb{P}(Z \leq |z_0|)] & \text{tweezijdige toets:} & H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0; \\ 1-\mathbb{P}(Z \leq z_0) & \text{van boven begrensde toets:} & H_0 : \mu = \mu_0, H_1 : \mu > \mu_0; \\ \mathbb{P}(Z \leq z_0) & \text{van beneden begrensde toets:} & H_0 : \mu = \mu_0, H_1 : \mu < \mu_0. \end{cases} \tag{66}$$

Hier is z_0 de toetsingsgrootheid en Z_0 de schatter daarvan, dat wil zeggen, de normalisaties van $\bar{x}$ en $\bar{X}$:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}. \tag{67}$$

De P-waarde geeft een getalswaarde aan de de kwaliteit van het bewijs dat we hebben voor het accepteren of afwijzen van de hypothese. In een experiment streven we naar een lage P-waarde.

Voorbeeld 30 *Als $\mu_0 = 10$, $\sigma = 2$, $\bar{x} = 12$ en $n = 16$ dan moet ons acceptatiegebied voor een tweezijdige alternatieve hypothese $H_1 : \mu \neq 10$ dus minimaal $[8,12]$ zijn om de nulhypothese $H_0 : \mu = 10$ te kunnen accepteren. Het significantieniveau dat hierbij hoort is de P-waarde van de toets:*

$$P = 2[1-\mathbb{P}\left(Z \leq \frac{12-10}{2/4}\right)] = 2[1-\mathbb{P}(Z \leq 1)] = 2[1-0.841] = 0.318.$$

5 Toetsen

Tot nu toe hebben we besproken hoe een hypothese opgezet dient te worden, dus hoe H_1 gekozen dient te worden. Ook hebben we gezien hoe we door α en β uit te rekenen in getallen kunnen uitdrukken hoe goed de toets is. Nu zullen we zien hoe we een toets kunnen opzetten zodanig dat onze type I fout hooguit α is.

Om een hypothese te toetsen voeren we een experiment n keer uit om de waarde van $\bar{x}$ te bepalen. Daarna vinden we het $100(1-\alpha)\%$- betrouwbaarheidsinterval voor μ, gebruik makende van de gemeten waarde $\bar{x}$. Als μ nu in het betrouwbaarheidsinterval ligt, kunnen we H_0 accepteren.

We kunnen het ook anders aanpakken door niet het betrouwbaarheidsinterval te berekenen, maar de toetsingsgrootheid z_0 (*test statistic*) te berekenen. Deze waarde is genormaliseerd, dus kunnen we dan kijken of z_0 in het genormaliseerde acceptatiegebied ligt. Deze methode heeft als voordeel dat die vaak iets sneller uit te voeren is, en dat de P-waarde direct duidelijk wordt.

Als er in de opgave niet specifiek om één van de twee methodes gevraagd wordt ben je vrij om zelf te kiezen. Bij veel opgaven is het toetsen echter opgedeeld in stappen, zodat de keuze al voor je gemaakt wordt.

We geven kort uitleg over zes verschillende situaties en we geven voor ieder ook een voorbeeld.

5.1 Toetsen met een bekende variantie

Wanneer we weten dat de uitkomst van onze experimenten normaal verdeeld zijn met een bekende variantie σ^2, dan weten we hoe we het betrouwbaarheidsinterval moeten berekenen.

Om niet in herhaling te vallen, laten we hier een andere aanpak zien dan we eerder hebben gedaan. We gaan gebruik maken van de toetsingsgrootheid.

De toetsingsgrootheid wordt gegeven door:

$$z_0 \quad = \quad \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}. \tag{68}$$

Verder berekenen we het genormaliseerde acceptatiegebied. Deze hangt af van onze keuze voor H_1:

$$\text{tweezijdig: } H_1 : \mu \neq \mu_0 \qquad -z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}; \tag{69}$$

$$\text{van boven begrensd: } H_1 : \mu < \mu_0 \qquad z_0 \geq -z_\alpha; \tag{70}$$

$$\text{van beneden begrensd: } H_1 : \mu > \mu_0 \qquad z_0 \leq z_\alpha. \tag{71}$$

We accepteren H_0 als z_0 nu in het acceptatiegebied ligt voor een gegeven α.

De toetsingsgrootheid z_0 vertelt ons ook wat de P-waarde is die hoort bij $\bar{x}$. De P-waarde is de α zodanig dat $|z_0| = z_{\alpha/2}$ (voor een tweezijdig interval) of $|z_0| = z_\alpha$ (voor een eenzijdig interval).

Voorbeeld 31 *We willen toetsen met een zekerheid van* 95% *of de gemiddelde lengte van een volwassen Franse man* 1.76 *meter is. We weten dat lichaamslengte normaal verdeeld is met een standaarddeviatie van* $\sigma = 8$ *cm. We meten* 100 *Franse mannen en komen uit op een gemiddelde lengte van* 1.73 *meter.*

Uit de tekst maken we op dat $\mu_0 = 176$ *cm,* $\sigma = 8$ *cm,* $\bar{x} = 173$ *cm,* $n = 100$ *en* $\alpha = 0.05$ *(Merk op dat we eerst de gegeven eenheden gelijk moeten maken, in dit geval dus de lengtes allemaal omrekenen naar meters of centimeters.) Omdat we hier een waarde willen verifiëren gebruiken we een tweezijdige toets:*

$$\begin{aligned} H_0 &: \quad \mu = \mu_0; \\ H_1 &: \quad \mu \neq \mu_0. \end{aligned}$$

We kennen de variantie, en we gebruiken een tweezijdige toets, dus kunnen we opzoeken dat $z_{\alpha/2} = z_{0.025} = 1.96$. *Nu schrijven we de toetsingsgrootheid:*

$$z_0 \quad = \quad \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{173-176}{8/\sqrt{100}} = -3.75. \tag{72}$$

Het acceptatiegebied is $-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}$. *Maar* z_0 *ligt niet in dit gebied dus wijzen we* H_0 *af.*

We kunnen het probleem ook oplossen door direct naar het betrouwbaarheidsinterval te kijken:

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{73}$$

$$= \left[173 - 1.96\frac{8}{\sqrt{100}}, \bar{x} + 1.96\frac{8}{\sqrt{100}}\right] = [171.4, 174.5]. \tag{74}$$

Ook hier ligt μ_0 *niet in het acceptatiegebied, en we wijzen dus* H_0 *ook hier af.*

Voorbeeld 32 *Met een zekerheid van* 90% *willen we weten of de gemiddelde lengte van een sneeuwman kleiner is dan* 140 *cm. We bellen het CBS en horen daar dat de variantie* 100 cm^2 *is. In de winter gaan we erop uit, en meten we* 56 *sneeuwmannen. Het gemiddelde van die metingen is* 135 *cm.*

Uit de tekst maken we op dat $\mu_0 = 140$ *cm,* $\sigma = 10$ *cm,* $\bar{x} = 135$, $n = 56$ *en* $\alpha = 0.1$. *We willen dus bepalen of sneeuwmannen kleiner dan* 140 *cm zijn, dus gebruiken we een eenzijdige toets met hypothese*

$$\begin{aligned} H_0 &: \quad \mu = \mu_0; \\ H_1 &: \quad \mu < \mu_0. \end{aligned}$$

We beginnen met het berekenen van het betrouwbaarheidsinterval voor de gegeven data. Omdat het een eenzijdige toets is berekenen we

$$\mu \leq \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} = 135 + z_{0.1}\frac{\sqrt{100}}{\sqrt{56}} = 135 + 1.28\frac{10}{\sqrt{56}} = 136.1. \tag{75}$$

Dus we weten dat het ware gemiddelde met een zekerheid van 90% *kleiner is dan* 136.1 *cm, en dus wijzen we de hypothese af.*

Nog een keer hetzelfde probleem, met de toetsingsgrootheid:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{135 - 140}{\sqrt{100}/\sqrt{56}} = -3.74. \tag{76}$$

Het acceptatiegebied is $z_0 \geq -z_\alpha = -1.28$. *Sinds* $z_0 = -3.74$ *kleiner is dan* -1.28 *wijzen we* H_0 *weer af, dus de gemiddelde sneeuwman is kleiner dan* 140 *cm.*

5.2 Toetsen met een onbekende variantie

Omdat we de variantie niet kennen moeten we de t-verdeling gebruiken. We kunnen het betrouwbaarheidsinterval dus uitrekenen zoals we dat eerder hebben gedaan voor de t-verdeling, of we rekenen de toetsingsgrootheid uit:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{77}$$

waar s^2 de steekproefvariantie is van onze meting. Het t-toets acceptatiegebied wordt dus gegeven door

$$\text{tweezijdig: } H_1 : \mu \neq \mu_0 \qquad -t_{\alpha/2,n-1} \leq t_0 \leq t_{\alpha/2,n-1}; \tag{78}$$

$$\text{van boven begrensd: } H_1 : \mu < \mu_0 \qquad t_0 \geq -t_{\alpha,n-1}; \tag{79}$$

$$\text{van beneden begrensd: } H_1 : \mu > \mu_0 \qquad t_0 \leq t_{\alpha,n-1}. \tag{80}$$

Ook kunnen we weer kijken of t_0 in het genormaliseerde acceptatiegebied ligt om te bepalen of we H_0 accepteren.

Voorbeeld 33 *Stel dat je net je master hebt afgerond en nu op zoek bent naar een baan als architect. In het sollicitatiegesprek krijg je een baan aangeboden met een salaris van* 45,000 *Euro per jaar, met de opmerking dat dit het gemiddelde salaris is voor iemand met een Master in bouwkunde. Je wil controleren of dit waar is, dus je vraagt* 10 *willekeurig gekozen ex-studiegenoten naar hun eerste salaris. Daaruit blijkt dat het gemiddelde* 50,000 *Euro per jaar is, met een steekproef standaardeviatie van* 2,500 *Euro. Als je met een zekerheid van, zeg,* 97.5% *kan aantonen dat het gemiddelde salaris meer is dan* 45,000 *Euro, dan kun je die statistiek laten zien aan het bedrijf waar je solliciteert, om een beter salaris te bedingen (hoewel de wijsneus uithangen in dit soort situaties meestal niet echt in je voordeel werkt...).*

Probeer zelf anders eerst eens het antwoord te vinden voor je verder leest.

Uit de tekst kunnen we de volgende waarden halen: $\mu_0 = 45{,}000$, $s = 2500$, $\bar{x} = 50{,}000$, $n = 10$ *en* $\alpha = 0.025$. *We willen laten zien dat het ware gemiddelde hoger ligt dan* μ_0. *Dit betekent dat onze hypothese als volgt is*

$$\begin{aligned} H_0 &: \quad \mu = \mu_0 \\ H_1 &: \quad \mu > \mu_0 \end{aligned}$$

We beginnen met het berekenen van het 97.5%*-betrouwbaarheidsinterval voor* μ *gegeven* $\bar{x}$. *We kennen de variantie niet, dus zoeken we op dat* $t_{\alpha,n-1} = t_{0.025,9} = 2.262$ *en we berekenen*

$$\bar{x} - t_{\alpha,n-1}\frac{\sigma}{\sqrt{n}} \quad = \quad 50000 - 2.262\frac{2500}{\sqrt{10}} = 50000 - 1788 = 48211 \geq \mu$$

onze $\mu_0 = 45000$ *is een stuk kleiner dan* 48211 *dus wijzen we de hypothese* H_0 *af.*

De andere variant van dezelfde berekening:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{50000 - 45000}{2500/\sqrt{10}} = 6.32.$$

Het acceptatiegebied voor t_0 *is eenzijdig met*

$$\begin{aligned} H_1 : \mu > \mu_0 \qquad & \qquad t_o \leq t_{\alpha,n-1} \\ \text{en } t_0 = 6.32 \quad \not\leq \quad & 2.25 = t_{0.025,9} \end{aligned}$$

Dus wijzen we nogmaals H_0 *af.*

De bewering in het sollicitatiegesprek klopt dus niet.

5.3 Toetsen voor de proportie van een populatie

We kunnen in plaats van het gemiddelde μ natuurlijk ook toetsen of een polulatieproportie p_0 klopt of niet. We hebben al gezien hoe we voor een populatieproportie het betrouwbaarheidsinterval berekenen. We gebruikten daarvoor de variantie $\sigma^2 = \hat{p}(1-\hat{p})$ en $z_{\alpha/2}$. In deze situatie gebruiken we de toetsingsgrootheid

$$z_0 \quad = \quad \frac{\bar{x} - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{81}$$

en gebruiken we hetzelfde acceptatiegebied als in sectie 5.1.

Voorbeeld 34 *We gaan door met de gegevens van Voorbeeld 27. Daar wilden we weten wat de proportie van linkshandige Nederlandse mannen is met een zekerheid van* 95%. *We vroegen het aan* 216 *Nederlandse mannen en* 18 *van hen waren linkshandig. Nu willen we de bewering toetsen dat de ware proportie* 11% *is. In Voorbeeld 27 zagen we al dat* 0.11 *in het betrouwbaarheidsinterval ligt, dus kunnen we* H_0

gewoon accepteren. Om het verhaal compleet te maken berekenen we het nog een keer, maar nu gebruik makend van de toetsingsgrootheid:

$$z_0 = \frac{\bar{x} - np_0}{\sqrt{np_0(1-p_0)}} = \frac{18 - 216 \times 0.11}{\sqrt{216 \times 0.11 \times 0.89}} = -1.25$$

De acceptatiecriteria zijn dat $-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}$ *met* $z_{\alpha/2} = 1.96$. *Dus ook met deze methode accepteren we* H_0.

5.4 Vergelijken van twee populaties met bekende varianties

Soms willen we twee experimenten met toevallige uitkomsten X en Y vergelijken. We zijn voornamelijk geïnteresseerd in het verschil tussen de gemiddelden μ_X en μ_Y.

Hier gaan we ervan uit dat we de varianties σ_X^2 en σ_Y^2 van X en Y kennen. Het eerste experiment voeren we n_X keer uit en dit levert het steekproefgemiddelde $\bar{X}$ op. Het tweede experiment voeren we n_Y keer uit, met een steekproefgemiddelde van $\bar{Y}$. De gezamenlijke variantie van $\bar{X} - \bar{Y}$ wordt dan gegeven door

$$\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}. \tag{82}$$

We kunnen dus het $100(1-\alpha)\%$-betrouwbaarheidsinterval voor $\mu_X - \mu_Y$ als volgt berekenen:

$$\left[\bar{x} - \bar{y} - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}, \bar{x} - \bar{y} + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right]. \tag{83}$$

De situatie voor eenzijdige betrouwbaarheidsintervallen is analoog.
De toetsingsgrootheid wordt gegeven door

$$z_0 = \frac{\bar{x} - \bar{y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \tag{84}$$

en we gebruiken dezelfde acceptatiegebieden als in sectie 5.1.

Voorbeeld 35 *Er zijn twee dartspelers, meneer Iks en meneer Eij. Wij denken dat meneer Iks een betere dartspeler is, wat wil zeggen, hij gooit een hoger gemiddelde. Het verschil tussen de twee is alleen zo klein dat we er wel* 95% *zeker van willen zijn (er staan reputaties op het spel). Daarom laten we beide heren* 10 *rondjes spelen. Meneer Iks heeft een gemiddelde score van* 46 *punten, en meneer Eij scoort gemiddeld* 44 *punten. We hebben deze darters al een tijdje in de gaten gehouden, en zo weten we dat de variantie in de scores van meneer Iks* 10 *is, en van meneer Eij* 8.

Vanwege hun namen noemen we het steekproefgemiddelde van meneer Iks maar $\bar{x}$, *zodat* $\bar{x} = 46$. *Meneer Eij's steekproefgemiddelde is dan* $\bar{y} = 44$. *Verder weten we dat* $\sigma_X^2 = 10$, $\sigma_Y^2 = 8$ *en* $\alpha = 0.05$. *We willen laten zien dat Iks de betere speler is, dus doen we een eenzijdige toets, waar we willen laten zien dat* $\mu_X - \mu_Y$ *groter is dan 0.*

$$\begin{aligned} H_0 &: \quad \mu_x - \mu_y = 0; \\ H_1 &: \quad \mu_x - \mu_y > 0. \end{aligned}$$

We zoeken op dat $z_\alpha = z_{0.05} = 1.65$ *en berekenen dan het linker betrouwbaarheidsinterval*

$$\bar{x} - \bar{y} - z_\alpha\sqrt{\frac{\sigma_X^2}{n_X} - \frac{\sigma_Y^2}{n_Y}} = 46 - 44 - 1.65\sqrt{\frac{10}{10} + \frac{8}{10}} = -0.21 \leq \mu_x - \mu_y.$$

We moeten H_0 *dus accepteren, en we weten dus nog steeds niet wie de betere darter is.*

Als we de toetsingsgrootheid willen gebruiken berekenen we

$$z_0 = \frac{46-44}{\sqrt{\frac{10}{10}+\frac{8}{10}}} = 1.49.$$

Merk op dat het acceptatiecriterion dan $z_0 \leq z_\alpha = 1.65$ *is, dus ook hier moeten we* H_0 *accepteren.*

5.5 Vergelijken van twee populaties met onbekende varianties

Ook bij het vergelijken van twee populaties kan het het geval zijn dat we niet precies de varianties kennen van de experiment X en Y. Als dit het geval is moeten we ons afvragen of de twee varianties gelijk zijn of niet. Wij zullen hier alleen de situatie behandelen dat dat zo is, maar om goed voorbereid te zijn op het tentamen is het misschien ook belangrijk om te weten wat er moet gebeuren als ze niet hetzelfde zijn (zie hiervoor pagina 358 van Montgomery & Runger).
Wij nemen dus aan dat beide eenzelfde variantie zullen hebben: $\sigma_X^2 = \sigma_Y^2$. We kunnen dan de schatters voor de variantie, S_X^2 en S_Y^2 gebruiken om de gezamenlijke schatter (*pooled estimator*) S_p^2 te bepalen:

$$S_p^2 = \frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2} \tag{85}$$

Met deze schatter kunnen we het betrouwbaarheidsinterval voor $\mu_X - \mu_Y$ dan ook bepalen:

$$\left[\bar{x}-\bar{y}-t_{\alpha/2,n_X+n_Y-2}S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}, \bar{x}-\bar{y}+t_{\alpha/2,n_X+n_Y-2}S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}\right] \tag{86}$$

en op een zelfde manier kunnen de eenzijdige varianten berekend worden. De toetsingsgrootheid is

$$t_0 = \frac{\bar{x}-\bar{y}-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \tag{87}$$

en we gebruiken t_{α,n_X+n_Y-2} (resp. $t_{\alpha/2,n_X+n_Y-2}$) voor dezelfde acceptatiegebieden als in sectie 5.2.

Voorbeeld 36 *We hebben een kikker en een pad, en we willen weten welke gemiddeld verder springt. Voor de kikker meten we* 16 *sprongen en* 9 *voor de pad. De kikker sprong (noem dit* X*) gemiddeld* 30 *cm, met een steekproef standaarddeviatie van* 10*. De pad sprong (noem dit* Y*) gemiddeld* 27 *cm met een steekproef standaarddeviatie van* 6 *cm. Is er voldoende bewijs voor de bewering dat met een type I fout van maximaal* 5% μ_x *en* μ_Y *hetzelfde zijn?*

Uit de tekst halen we $\bar{x} = 30$, $S_X = 10$, $n_x = 16$, $\bar{y} = 27$, $S_Y = 6$, $n_y = 9$ *en* $\alpha = 0.05$*. We willen aantonen dat ze gemiddeld even ver springen dus kiezen we*

$$H_0: \quad \mu_x - \mu_y = 0;$$
$$H_1: \quad \mu_x - \mu_y \neq 0.$$

De gezamenlijke variantie S_p^2 *is dan*

$$S_p^2 = \frac{(16-1)10^2 + (9-1)6^2}{16+9-2} = 77.74.$$

De toetsingsgrootheid wordt gegeven door

$$t_0 = \frac{30-27}{\sqrt{77.74}\sqrt{\frac{1}{16}+\frac{1}{9}}} = 0.81.$$

De acceptatiecriteria zijn $-t_{\alpha/2,n-1} \leq t_0 \leq t_{\alpha/2,n-1}$ *en we vinden in de tabel dat* $t_{\alpha/2,n_X+n_Y-2}$=$t_{0.025,23} = 2.069$*. Dus* t_0 *volstaat om* H_0 *te accepteren.*

5.6 Vergelijken van twee populatieproporties

Het laatste dat we zullen doen is het vergelijken van twee proporties, p_X en p_Y. In sectie 5.3 geven we de variantie van proporties:

$$\sigma_X^2 = p_X(1-p_X) \qquad \sigma_Y^2 = p_Y(1-p_Y). \tag{88}$$

We gebruiken hier de formule

$$\sigma_{XY} = \sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{p_X(1-p_X) + p_Y(1-p_Y)} \tag{89}$$

voor de standaarddeviatie voor $p_X - p_Y$. Met deze σ_{XY} kan het betrouwbaarheidsinterval $p_X - p_Y$ berekend worden

$$\left[\hat{p}_X - \hat{p}_Y - z_{\alpha/2}\sigma_{XY}, \hat{p}_X - \hat{p}_Y + z_{\alpha/2}\sigma_{XY}\right]. \tag{90}$$

waar $\hat{p}_X$ en $\hat{p}_Y$ de schatters zijn voor de populatie proportie. De toetsingsgrootheid wordt gegeven door

$$z_0 = \frac{\hat{p}_X - \hat{p}_Y - (p_X - p_X)}{\sigma_{XY}} \tag{91}$$

en we gebruiken het dan acceptatiegebied zoals gegeven in sectie 5.1.

Voorbeeld 37 *We willen weten of vrouwen vaker linkshandig zijn dan mannen. We hadden het al aan* 216 *Nederlandse mannen gevraagd, en van hen waren* 18 *linkshandig. Nu vragen we het ook aan* 104 *Nederlandse vrouwen, en* 15 *van hen blijken linkshandig te zijn.*
We schrijven p_X voor de ware proportie van mannen die linkshandig zijn, en p_Y voor de ware proportie van linkshandige vrouwen.

Uit de tekst kunnen we de schatters voor de proportie opmaken: $\hat{p}_X = 18/216$ en $\hat{p}_Y = 15/104$. We willen nagaan of ze hetzelfde zijn met een zekerheid van 98%*. We gebruiken dus een tweezijdige toets met*

$$\begin{aligned} H_0 &: \quad p_X - p_Y = 0; \\ H_1 &: \quad p_X - p_Y \neq 0. \end{aligned}$$

We beginnen met de variantie te berekenen:

$$\sigma_{XY}^2 = \frac{\frac{18}{216}\left(1-\frac{18}{216}\right)}{216} + \frac{\frac{15}{104}\left(1-\frac{15}{104}\right)}{104} \approx 0.0015.$$

De toetsingsgrootheid wordt dan gegeven door

$$z_0 \quad = \quad \frac{\hat{p}_X - \hat{p}_Y - (p_X - p_Y)}{\sigma_{XY}} = \frac{\frac{18}{216} - \frac{15}{104}}{\sqrt{0.0015}} = -1.55.$$

Het acceptatiecriterion is dan $-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}$ en in ons geval hebben we $\alpha = 0.02$. We zoeken de waarde op: $z_{\alpha/2} = 2.33$.
Als z_0 tussen $-z_{\alpha/2}$ en $z_{\alpha/2}$ ligt moeten we H_0 accepteren. Dat zou betekent dus dat Nederlandse mannen en Nederlandse vrouwen dezelfde kans hebben om linkshandig te zijn.
Als laatste berekenen we ook nog even het betrouwbaarheidsinterval voor $p_x - p_y$:

$$\left[\frac{18}{216} - \frac{15}{104} - 2.33 \times 0.039, \frac{18}{216} - \frac{15}{104} + 2.33 \times 0.039\right] = [-0.15, 0.03]. \tag{92}$$

5.7 Gepaarde t-toets

Het kan voorkomen dat we observaties doen in paren. Denk bijvoorbeeld aan een observatie vóór een handeling en ook één er na. Of de stand van de AEX en de Dow Jones op hetzelfde tijdstip. Gepaarde observaties zijn interessant, omdat elk paar waarneming onder dezelfde omstandigheden gedaan wordt,

terwijl die omstandigheden wel tussen verschillende paren kan verschillen. Stel dat we een n paren observeren, (X_i, Y_i) voor $i = 1, 2, \ldots, n$. Dan schrijven we D_i voor het verschil tussen de waarden van het paar (X_i, Y_i). We zijn nu geïnteresseerd in de gemiddelde waarde $\bar{D}$ van de waarden D_i. We nemen aan dat de gemiddelde waarde van de populatie van X gegeven wordt door μ_X met variantie σ_X^2 en de gemiddelde waarde van de populatie van Y gegeven wordt door μ_Y met variantie σ_Y^2. We nemen aan dat de D_i's normaal verdeeld zijn, dus is het gemiddelde

$$\mu_D = \mathbb{E}[X - Y] = \mathbb{E}[X] - \mathbb{E}[Y] = \mu_X - \mu_Y. \tag{93}$$

en de variantie σ_D^2 heeft schatter

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2. \tag{94}$$

We gebruiken de t-distributie als we gebruik maken van de steekproefvariantie. Het tweezijdige $100(1-\alpha)\%$-betrouwbaarheidsinterval op het verschil in gemiddelden μ_D wordt gegeven door

$$\bar{d} - t_{\alpha/2,n-1} \frac{s_D}{\sqrt{n}} \leq \mu_D \leq \bar{d} + t_{\alpha/2,n-1} \frac{s_D}{\sqrt{n}} \tag{95}$$

waar $t_{\alpha/2,n-1}$ op dezelfde manier wordt bepaald als in Sectie 5.2. De bijbehorende toetsingsgrootheid is

$$T_0 = \frac{\bar{D} - (\mu_X - \mu_Y)}{S_D/\sqrt{n}}. \tag{96}$$

Voorbeeld 38 *We vragen ons af of het echt waar is dat je 's ochtends een centimeter langer bent dan 's avonds. Dus meten we 10 mensen zowel 's ochtends als 's avonds. Hieruit volgt dat ons steekproefgemiddelde van het verschil van* 0.7 *cm, en de steekproef standaarddeviatie* 0.2 *cm. We willen met* 95% *zekerheid weten dat de krimp inderdaad* 1 *cm bedraagt.*

We weten $\bar{d} = 0.7$, $\mu_D, 0 = 1$, $s_D = 0.2$, $n = 10$ *en* $\alpha = 0.05$. *Onze hypothese* $H_0 : \mu_D = 1$ *met alternatieve hypothese* $H_1 : \mu_D \neq 1$. *Uit de tabel voor de t-distributie halen we de waarde*

$$t_{\alpha/2,n-1} = t_{0.025,9} = 2.262.$$

Onze toetsingsgrootheid is

$$t_0 = \frac{0.7 - 1}{0.3/\sqrt{1}} = -3.16.$$

Omdat het niet zo is dat $-t_{\alpha/2,n-1} \leq t_0 \leq t_{\alpha/2,n-1}$, *verwerpen we de nulhypothese.*
We kunnen ook het acceptatiegebied uitrekenen:

$$\begin{aligned} 0.6 - 2.262\frac{0.3}{\sqrt{10}} &\leq \mu_D \leq 0.6 + 2.262\frac{0.3}{\sqrt{10}} \\ 0.385 \leq &\quad \mu_D \quad \leq 0.815. \end{aligned}$$

Ook hier geld dus weer dat 1 *niet tussen* 0.385 *en* 0.815 *ligt, en dus verwerpen we ook met deze methode de nulhypothese.*