Wegwijzer in de statistiek

Download gratis op Studentensupport.nl

In samenwerking met
Selexyz.nl

Leo Strijbosch

Wegwijzer in de statistiek

Studentensupport

Studentensupport.nl

Wegwijzer in de statistiek

Download gratis op www.studentensupport.nl
ISBN 87-7681-158-1

Studentensupport
Studentensupport.nl

Inhoudsopgave

Voorwoord

Statistiek geldt voor veel studenten wereldwijd als een struikelblok. In veel eerstejaars programma's worden, doorgaans omvangrijke, Engelse tekstboeken gebruikt met uitvoerige uitleg en veel voorbeelden. Dit compendium, met name bedoeld voor studenten economie, en sociale en medische wetenschappen, beoogt in eerste instantie naast de gebruikte leerboeken hulp te bieden door (een belangrijk deel van de) gebruikelijke stof op compacte en overzichtelijke wijze te presenteren. Ook het feit dat het een Nederlandse tekst is, kan voor veel studenten aantrekkelijk zijn. De gebruikelijke statistische methoden en toetsen worden na een korte uitleg op een kookboekachtige wijze gepresenteerd, waarbij doorgaans één of meer voorbeelden de werkwijze nog eens illustreren. De diverse appendices vergroten bovendien de toegankelijkheid van de aangeboden leerstof.

1. Grondbegrippen van de kansrekening

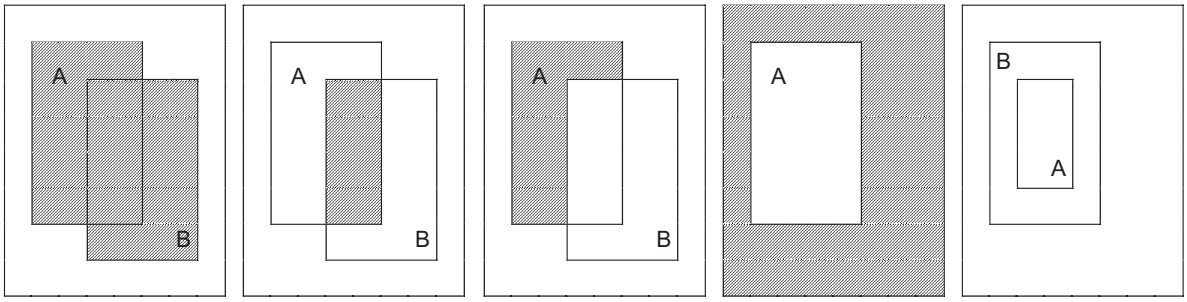
1.1 Inleiding

Statistiek gaat over het verkrijgen en gebruiken van informatie indien er sprake is van onzekerheid. De moderne kansrekening is gebaseerd op het model van Kolmogorov (1933) en de daarvan afgeleide theoretische waarschijnlijkheidsleer. Deze verschaft een (kans)model voor de situatie dat vergelijkbare oorzaken een aantal verschillende gevolgen kan hebben. Zo kunnen 2 worpen met een munt (ook al proberen we de manier van werpen gelijk te houden) 2 verschillende uitkomsten opleveren. Voor het modelleren van dit soort *random* (of: *stochastische*) *experimenten* hebben we een zgn. kansruimte nodig. Voor een algemene beschrijving introduceren we de volgende symbolen in de vorm van voorbeelden (elementaire kennis van de verzamelingenleer wordt bekend verondersteld).

$A \cup B$	**vereniging** van A en B	(alle elementen uit A en/of B)
$A \cap B$	**doorsnede** van A en B	(alle elementen die zowel in A als in B zitten)
A^c	**complement** van A	(alle elementen die niet in A zitten)
$A - B$	verschil van A en B	(alle elementen die wel in A maar niet in B zitten)
$A \subset B$	A is een **deelverzameling** van B	(alle elementen uit A zitten ook in B)

Van John Venn (1834-1923) is het zogenaamde **Venndiagram** afkomstig dat gebruikt kan worden om allerlei rekenregels en eigenschappen van verzamelingen grafisch zichtbaar te maken. Deze techniek is in Figuur 1.1 gebruikt om bovenstaande eigenschappen te verduidelijken.

Figuur 1.1:
Venndiagrammen



1.2 Kansruimte, kansfunctie, uitkomstenruimte, gebeurtenis

Voor een wiskundige beschrijving van een **random experiment** hebben we de volgende begrippen nodig.

Uitkomstenruimte:
De verzameling van alle mogelijke uitkomsten van een random experiment. We gebruiken hiervoor het symbool Ω. Het complement van Ω is leeg en noteren we met $\varnothing$; dus $\Omega^c = \varnothing$.

Uitkomst, **elementaire gebeurtenis**:
Elk element van Ω is een mogelijke *uitkomst*. Vaak wordt hiervoor het symbool ω gebruikt. Voor "is een element van" wordt de notatie "$\in$" gebruikt (dus: $\omega \in \Omega$).

Gebeurtenis:
Deelverzamelingen van Ω heten gebeurtenissen.

Een **kansruimte** is een paar (Ω, P) bestaande uit een verzameling Ω en een **kansfunktie** P, die aan iedere deelverzameling $A \subset \Omega$ een reëel getal $P(A)$ op het interval [0;1] toevoegt zodanig dat aan de volgende twee axioma's is voldaan.

Axioma's kansruimte
1. $P(\Omega) = 1$
2. $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$ indien de deelverzamelingen $A_1, A_2, \ldots$ van Ω paarsgewijs disjunct zijn (elkaar uitsluiten).

Twee gebeurtenissen A en B heten **disjunct** als $A \cap B = \varnothing$. Voor een gebeurtenis A heet $P(A)$ de **kans** op A. Uit de twee axioma's kunnen we (o.a.) de volgende eigenschappen afleiden.

Als $A \subset B$ dan $P(B-A) = P(B) - P(A) = P(A^c \cap B)$

$P(A^c) = 1 - P(A)$; bijzonder geval voor $A = \Omega : P(\varnothing) = 0$

Als $A \subset B$ dan $P(A) \leq P(B)$

$P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n)$

$P\left(\bigcap_{n=1}^{k} A_n\right) \geq 1 - \sum_{n=1}^{k} P(A_n^c)$, de ongelijkheid van Bonferroni

Als $A_1,...,A_n$ paarsgewijs disjunct zijn dan $P(A_1 \cup ... \cup A_n) = P(A_1) + ... + P(A_n)$

Een speciaal geval is de situatie dat Ω een eindig aantal elementen bevat, zeg N, die alle met gelijke kans voorkomen. In dat geval geldt $P(\omega) = 1/N$ voor iedere $\omega \in \Omega$. Bovendien geldt dat $P(A) = \frac{|A|}{N}$, waarbij $|A|$ het aantal elementen in A is.

Voorbeeld 1.1 Beschouw de verzameling $\Omega = \{1,2,3,4,5,6\}$. Definieer voor iedere deelverzameling $A \subset \Omega$

$$P(A) = \frac{|A|}{6}$$

Het paar (Ω, P) is dus een kansruimte dat model kan staan voor “een worp met een dobbelsteen”.

Voorbeeld 1.2 Beschouw nu de verzameling $\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$. Definieer voor iedere deelverzameling $A \subset \Omega$

$$P(A) = \frac{|A|}{36}$$

Het paar (Ω, P) staat nu model voor “twee worpen met een dobbelsteen”. De deelverzameling $A = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$ is de gebeurtenis “twee gelijke ogen”.

1.3 Voorwaardelijke kans en de regel van Bayes

Voor twee gebeurtenissen A en B wordt de **voorwaardelijke kans** op A gegeven B (met $P(B) > 0$) gedefinieerd als

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Losjes gezegd is $P(A \mid B)$ de kans dat A optreedt indien we al weten dat B opgetreden is. De vorige definitie anders geschreven levert de zogenaamde **vermenigvuldigingsregel** (**productregel**) op:

$$P(A \cap B) = P(B) \times P(A \mid B) = P(A) \times P(B \mid A)$$

Voorbeeld 1.3 In de finale van het wereldkampioenschap voetbal 2006 speelt Italië tegen de winnaar van de halve finale tussen Portugal en Frankrijk. Een bookmaker schat de kans dat Portugal de halve finale wint op 60%. De kans dat Italië Portugal in de finale verslaat wordt geschat op 20%, terwijl de kans dat Italië Frankrijk in de finale verslaat wordt geschat op 30%. De bookmaker berekent, gebruik makend van bovenstaande stelling de kans dat Italië de finale wint als volgt

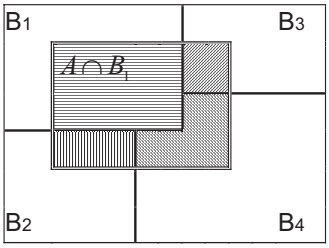
$$
\begin{aligned}
&P(\text{Italië wint finale}) = \\
&= P(\text{Portugal wint halve finale}) \times P(\text{Italië wint finale} \mid \text{Portugal wint halve finale}) + \\
&\quad P(\text{Frankrijk wint halve finale}) \times P(\text{Italië wint finale} \mid \text{Frankrijk wint halve finale}) = \\
&= 0{,}6 \times 0{,}2 + 0{,}4 \times 0{,}3 = 24\%
\end{aligned}
$$

Een gevolg hiervan is de **regel van Bayes**

$$P(B \mid A) = \frac{P(A \mid B) \times P(B)}{P(A)}$$

Als $B_1,...,B_n$ een zogenaamde **dissectie** is, d.w.z. als $B_1,...,B_n$ paarsgewijs disjuncte deelverzamelingen van Ω zijn die met elkaar verenigd weer Ω vormen (dus $\bigcup_{i=1}^{n} B_i = \Omega$) en die elk een positieve kans hebben ($P(B_i) > 0, \quad i = 1,...,n$), dan geldt voor iedere gebeurtenis A

$$A = \bigcup_{i=1}^{n} (A \cap B_i)$$ (in voorbeeld hiernaast is $n = 4$)



en

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \times P(B_i)$$

en, door combinatie van de vorige resultaten, de meer uitgebreide vorm van de **regel van Bayes**

$$P(B_i \mid A) = \frac{P(A \mid B_i) \times P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j) P(B_j)}$$

1.4 Onafhankelijke gebeurtenissen

Twee gebeurtenissen A en B heten **onafhankelijk** als ('**productregel voor kansen**')

$$P(A \cap B) = P(A) \times P(B)$$

Equivalent hiermee is de voorwaarde $P(A \mid B) = P(A)$, dus de kans op A is gelijk aan de kans op A gegeven B.

Twee gebeurtenissen zijn onafhankelijk als de kans op de ene gebeurtenis niet afhangt van de wetenschap dat de andere gebeurtenis (wel of niet) plaats vindt.

Voorbeeld 1.4 We gooien een rode en een zwarte dobbelsteen. Beschouw de gebeurtenissen

A: de rode dobbelsteen toont een 6
B: de zwarte dobbelsteen toont een 6

Aangezien

$$P(A \cap B) = \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = P(A) \times P(B)$$

zijn A en B onafhankelijk. De kans dat de worp met de rode dobbelsteen een 6 oplevert, wordt niet beïnvloed door het resultaat van de worp met de zwarte dobbelsteen.

Voorbeeld 1.5 We gooien een rode en een zwarte dobbelsteen. Beschouw de gebeurtenissen

A: de rode en de zwarte dobbelsteen geven hetzelfde aantal ogen
B: het aantal ogen van de rode en de zwarte dobbelsteen is samen 10

Aangezien $P(A)=\frac{1}{6}$, maar, vanwege $B=\{(4,6),(5,5),(6,4)\}$, $P(A\mid B)=\frac{1}{3}$, zijn A en B **afhankelijk**. De kans op een gelijk aantal ogen neemt toe als we weten dat de som van het aantal ogen 10 is.

1.5 De somregel voor kansen

Voor 2 willekeurige gebeurtenissen A en B geldt (**somregel**)

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

Omdat door de optelling van $P(A)$ en $P(B)$ de term $P(A\cap B)$ dubbel wordt geteld, moet deze er weer één keer van worden afgetrokken. Via een Venndiagram zien we ook onmiddellijk dat $P(A\cup B)=1-P(A^c\cap B^c)$. De somregel voor 3 willekeurige gebeurtenissen A, B en C is

$$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$$

Voorbeeld 1.6 Wat is de kans op minstens één 6 als we 3 keer met een dobbelsteen werpen? Definieer A_1 als de gebeurtenis dat de 1e worp een 6 oplevert, en definieer A_2 en A_3 analoog. Volgens de somregel is de gevraagde kans

$$P(A_1\cup A_2\cup A_3)=\frac{1}{6}+\frac{1}{6}+\frac{1}{6}-\frac{1}{6^2}-\frac{1}{6^2}-\frac{1}{6^2}+\frac{1}{6^3}\approx 42\%$$

Vaak kunnen kansen op verschillende manieren worden berekend. In dit geval is een alternatieve berekeningswijze: $1-P(A_1^c\cap A_2^c\cap A_3^c)=1-\left(1-\frac{1}{6}\right)^3=1-\left(\frac{5}{6}\right)^3\approx 42\%$.

De uitgebreide somregel voor *n* willekeurige gebeurtenissen is

$$P(\textstyle\bigcup_{i=1}^n A_i)=\sum_{i=1}^n P(A_i)-\sum_{i<j}P(A_i\cap A_j)+\sum_{i<j<k}P(A_i\cap A_j\cap A_k)-\ldots+(-1)^{n+1}P(\textstyle\bigcap_{i=1}^n A_i)$$

Voor de laatste term geldt het plusteken voor oneven *n*, en het minteken voor even *n*.

Voorbeeld 1.7 We trekken 5 willekeurige kaarten uit een stok van 52 speelkaarten. Beschouw de gebeurtenis *B* dat alle 'kleuren' onder deze 5 kaarten voorkomen. We willen nu $P(B)$ bepalen. Definieer A_r als de gebeurtenis dat er *geen* ruiten voorkomt onder de 5 getrokken kaarten; definieer analoog A_h, A_k, A_s. Nu geldt $P(B)=1-P(B^c)=1-P(A_r\cup A_h\cup A_k\cup A_s)$. Zie verder Voorbeeld 1.8.

1.6 Combinatoriek (wiskundige rekenregels voor het tellen)

Er zijn $n! = 1 \times 2 \times ... \times n$ ($n \geq 0$, n geheel; $0! = 1$; spreek uit: '**n-faculteit**') manieren om n verschillende objecten te rangschikken. Elk van die manieren noemen we een **permutatie**. Indien niet alle objecten verschillend zijn, bijvoorbeeld, er zijn n_i identieke objecten van type i, met $\sum_{i=1}^{r} n_i = n$, dan is het aantal permutaties gelijk aan

$$\binom{n}{n_1\, n_2\, \ldots\, n_r} = \frac{n!}{n_1! \times n_2! \times ... \times n_r!}$$

De getallen $\binom{n}{n_1\, n_2\, \ldots\, n_r}$ noemen we ook **multinomiaal coëfficiënten**.

Een **variatie** van k objecten uit n verschillende objecten is een *groep* van k objecten uit deze n objecten in een bepaalde *volgorde*. Het aantal variaties van k objecten uit n verschillende objecten is

$$\frac{n!}{(n-k)!}$$

Een **combinatie** van *k* objecten uit *n* verschillende objecten is een *groep* van *k* objecten uit deze *n* objecten. Let wel: volgordeverwisselingen binnen zo'n groep leveren geen andere combinatie op. Het aantal combinaties van *k* objecten uit *n* verschillende objecten is

$$\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k! \times (n-k)!}$$

De getallen $\binom{n}{k}$ noemen we ook **binomiaal coëfficiënten**, overeenkomend met de multinomiaal coëfficiënten voor $r = 2$ (zie boven). De zogenaamde driehoek van Pascal (Figuur 1.2) geeft een rangschikking van de binomiaal coëfficiënten die allerlei bijzondere eigenschappen ervan laat zien, zoals $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$, $n > 1; 1 < k < n$. De top is $\binom{0}{0}$, en bijvoorbeeld de 3e rij $\binom{2}{0}\binom{2}{1}\binom{2}{2}$.

Figuur 1.2:
De driehoek van Pascal

```
                                1
                            1       1
                        1       2       1
                    1       3       3       1
                1       4       6       4       1
            1       5       10      10      5       1
        1       6       15      20      15      6       1
    1       7       21      35      35      21      7       1
..      ..      ..      ..      ..      ..      ..      ..      ..
```

Voorbeeld 1.8 Vervolg van Voorbeeld 1.7. Toepassing van de uitgebreide somregel geeft

$$P(B) = 1 - \left[4 \times \frac{\binom{39}{5}}{\binom{52}{5}} - 6 \times \frac{\binom{26}{5}}{\binom{52}{5}} + 4 \times \frac{\binom{13}{5}}{\binom{52}{5}} \right] \approx 0{,}264$$

Een alternatieve berekeningswijze geeft

$$P(B) = P(RRHKS) \times \frac{5!}{2!} \times 4 = \frac{13}{52} \times \frac{12}{51} \times \frac{13}{50} \times \frac{13}{49} \times \frac{13}{48} \times \frac{5!}{2!} \times 4 \approx 0{,}264$$

waarbij $P(RRHKS)$ de kans is dat we achtereenvolgens twee ruiten, een harten, een klaveren en een schoppen trekken. De factor 5!/2! vertegenwoordigt alle **permutaties** van *RRHKS*, terwijl de factor 4 ervoor zorgt dat ook alle mogelijkheden met 2 harten, 2 klaveren en 2 schoppen worden meegenomen.

Een **herhalingscombinatie** van k objecten uit n verschillende objecten is een *groep* van k objecten uit deze n objecten, waarbij elk object meer dan éénmaal mag voorkomen (trekken '**met teruglegging**'). Ook hier is alleen de samenstelling van de groep belangrijk, niet de volgorde. Het aantal herhalingscombinaties van k objecten uit n verschillende objecten is

$$\binom{n+k-1}{k}$$

Tot slot kunnen we ook nog k objecten uit n verschillende objecten nemen '**met teruglegging**' en waarbij de volgorde er wel toe doet. Het aantal mogelijke **variaties met herhaling** dat we zo kunnen krijgen is n^k.

Samengevat:

	Ordening is belangrijk	Ordening is onbelangrijk
Zonder teruglegging	$\frac{n!}{(n-k)!}$ (variaties)	$\binom{n}{k}$ (combinaties)
Met teruglegging	n^k (variaties met herhaling)	$\binom{n+k-1}{k}$ (herhalingscombinaties)

Voorbeeld 1.9 We trekken 5 kaarten uit een stok van 52 speelkaarten. De kans dat deze 5 kaarten precies 2 azen bevat, wordt gegeven door

$$\frac{\binom{4}{2}\times\binom{48}{3}}{\binom{52}{5}} = \frac{\frac{4!}{2!\times 2!}\times\frac{48!}{3!\times 45!}}{\frac{52!}{5!\times 57!}} = \frac{\frac{4\times 3}{2}\times\frac{48\times 47\times 46}{3\times 2}}{\frac{52\times 51\times 50\times 49\times 48}{5\times 4\times 3\times 2}} = \frac{47\times 46\times 5\times 4\times 3\times 2}{52\times 51\times 50\times 49} = 0{,}03993 \approx 4\%$$

2. Populatie; Steekproef; Stochastische variabele; Kansverdeling

2.1 Populatie en aselecte steekproef

Eén van de doelstellingen van de statistiek is het verkrijgen van informatie over de werkelijkheid via waarnemingen aan enkele elementen die informatie over die werkelijkheid bevatten. Kenmerkend, vervolgens, is dat we aan de hand van die waarnemingen beweringen willen doen over die werkelijkheid. Een eenvoudig voorbeeld is de productie van gloeilampen. In een fabriek worden volgens een bepaalde technologie op daartoe ontworpen machines gloeilampen geproduceerd; de machine wordt op de eerste dag van iedere maand opnieuw afgeregeld. Men wil weten wat de gemiddelde levensduur is van zo'n gloeilamp. Alle gloeilampen tezamen die de desbetreffende machine tussen gedurende een maand produceert, noemen we de te onderzoeken **populatie** (na een maand zou de gemiddelde levensduur door gebruik van andere grondstoffen en/of andere machine-instellingen kunnen veranderen). We zullen een deel van de geproduceerde gloeilampen moeten laten branden tot ze het begeven en per lamp registreren wat de levensduur is. Hoe preciezer we over de gemiddelde levensduur een uitspraak willen doen, des te groter zal het aantal te onderzoeken gloeilampen moeten zijn. De verzameling van te onderzoeken gloeilampen noemen we een **steekproef**. Om statistisch verantwoorde uitspraken te mogen doen, zal de steekproef **aselect** moeten zijn, d.w.z. alle in de desbetreffende maand geproduceerde gloeilampen moeten in principe een gelijke kans hebben om in die steekproef terecht te komen. Het is evident dat iedere nieuwe steekproef weer tot andere resultaten kan leiden (de levensduren van gloeilampen zijn verschillend). Daarom hebben we de **kansrekening** nodig om uitspraken te kunnen doen over het werkelijk gemiddelde op basis van één steekproef.

2.2 Stochastische variabele

Als gegeven is een kansruimte (Ω, P) dan wordt een **stochastische variabele**, X, gedefinieerd als een een functie op Ω die aan iedere uitkomst $\omega \in \Omega$ een reëel getal toevoegt. Een stochastische variabele is bedoeld om gebeurtenissen te beschrijven. In het voorbeeld van §2.1 zouden we dus de volgende stochastische variabele kunnen definiëren ($\Omega =$ {alle in 1 maand geproduceerde gloeilampen}; P is analoog aan die van de voorbeelden 1.1 en 1.2.

$X =$ "levensduur, afgerond op gehele uren, van een gloeilamp indien men deze continu laat branden bij kamertemperatuur"

Deze definitie maakt meteen duidelijk dat in dit verband ook andere stochastische variabelen gedefinieerd kunnen worden; bijvoorbeeld, hoewel minder relevant: de grootste diameter van het glas van de gloeilamp. Aan de steekproefelementen wordt dus een of andere (*nauwkeurig omschreven*) **eigenschap** gemeten die we uitdrukken in een getal. In dit geval zouden we geïnteresseerd kunnen zijn in de kans op de gebeurtenis $|X - 1000| < 50$, m.a.w. hoe groot is de kans dat de levensduur van een gloeilamp ligt tussen 950 en 1050 uur. Voor stochastische variabelen zullen in dit boek doorgaans hoofdletters als (bijvoorbeeld X, Y, Z) worden gebruikt (in sommige boeken ziet men ook een notatie met onderstreepte kleine letters zoals $\underline{x}, \underline{y}, \underline{z}$).

2.3 Kansverdeling

Bij de statistische analyse, vervolgens, veronderstelt men vaak dat de mogelijke waarden van zo'n stochastische variabele worden vastgelegd door een bepaalde **kansverdeling** die precies beschrijft met welke kans de stochastische variabele, X, waarden aanneemt in een willekeurige verzameling $A \subset \mathbf{R}$ (**R** is de verzameling van reële getallen); deze kans noteren we dan met $P_X(A)$, of $P(X \in A)$. Die kansverdeling moet men meestal opvatten als een **model** voor de werkelijke verdeling, die men immers niet kent. Altijd geldt uiteraard $0 \le P(X \in A) \le 1$. De bedoeling van de statistische analyse, vervolgens, is dan om uitspraken te doen over de **parameter(s)** die de theoretische kansverdeling bepalen. Veel gebruikte modellen zijn de **binomiaal verdeling** (Hoofdstuk 7) met parameters n en p, en de **normale verdeling** (Hoofdstuk 14) met parameters μ en σ^2. Uitspraken over een parameter van een bepaalde kansverdeling hebben de vorm van

- Het **schatten** van de waarde van de parameter
- Het **toetsen** van een hypothese over de parameter
- Het bepalen van een **betrouwbaarheidsinterval** voor de parameter

2.4 Discrete kansverdelingen

Een stochastische variabele, X, heet **discreet** als deze slechts een eindig aantal of aftelbaar veel waarden kan aannemen (bijvoorbeeld: een waarde uit $\{1,2,3,4,5,6,7,8,9,10\}$, uit $\{0,1,2,3,...,\infty\}$, of uit $\{1,1/2,1/3,1/4,...\}$ enz.). De kansverdeling van een discrete stochastische variabele die waarden kan aannemen uit de verzameling K, wordt volledig bepaald door de kansen $p(k) = P(X = k)$. De functie,

p, die aan iedere mogelijke waarde van X een getal uit het interval $[0;1]$ toevoegt, zodanig dat $\sum_{k\in K} p(k) = 1$, noemen we een **kansfunctie**. De kans dat X een waarde aanneemt uit de verzameling $A \subset K$ wordt gegeven door

$$P_X(A) = P(X \in A) = \sum_{k\in A} P(X = k)$$

In het bijzonder geldt

$$F(x) = P(X \le x) = \sum_{a\le x; a\in K} P(X = a) = 1 - P(X > x)$$

en, analoog,

$$P(X > x) = \sum_{a> x; a\in K} P(X = a) = 1 - F(x)$$

waarbij $F(x)$ de **kansverdeling** (meer volledig: de **cumulatieve (kans)verdeling(sfunctie)**) is van X. Uit de definitie van $F(x)$ volgt dat deze een niet-dalende functie is van x op het interval $[0;1]$. Voorbeelden van discrete kansverdelingen worden besproken in de Hoofdstukken 7 t/m 10. Overigens wordt in plaats van de notatie $P_X(A)$ en $F_X(x)$ vaak eenvoudigweg $P(A)$ en $F(x)$, respectievelijk, geschreven, mits dat niet tot verwarring leidt.

2.5 Continue kansverdelingen

De cumulatieve verdelingsfunctie van een **continue stochastische variabele**, X, wordt bepaald door de zogenaamde (**kans)dichtheid(sfunctie)**, $f_X(x)$, volgens

$$P(X \in A) = \int_{x\in A} f(x)dx, \quad (A \subset \mathbf{R})$$

In het bijzonder geldt natuurlijk

$$P(a \le X \le b) = \int_a^b f(x)dx$$

en, voor de cumulatieve verdelingsfunctie,

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Voor een continue stochastische variabele, X, geldt $P(X = x) = 0$ zodat $P(X \le x) = P(X < x)$. Nuttige rekenregels zijn nog

$$P(|X| \le x) = P(|X| < x) = F(x) - F(-x)$$

$$P(|X| \ge x) = P(|X| > x) = F(-x) + 1 - F(x)$$

Voorbeelden van continue kansverdelingen komen aan bod in Hoofdstukken 11 t/m 16.

2.6 Stochastische vectoren, simultane kansdichtheid en kansverdeling, onafhankelijke stochastische variabelen

In veel onderzoeken spelen meer dan één variabelen tegelijkertijd een rol. Op dezelfde kansruimte (Ω, P) worden dan verschillende stochastische variabelen, X_1 t/m X_k, gedefinieerd. De combinatie van deze stochastische variabelen is de **stochastische vector**

$$\mathbf{X} = (X_1, ..., X_k)$$

Als voorbeeld kan men denken aan $\Omega =$ "alle op 1-1-2006 geregistreerde Nederlanders" waarbij $X_1 =$ "gewicht in gram", $X_2 =$ "lengte in cm", etc. Bij statistische analyses is dan de **gezamenlijke (cumulatieve) kansverdeling**, $F_{X_1,...,X_k}(x_1,...,x_k) = P(X_1 \le x_1 \wedge ... \wedge X_k \le x_k)$, van **X** relevant (het symbool '$\wedge$' staat voor 'én'). Als $n = 2$ spreken we over een **bivariate cumulatieve kansverdeling**. De cumulatieve kansverdelingen $F_{X_i}(x_i)$ noemen we in dit verband de **marginale cumulatieve kansverdelingen**. De vector **X** is alleen een **k-dimensionale continue stochastische variabele** als een functie

$$f_{X_1,...,X_k}(x_1,...,x_k) \ge 0$$

de zogenaamde **gezamenlijke kansdichtheidsfunctie**, bestaat zodanig dat

$$F_{X_1,...,X_k}(x_1,...,x_k) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_k} f_{X_1,...,X_k}(u_1,...,u_k) du_1 ... du_k$$

Als **X** alleen **discrete** variabelen bevat dan noemen we

$$f_{X_1,...,X_k}(x_1,...,x_k) = P(X_1 = x_1 \wedge ... \wedge X_k = x_k)$$

de **gezamenlijke discrete kansdichtheidsfunctie** en een afzonderlijke kansfunctie $f_{X_i}(x_i)$ een **marginale kansfunctie** in dit verband.

Twee stochastische variabelen, *X* en *Y*, zijn **onafhankelijk** als de gebeurtenissen $X \in A$ en $Y \in B$ onafhankelijk zijn, d.w.z. als

$$P(X \in A \wedge Y \in B) = P(X \in A) \times P(Y \in B)$$

voor alle mogelijke verzamelingen $A \subset \mathbf{R}$ en $B \subset \mathbf{R}$. Dit begrip van **onafhankelijkheid** is op een voor de hand liggende wijze uit te breiden naar meer dan 2 stochastische variabelen.

3. Verwachtingswaarde en variantie

3.1 Verwachtingswaarde van een stochastische variabele

Een zeer belangrijk begrip in de statistiek is de **verwachtingswaarde** (ook wel: **verwachting**) van een stochastische variabele. Voor een *discrete* variabele, *X*, is de verwachtingswaarde, μ_X (of $E(X)$), gedefinieerd door

$$\mu_X = E(X) = \sum_{k \in K} k \times p_X(k)$$

als p_X de kansfunctie van *X* is, en *K* de verzameling van alle mogelijke waarden die *X* kan aannemen. Voor een *continue* variabele, *X*, is de verwachtingswaarde, μ_X, gedefinieerd door

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \times f_X(x) dx$$

als f_X de kansdichtheidsfunctie is van *X*. Beide notaties, $E(X)$ en μ_X, worden door elkaar gebruikt ('*E*' is een afkorting van 'Expected value'). De verwachtingswaarde is op te vatten als een **gewogen gemiddelde** van de mogelijke waarden van *X* waarbij de gewichten bepaald worden door de kans(dichtheids)functie: waarden die met een grotere kans voorkomen krijgen ook een groter gewicht in de verwachtingswaarde. In plaats van "verwachtingswaarde van *X*" wordt ook wel gesproken over het "**gemiddelde van** *X*" of zelfs het "gemiddelde van de kansverdeling van *X*". De verwachtingswaarde van *X* geeft aan waar het *centrum van de kansverdeling* is gelocaliseerd.

Voorbeeld 3.1 Een *zuivere* dobbelsteen wordt gekenmerkt door gelijke kansen op elk van de resultaten 1,2,3,4,5, of 6 van een worp. De verwachtingswaarde van $X =$ "het aantal ogen na een worp" is derhalve $\mu_X = \sum_{k=1}^{6} k/6 = 3{,}5$.

Voorbeeld 3.2 Stel dat de kansdichtheid van een continue stochastische variabele *X* is gedefinieerd door $f_X(x) = 2$ voor $0 \le x \le 0{,}5$, en $f_X(x) = 0$ elders. Dan geldt $\mu_X = \int_0^{0,5} 2x dx = x^2 \Big|_0^{0,5} = 0{,}25$.

3.2 Variantie en standaardafwijking van een stochastische variabele

De **variantie**, σ_X^2 (of $Var(X)$), van een stochastische variabele, *X*, is een maat voor de *spreiding*. In woorden is de variantie de verwachtingswaarde van het gekwadrateerde verschil tussen *X* en μ_X. De definitie is

$$\sigma_X^2 = Var(X) = E\left[(X - \mu_X)^2\right] = E\left[X^2 - 2\mu_X X + \mu_X^2\right] = E(X^2) - 2\mu_X^2 + \mu_X^2 = E(X^2) - \mu_X^2$$

(Merk op dat de laatste uitwerking de belangrijke eigenschap $E(X^2) = \mu_X^2 + \sigma_X^2$ laat zien.) Voor een *discrete* variabele *X* geldt (p_X en *K* als in §3.1)

$$Var(X) = \sum_{k \in K} (k - \mu_x)^2 \times p_X(k)$$

Voor een *continue* variabele X geldt (f_X als in §3.1)

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 \times f_X(x) dx$$

Dat de variantie een **spreidingsmaat** is, blijkt uit de weging van de gekwadrateerde verschillen (van X met zijn verwachtingswaarde) met de corresponderende kans(dichtheid).

Een handiger spreidingsmaat is de standaardafwijking omdat deze is uitgedrukt in dezelfde eenheid als die van de variabele. De **standaardafwijking**, σ_X, van een stochastische variabele X is gedefinieerd door

$$\sigma_X = +\sqrt{Var(X)} = +\sqrt{\sigma_X^2}$$

dus als de positieve wortel van de variantie.

Voorbeeld 3.3 Vervolg van voorbeeld 3.1. De variantie van X is $\sigma_X^2 = \sum_{k=1}^{6}(k-3{,}5)^2/6 \approx 2{,}917$; en $\sigma_X \approx 1{,}708$.

Voorbeeld 3.4 Vervolg van voorbeeld 3.2. De variantie van X is $\sigma_X^2 = \int_0^2 (x-0{,}25)^2 \times 0{,}5\,dx = (x-0{,}25)^3/3\Big|_0^2 = \left[1{,}75^3 - (-0{,}25)^3\right]/3 \approx 1{,}792$; en $\sigma_X \approx 1{,}339$.

3.3 Rekenregels voor verwachtingswaarde en variantie

Voor de som van een aantal stochastische variabelen X_1 t/m X_n geldt algemeen

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

m.a.w., de verwachtingswaarde van de som is de som van de verwachtingswaarden. De variantie van de som van een aantal stochastische variabelen is echter *alleen* gelijk aan de som van de afzonderlijke varianties als X_1 t/m X_n *onafhankelijk* zijn, dus

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) \text{ mits } E(X_1 \times \ldots \times X_n) = E(X_1) \times \ldots \times E(X_n)$$

(zie §2.6 voor onafhankelijkheid van stochastische variabelen).

Verwachting en variantie van een **lineaire transformatie** $aX + b$ is gegeven door

$$E(aX+b) = aE(X) + b$$
$$Var(aX+b) = a^2 Var(X) \text{, en dus } \sigma_{aX+b} = a\sigma_X$$

De tweede vergelijking maakt duidelijk dat *optelling met een constante* voor de variantie van een stochastische variabele geen verschil maakt.

3.4 Covariantie en correlatiecoëfficiënt

De (populatie)**covariantie**, $Cov(X,Y)$, en de (populatie)**correlatiecoëfficiënt**, $\rho_{X,Y}$, zijn twee belangrijke grootheden voor de mate waarin twee stochastische variabelen, X en Y, een **lineaire relatie** met elkaar hebben (een *speciale vorm* van afhankelijkheid). De definities zijn

$$Cov(X,Y) = \sigma_{XY} = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E(X\times Y) - \mu_X \times \mu_Y$$

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

Een bijzonder geval is $Cov(X,X) = \sigma_{XX} = \sigma_X^2 = Var(X)$. De covariantie is 0 als X en Y onafhankelijk van elkaar zijn (immers dan is $E(X\times Y) = \mu_X \times \mu_Y$). De correlatiecoëfficiënt heeft altijd een waarde tussen –1 en 1. Als $0 < \rho_{X,Y} < 1$ dan hebben X en Y een **positieve lineaire relatie**. Als $-1 < \rho_{X,Y} < 0$ dan hebben X en Y een **negatieve lineaire relatie**. De lineaire relatie is *perfect* als $\left|\rho_{X,Y}\right| = 1$.

De variantie van de som van twee stochastische variabelen, X en Y, is

$$Var(X+Y) = Var(X) + Var(Y) + 2\times Cov(X,Y)$$

Bij onafhankelijkheid van X en Y valt de laatste term natuurlijk weg.

Voorbeeld 3.5 We gooien een rode en een zwarte dobbelsteen. X is het aantal ogen van de rode, en Y van de zwarte dobbelsteen. We bepalen de covariantie en correlatiecoëfficiënt tussen X en $Z = X + Y$. Daarvoor hebben we nodig

$$E(X) = 3{,}5$$
$$E(Z) = \sum_{i=1}^{6}\sum_{k=1}^{6}(i+k)/36 = 7$$
$$Var(X) = \sum_{i=1}^{6} i^2 - 3{,}5^2 \approx 2{,}917$$
$$Var(Z) = \sum_{i=1}^{6}\sum_{k=1}^{6}(i+k)^2/36 - 7^2 \approx 5{,}833$$
$$E(X\times Z) = \sum_{i=1}^{6}\left(i\times\sum_{k=1}^{6}(i+k)\right)/36 \approx 27{,}417$$

zodat

$$Cov(X,Z) = E(X\times Z) - \mu_X \times \mu_Z \approx 27{,}417 - 3{,}5\times 7 \approx 2{,}917$$

en

$$\rho_{X,Z} \approx 2{,}917/\sqrt{2{,}917\times 5{,}833} \approx 0{,}708$$

4. De wet van de grote aantallen

4.1 De ongelijkheid van Chebyshev

Voor een stochastische variabele met verwachtingswaarde μ_X en variantie σ_X^2 geldt

$$P\left(|X-\mu_X| \geq k \times \sigma_X\right) \leq \frac{1}{k^2},\ (k>0)$$

of, equivalent,

$$P\left(|X-\mu_X| < k \times \sigma_X\right) \geq 1-\frac{1}{k^2}$$

Dus, bijvoorbeeld, de kans dat X een waarde aanneemt die minder dan 2 keer zijn standaardafwijking verwijderd is van zijn verwachtingswaarde, is groter dan 0,75. Dit is de **ongelijkheid van Chebyshev**. Het bijzondere van deze ongelijkheid is dat informatie over de vorm van de kansverdeling van X niet nodig is; het geldt dus voor *iedere mogelijke* kansverdeling.

4.2 De zwakke wet van de grote aantallen

Met behulp van (een meer algemene vorm van) de ongelijkheid van Chebyshev kan bewezen worden dat het gemiddelde $\overline{X}_n$ (het **steekproefgemiddelde**, zie §5.2) van een aselecte steekproef $X_1,\ldots,X_n$ met de onbekende verwachtingswaarde, μ_X, willekeurig dicht benaderd kan worden door *n* (dus de steekproefomvang) groot genoeg te maken. Formeler: voor willekeurige waarden van $\varepsilon > 0$ en $0 < \delta < 1$ bestaat er een geheel getal *m* zodanig dat voor alle gehele getallen $n \geq m$ geldt

$$P\left(\left|\overline{X}_n - \mu_X\right| < \varepsilon\right) \geq 1 - \delta$$

Dit heet de **zwakke wet van de grote aantallen**. In het bewijs hiervan blijkt dat $n > \frac{\sigma_X^2}{\varepsilon^2 \delta}$.

Voorbeeld 4.1 Stel dat van de *onbekende* kansverdeling van *X* het gemiddelde μ_X onbekend is, maar de standaardafwijking bekend, $\sigma_X = 2$. Als we nu met een aselecte steekproef μ_X zo nauwkeurig willen schatten dat de kans op een absoluut verschil van minder dan $\varepsilon = 0{,}5$ minimaal 90% is (dus $\delta = 0{,}1$) dan dient de steekproefomvang minimaal $n = 4/(0{,}5^2 \times 0{,}1) = 160$ te zijn. Merk op dat als de kansverdeling *bekend* is met een kleinere steekproefomvang kan worden volstaan.

4.3 De Centrale Limietstelling

De **centrale limietstelling** (**CLS**) is één van de meest belangrijke stellingen uit de kansrekening. Deze stelling biedt ons een *benadering* van de kansverdeling van een steekproefgemiddelde, en luidt:

Als $X_1,\ldots,X_n$ een rij van elkaar onafhankelijke en gelijkverdeelde stochastische variabelen zijn met verwachtingswaarde μ_X en variantie σ_X^2, dan convergeert de kansverdeling van

$$Z_n = \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} = \frac{\overline{X}_n - \mu_X}{\sigma_X / \sqrt{n}}$$

naar de kansverdeling, Φ, van een standaardnormaal verdeelde variabele (zie §14.2).

Merk op dat Z_n verkregen wordt door het gemiddelde $\overline{X}_n$ van de stochasten X_1 t/m X_n te "**standaardiseren**", d.w.z. we verminderen $\overline{X}_n$ met zijn verwachtingswaarde, en delen het resultaat door zijn standaardafwijking. We kunnen van dit resultaat op diverse manieren gebruik maken om bepaalde kansen te benaderen m.b.v. de limietverdeling Φ.

Kansen m.b.t. het gestandaardiseerde gemiddelde:

$$P\left(a < Z_n < b\right) \approx \Phi(b) - \Phi(a)$$

Kansen m.b.t. het gemiddelde:

$$P(a < \overline{X}_n < b) \approx \Phi\left(\frac{b-\mu_X}{\sigma_X/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu_X}{\sigma_X/\sqrt{n}}\right)$$

Kansen m.b.t. de som:

$$P\left(a < \sum_{i=1}^{n} X_i < b\right) \approx \Phi\left(\frac{b-n\mu_X}{\sigma_X \times \sqrt{n}}\right) - \Phi\left(\frac{a-n\mu_X}{\sigma_X \times \sqrt{n}}\right)$$

Het opmerkelijke van de CLS is dat geen enkele voorwaarde wordt gesteld met betrekking tot de vorm van de kansverdeling van X_i ; dus ook voor *extreme* verdelingen geldt de stelling. Het volgende voorbeeld illustreert dit.

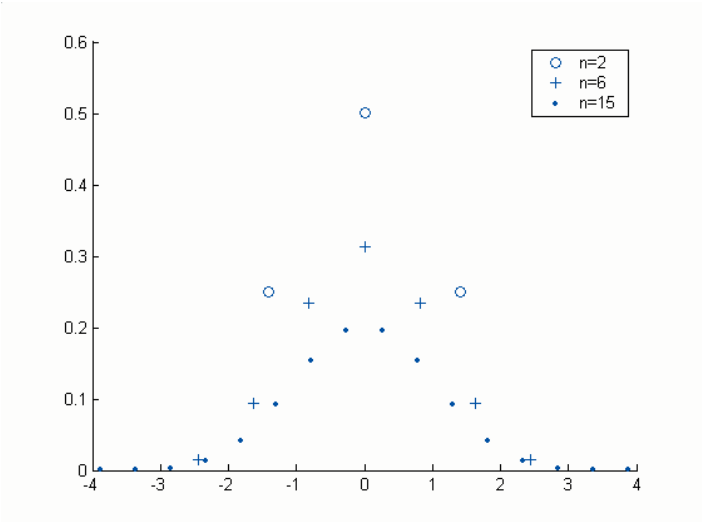
Voorbeeld 4.2 Zij X een stochast die gegeven is door $P(X=-1) = P(X=1) = 0{,}5$. Denk bijvoorbeeld aan de worp met een munt waarbij aan de uitkomsten 'kop' en 'munt' respectievelijk de getallen -1 en 1 worden gekoppeld. Uiteraard geldt $\mu_X = -1 \times 0{,}5 + 1 \times 0{,}5 = 0$ en $\sigma_X^2 = E(X^2) - \mu_X^2 = 1$. Stel we gooien (op onafhankelijke wijze) *n* keer met die munt en registreren met X_1 t/m X_n de bijbehorende resultaten. Beschouw de (steekproef)verdeling van $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Het blijkt dat, bijvoorbeeld,

$$P(\overline{X}_2 = 1) = P(\overline{X}_2 = -1) = 1/4\,;\ P(\overline{X}_2 = 0) = 2/4$$
$$P(\overline{X}_3 = 1) = P(\overline{X}_3 = -1) = 1/8\,;\ P(\overline{X}_3 = 1/3) = P(\overline{X}_3 = -1/3) = 3/8$$
$$P(\overline{X}_4 = 1) = P(\overline{X}_4 = -1) = 1/16\,;\ P(\overline{X}_4 = 2/4) = P(\overline{X}_4 = -2/4) = 4/16\,;\ P(\overline{X}_4 = 0) = 6/16$$

Figuur 4.1 laat ter illustratie voor de waarden $n = 2,\ 6,\ \text{en}\ 15$ de kansverdeling zien van

$$Z_n = \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} = \frac{\overline{X}_n}{1/\sqrt{n}} = \overline{X}_n\sqrt{n}$$

Figuur 4.1:
Illustratie van de Centrale Limietstelling (Voorbeeld 4.2)



5. Beschrijvende statistiek

5.1 Klassificatie van variabelen

Wanneer men stochastische variabelen of, analoog, steekproefgegevens gaat *beschrijven*, dient men rekening te houden met het karakter ervan. Men onderscheidt achtereenvolgens **nominale**, **ordinale**, **interval** en **ratio** variabelen. Bij *nominale* variabelen onderscheiden de mogelijke waarden zich alleen door de *naam*; voorbeelden zijn '*politieke voorkeur*', '*religie*', '*soort auto*' e.d. Het kenmerkende van deze variabelen is dat de mogelijke waarden ervan niet eenduidig zijn te ordenen. Een stapje hoger in de hiërarchie zijn de *ordinale* variabelen; deze onderscheiden zich van de nominale variabelen doordat de afzonderlijke waarden wel eenduidig zijn te *ordenen*; voorbeelden zijn '*mate van tevredenheid*' of '*product kwaliteit*' (met *niveaus*: uitstekend, goed, matig, slecht) e.d. *Interval* variabelen kenmerken zich door de eigenschap dat het *verschil* tussen twee mogelijke waarden van zo'n variabele op de getallenrechte ook echt betekenis heeft; hercodering van een intervalvariabele door een lineaire transformatie verandert de betekenis van die variabele niet wezenlijk. Een klassiek voorbeeld is '*temperatuur*'. Deze wordt gemeten in graden Fahrenheit (*F*), Celsius (*C*) of Kelvin (*K*); hoewel deze temperatuurschalen een verschillend nulpunt hebben en de één een lineaire transformatie is van de ander ($F = 32 + 1{,}8 \times C$, $K = 273 + C$) beschrijven ze dezelfde informatie over de temperatuur. De *ratio* variabele is de hoogste in hiërarchie: bij zo'n variabele biedt ook de afstand tot 0 zinvolle informatie zodat ook de *verhouding* van twee waarden betekenis heeft; voorbeelden zijn '*gewicht*', '*leeftijd*', e.d.

5.2 Locatiematen: modus, (steekproef)gemiddelde, mediaan

Beschrijvende statistiek houdt zich bezig met het samenvatten van de eigenschappen van een kansverdeling of (meestal) van steekproefgegevens in één of enkele getallen (maten), tabellen of grafieken. Bij samenvatting in getallen onderscheid men grofweg maten voor *locatie* en *spreiding* (§5.3). **Locatiematen** zeggen wat over waar het *centrum van de verdeling* zich bevindt (gelocaliseerd is). De enig zinvolle locatiemaat voor *nominale* variabelen is de **modus**: de waarde (of de waarden) die het meest frequent voorkomen. Bij *ordinale* variabelen kan daarnaast ook de **mediaan** gebruikt worden als locatiemaat; deze legt vast voor welke waarde geldt dat 50% van de verdeling (of van de steekproefgegevens) kleiner is (en dus ook 50% groter). Volgens een gebruikelijke definitie is de **mediaan** van een rij *geordende* waarden $x_1,...,x_n$ gegeven door $x_{(n+1)/2}$ als *n oneven* is, en $(x_{n/2} + x_{n/2+1})/2$ voor *even n*. Het **gemiddelde** van een rij waarden $x_1,...,x_n$ is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Als $x_1,...,x_n$ de waarden zijn uit een steekproef dan noemen we $\bar{x}$ het **steekproefgemiddelde**. Modus, mediaan en gemiddelde zijn zinvolle maten voor *interval* en *ratio* variabelen.

Voorbeeld 5.1 Beschouw de volgende steekproef uit de kansverdeling van een intervalvariabele: 12, 18, 17, 18, 10, 8. De modus gelijk is aan 18; de steekproefomvang is $n = 6$, zodat de mediaan gelijk is aan het gemiddelde van de 3e en 4e waarneming indien geordend naar grootte, dus $(12 + 17)/2 = 14{,}5$. Het steekproefgemiddelde is natuurlijk $83/6$.

Merk op dat locatie- en spreidingsmaten van stochastische variabelen vaak kunnen worden uitgedrukt in de parameters van de bijbehorende kansverdeling (zie Hoofdstukken 7 t/m 15); anders dienen technieken zoals in §3.1 en §3.2 gebruikt te worden.

5.3 Spreidingsmaten: bereik, (steekproef)variantie en standaard-afwijking, variatiecoëfficiënt

Spreidingsmaten zijn alleen zinvol voor de *kwantitatieve* interval en ratio variabelen. Zij hebben als doel om aan te geven in welke mate de waarden van een verdeling of steekproef uiteenlopen. Daartoe worden verscheidene maten gebruikt. De eerste, voor de hand liggende, maat is het **bereik** gedefinieerd als het verschil tussen de grootste en de kleinste waarde (waarneming). Bij veel verdelingen is het bereik oneindig en is dus onbruikbaar; bij steekproeven is deze maat erg gevoelig voor **uitschieters** (extreme waarnemingen; **uitbijters**), hetgeen een onwenselijke eigenschap is. Geschikter zijn maten die gebaseerd worden op kwartielen, of op verschillen met het gemiddelde.

Men onderscheid 3 **kwartielen**: het 1e, 2e en 3e kwartiel, vaak genoteerd als respectievelijk $Q_1, Q_2,$ en Q_3. Het 2e kwartiel is hetzelfde als de in §5.2 gedefinieerde mediaan; het 1e (3e) kwartiel is de mediaan van alle waarden of waarnemingen kleiner (groter) dan het 2e kwartiel. De 3 *kwartielen* splitsen dus de steekproefgegevens of de verdeling in 4 *kwarten* van elk 25%. De zogenaamde **interkwartielafstand** is het verschil tussen het 3e en het 1e kwartiel: $Q_3 - Q_1$.

De meest gebruikte spreidingsmaat is de **standaardafwijking** (§3.2). In de ongelijkheid van Chebyshev (§4.1) speelt deze een prominente rol: bijvoorbeeld de kans dat een waarneming verder dan 3σ verwijderd is van de verwachtingswaarde, μ, is kleiner dan 1/9. Bij een aselecte steekproef $x_1,...,x_n$ rekent men naast het steekproefgemiddelde meestal ook de **steekproefvariantie**, s^2, uit. De formule daarvoor is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n \times \bar{x}^2\right)$$

De (positieve) wortel hieruit heet de **steekproefstandaardafwijking**. Merk op dat de laatste gelijkheid de zogenaamde **rekenformule** voor de steekproefvariantie is; soms is deze handiger dan de definitieformule. Als uitschieters in een steekproef gedefinieerd worden als waarnemingen die verder verwijderd zijn van het steekproefgemiddelde dan 3 keer de steekproefstandaardafwijking, dan geldt ook dat we nooit meer dan *n*/9 uitschieters in een steekproef zullen aantreffen.

Als men, tot slot, de spreiding wil vergelijken tussen verschillende verdelingen, of steekproeven, dan gebruikt men vaak de zogenaamde **variatiecoëfficiënt** (met Griekse letter ν ('nu') als symbool). Dit is een dimensieloze grootheid die ontstaat door de standaardafwijking te delen door het corresponderende gemiddelde; dus $\nu = \sigma / \mu$.

5.4 Maten voor lineaire samenhang: steekproefcovariantie, en steekproefcorrelatiecoëfficiënt

De **covariantie**, σ_{XY}, en de **correlatiecoëfficiënt**, $\rho_{X,Y}$, tussen twee stochastische variabelen, X en Y, (zie §3.4) kunnen geschat worden op basis van een gepaarde steekproef $(x_1, y_1),...,(x_n, y_n)$. De volgende formules geven respectievelijk de **steekproefcovariantie**, s_{xy}, en de **steekproefcorrelatiecoëfficiënt**, r_{xy}:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right)$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

De laatste gelijkheid in de formule voor s_{xy} bevat weer de **rekenformule**.

6. Het toetsen van een hypothese

6.1 Nulhypothese en alternatieve hypothese; toetsingsgrootheid

In de komende hoofdstukken zullen verscheidene toetsprocedures de revue passeren. Hier bespreken we het algemene *raamwerk* voor het toetsen van een hypothese. Een toetsprocedure stelt ons in staat om (onder zeker omstandigheden) na te gaan of een bepaalde bewering of een bepaald vermoeden (*waarschijnlijk*) juist is op basis van steekproefgegevens. Een volledige toetsprocedure omvat de volgende stappen en redeneringen:

Een onderzoeker formuleert een *bewering* met betrekking tot een onbekende parameter, zeg θ, van een kansverdeling, bijvoorbeeld de proportie, p, van de elementen van een populatie die voldoen aan een bepaalde eigenschap, of het gemiddelde, μ, van een populatie, enzovoorts. Deze bewering noemt men de **onderzoekshypothese**, en vindt men in de toetsprocedure terug als de zogenaamde **alternatieve hypothese**, aangeduid met H_1 (we zien ook wel: H_a). We onderscheiden meestal 3 mogelijke alternatieve hypothesen: $H_1 : \theta \neq \theta_0$, $H_1 : \theta < \theta_0$, en $H_1 : \theta > \theta_0$; deze hypothesen noemen we respectievelijk '**tweezijdig**', '**(links)eenzijdig**' en '**(rechts)eenzijdig**'. We zijn slechts echter bereid ervan uit te gaan dat de alternatieve hypothese juist is als de steekproefgegevens 'sterke' aanwijzingen bevatten in die richting. We gebruiken daarvoor een zogenaamde **toetsingsgrootheid** die een functie is van de steekproefgegevens uit een **aselecte steekproef** en waarvan we de (eventueel benaderende) kansverdeling kennen onder de bijzondere *veronderstelling* dat $\theta = \theta_0$. We noemen $H_0 : \theta = \theta_0$ de **nulhypothese**. Dus kortom: als de **waarde van de toetsingsgrootheid** in een concreet steekproefonderzoek een *extreme* waarde aanneemt in de kansverdeling van de toetsingsgrootheid 'onder H_0' dan **verwerpen** we de nulhypothese ten gunste van de alternatieve (onderzoeks)hypothese.

6.2 Onbetrouwbaarheidsdrempel, fout van de eerste soort; kritieke waarde

Afhankelijk van het type alternatieve hypothese moet men i.h.a. voor 'extreem' lezen respectievelijk: 'extreem klein of extreem groot' (bij $H_1 : \theta \neq \theta_0$), 'extreem klein' (bij $H_1 : \theta < \theta_0$), of 'extreem groot' (bij $H_1 : \theta > \theta_0$). Vanaf welke waarde een toetsingsgrootheid 'extreem' genoemd mag worden, hangt af van een afspraak die we vòòr de uitvoering van de toets willen maken ten aanzien van de *kans* dat we, ondanks dat H_0 waar is, toch de *onjuiste beslissing* nemen om H_0 te verwerpen. Deze kans, de zogenaamde **onbetrouwbaarheidsdrempel** of '**fout van eerste soort**', is typisch 'klein', en wordt aangeduid met de Griekse letter α . Een zeer gebruikelijke waarde voor α is 0,05. Op basis van deze α , het type alternatieve hypothese, en de kansverdeling van de toetsingsgrootheid onder H_0 kunnen we de zogenaamde **kritieke waarde(n)** bepalen. Als de waarde van de toetsingsgrootheid extremer is dan deze kritieke waarde(n) dan is de eindconclusie van het onderzoek: 'verwerp H_0', en anders: 'verwerp H_0 niet'.

De naamgeving 'fout van eerste soort' suggereert dat er ook een '**fout van de tweede soort**' is; dat is de kans dat we de nulhypothese *ten onrechte niet* verwerpen. Deze kans kan in een concreet geval

alleen worden bepaald indien de alternatieve hypothese verder gespecificeerd wordt. We laten de details hier achterwege en verwijzen naar de diverse tekstboeken.

6.3 Overschrijdingskans

Zeer gebruikelijk (met name door statistische softwarepakketten) is om het resultaat van een toetsprocedure te presenteren als een zogenaamde **overschrijdingskans**. Dit is de kans om, gegeven dat H_0 waar is, op basis van een aselecte steekproef een waarde van de toetsingsgrootheid te vinden die gelijk is aan of nòg extremer is dan de reeds gevonden waarde in het onderzoek. Deze overschrijdingskans noemen we ook wel de **p-waarde**. De algemene regel aangaande het doen van een uitspraak over H_0 is dan: 'Verwerp H_0 als de p-waarde kleiner is dan α ' (en anders niet).

Softwarepakketten presenteren doorgaans de overschrijdingskans die hoort bij een *tweezijdige* alternatieve hypothese. Het is dan even opletten hoe deze informatie gebruikt dient te worden om een uitspraak te doen in geval van een éénzijdige alternatieve hypothese.

6.4 Een eenvoudig voorbeeld

Men vertrouwt de 'zuiverheid' van een bepaalde dobbelsteen niet, met name vermoedt men dat de relatieve frequentie, *p*, van het aantal 'zessen' hoger ligt dan de gebruikelijke 1/6. Men wil een onderzoek daartoe met een onbetrouwbaarheiddrempel van 1%. Voor het onderzoek formuleert men de **hypothesen**:

$$H_0 : p = 1/6 \text{ vs. } H_1 : p > 1/6;\ \alpha = 0{,}01$$

Het onderzoek zal bestaan uit $n = 20$ worpen (men schat dat deze **steekproefomvang** genoeg is) met de desbetreffende dobbelsteen. We registreren dan het aantal zessen in de 20 worpen, en duiden het resultaat aan met de stochastische variabele X . Als toetsingsgrootheid kunnen we X nemen of ook X / n ; we nemen echter als **toetsingsgrootheid**:

$$X$$

Om de toets te kunnen uitvoeren dienen we de kansverdeling van X te kennen onder H_0 , d.w.z. als $p = 1/6$. X heeft dan een zogenaamde binomiaal verdeling met parameters $n = 20$ en $p = 1/6$ (zie Hoofdstuk 7). Nu weten we genoeg om de toets uit te voeren en voeren zorgvuldig de 20 worpen uit. Het resultaat is (bijvoorbeeld):

1, 6, 5, 4, 6, 6, 2, 6, 2, 4, 4, 5, 2, 1, 5, 4, 1, 6, 4, 6

De **waarde van de toetsingsgrootheid** is dus $x = 6$.

De **overschrijdingskans** is

$$P(X \geq 6 \mid p = 1/6) = \sum_{i=6}^{20} \binom{20}{i} \left(\frac{1}{6}\right)^{i} \left(\frac{1}{6}\right)^{20-i} \approx 0{,}1$$

De **kritieke waarde** ligt bij $X = 9$ (aangezien $P(X \geq 8 \mid p = 1/6) \approx 0{,}11$, en $P(X \geq 9 \mid p = 1/6) \approx 0{,}003$). Het resultaat van de toets is : 'verwerp H_0 *niet*' omdat de waarde van de toetsingsgrootheid niet in het **kritieke gebied** ($X \geq 9$) ligt, of, alternatief, omdat de overschrijdingskans groter is dan 0,05; op basis van dit onderzoek is er onvoldoende bewijs dat de desbetreffende dobbelsteen niet zuiver is.

7. Binomiaal verdelingen Bin(*n*,*p*)

7.1 Kansverdeling, parameters, verwachtingswaarde en variantie

De uitkomsten van een kansexperiment met slechts 2 mogelijke uitkomsten worden vaak genoteerd als respectievelijk *S* (**succes**) en *F* (**failure**) met $P(\{S\}) = p$ en $P(\{F\}) = q = 1 - p$; een voorbeeld is het opgooien van een munt met mogelijke uitkomsten $S =$ 'kop' en $F =$ 'munt', en $p = 0.5$ (mits de munt 'zuiver' is, en de worp voldoende 'wild'). We noemen zo'n experiment meestal een **binomiaal** (of: **Bernoulli**) **experiment**. Het *n* keer op onafhankelijke wijze herhalen van zo'n experiment geeft mogelijke uitkomsten $\omega = (X_1,...,X_n)$ waarbij $X_i \in \{S,F\}$, *i*=1,…,*n*. Het aantal uitkomsten ω waarbij de *n* individuele binomiaal experimenten in totaal *k* successen opleveren, is de **binomiaalcoëfficiënt** $\binom{n}{k}$. De kans dat het aantal keren succes, *Y*, in ω gelijk is aan *k*, $P(Y = k)$, is nu gelijk aan

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0,1,...,n\}; \quad 0 < p < 1$$

Een **binomiaal verdeling** wordt dus gekenmerkt door twee **parameters**, *n* en *p*. Als de kansverdeling van een stochastische variabele *Y* een binomiaal verdeling is dan schrijven we

$$Y \sim \text{Bin}(n, p)$$

Voorbeeld 7.1 Een student maakt zonder voorbereiding een multiple choice toets bestaande uit 20 vragen ($n = 20$) van elk 4 alternatieven waarvan er slechts 1 juist is ($p = 1/4$). Hij gokt alle antwoorden. Een voldoende vereist 11 goede antwoorden. Als *Y* het aantal correcte antwoorden is, is de kans dat hij een voldoende scoort gelijk aan

$$\sum_{k=11}^{20} \binom{20}{k} \times 0{,}25^k \times 0{,}75^{20-k} = 0{,}003942$$

Voorbeeld 7.2 We gooien 5 dobbelstenen. De kans op tenminste 3 gelijke ogen vinden we bijvoorbeeld door de kansen op precies 3, 4 en 5 énen op te tellen en met 6 te vermenigvuldigen.

$$6 \times \left[\binom{5}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 + \binom{5}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^1 + \binom{5}{5} \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^0 \right] \approx 0{,}213$$

Voorbeeld 7.3 We gooien 4 dobbelstenen. De kans op tenminste 2 gelijke ogen is te vinden volgens hetzelfde principe als in Voorbeeld 7.2, hoewel enigszins aangepast om dubbeltellingen te voorkomen:

$$\frac{6 \times \left[\binom{4}{2} \times 5 \times (4 + 0{,}5) + \binom{4}{3} \times 5 + \binom{4}{4} \right]}{6^4} = \frac{13}{18} \approx 0{,}72$$

In plaats van 5 schrijven we 4+0,5 omdat te voorkomen dat bijvoorbeeld de combinatie 1122 dubbel wordt geteld. Een veel eenvoudigere manier (die in Voorbeeld 7.2 *niet* werkt) is (zie ook §1.6):

$$1-\frac{6}{6}\times\frac{5}{6}\times\frac{4}{6}\times\frac{3}{6}=1-\frac{6!/2!}{6^4}=\frac{13}{18}\approx 0{,}72$$

Belangrijke karakteristieken van een binomiale variabele $Y \sim \text{Bin}(n, p)$ zijn

verwachtingswaarde	$\mu = n \times p$
variantie	$\sigma^2 = n \times p \times (1-p) = n \times p \times q$
standaardafwijking	$\sigma = \sqrt{n \times p \times q}$

7.2 Overschrijdingskansen

Als we beschikken over een aselecte steekproef $X_1,...,X_n$ uit een populatie waarin een fractie p van de elementen een bepaalde eigenschap heeft dan heeft het aantal elementen, Y, in die steekproef met de desbetreffende eigenschap een binomiaal verdeling met parameters n en p. Stel dat in een concreet experiment $Y = k$. Als p onbekend is dan is (afhankelijk van de situatie) het toetsen van een drietal nul- en alternatieve hypothesen interessant. Onderstaande tabel vermeldt deze met de bijbehorende **overschrijdingskansen**; hierbij is $Y_0 \sim \text{Bin}(n, p_0)$.

H_0	H_1	Overschrijdingskans
$p \le p_0$, of: $p = p_0$	$p > p_0$	$P(Y_0 \ge k) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$
$p \ge p_0$, of: $p = p_0$	$p < p_0$	$P(Y_0 \le k) = \sum_{i=0}^{k} \binom{n}{i} p_0^i (1-p_0)^{n-i}$
$p = p_0$	$p \ne p_0$	$\sum_{i \in A} \binom{n}{i} p_0^i (1-p_0)^{n-i}$, $A = \{j \mid P(Y_0 = j) \le P(Y_0 = k)\}$

Een nulhypothese wordt verworpen met **onbetrouwbaarheidsdrempel** α, indien de overschrijdingskans kleiner is dan α.

Voorbeeld 7.4 De producent van een machine die microchips produceert, beweert dat ten hoogste 15% van de geproduceerde chips onbruikbaar zijn. Stel dat er van de eerste 20 geproduceerde chips 6 defect zijn. De vraag of er op grond hiervan reden is om aan te nemen dat de producent geen gelijk heeft, kunnen we beantwoorden via het toetsen van $H_0 : p = 0.15$ vs. $H_1 : p > 0.15$. De bijbehorende overschrijdingskans is $\sum_{i=6}^{20} \binom{20}{i} \times 0{,}15^i \times 0{,}85^{20-i} = 1 - \sum_{i=0}^{5} \binom{20}{i} \times 0{,}15^i \times 0{,}85^{20-i} \approx 0{,}067$. Bij een onbetrouwbaarheidsdrempel $\alpha = 0.05$ kunnen we de nulhypothese dus *niet* verwerpen (bij $\alpha = 0.1$ natuurlijk *wel*).

7.3 Het benaderen van een binomiaal verdeling door een normale verdeling

Onder zekere voorwaarden kan men een binomiaal verdeling *benaderen* door een normale verdeling. Meestal stelt men als voorwaarde dat zowel $np > 5$ als $nq > 5$, of ook wel dat het interval $[np - 3\sqrt{npq}; np + 3\sqrt{npq}]$ geheel ligt in het interval $[0; n]$. Cumulatieve binomiaal kansen worden dan als volgt benaderd (Φ is de cumulatieve dichtheidsfunctie van de standaardnormale verdeling):

$$P(Y \le k) \approx \Phi\left(\frac{k + 0{,}5 - np}{\sqrt{npq}}\right), \quad \text{en} \qquad P(Y \ge k) \approx 1 - \Phi\left(\frac{k - 0{,}5 - np}{\sqrt{npq}}\right)$$

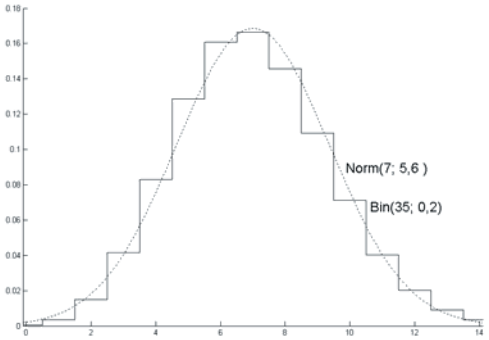
Een gevolg hiervan is

$$P(Y = k) = P(Y \le k) - P(Y \le k - 1) \approx \Phi\left(\frac{k + 0{,}5 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{k - 0{,}5 - np}{\sqrt{npq}}\right)$$

Het optellen, respectievelijk aftrekken, van 0,5, wordt ook wel **continuïteitscorrectie** genoemd.

Figuur 7.1 illustreert de benadering van een binomiaal verdeling met parameters $n = 35$ en $p = 0{,}2$ door een normale verdeling met parameters $\mu = np = 7$ en $\sigma^2 = np(1-p) = 5{,}6$.

Figuur 7.1:
Normale benadering van een binomiaal verdeling



Voorbeeld 7.5 Vervolg van Voorbeeld 7.4. Na twee weken heeft de machine 200 chips geproduceerd waarvan er 40 defect blijken te zijn. Aangezien het interval $40 \pm 3 \times \sqrt{24} = [25{,}3;\ 54{,}7]$ geheel binnen [0; 200] ligt, is een normale benadering verantwoord. De overschrijdingskans bij het toetsen van $H_0 : p = 0{,}15$ vs. $H_1 : p > 0{,}15$ wordt als volgt bepaald

$$P(Y_0 \geq 40) \approx 1 - \Phi\left(\frac{40 - 0{,}5 - 200 \times 0{,}15}{\sqrt{200 \times 0{,}15 \times 0{,}85}}\right) = 1 - \Phi(1{,}8812) \approx 1 - 0{,}97 = 0{,}03$$

Indien $\alpha = 0{,}05$ kunnen we de nulhypothese nu verwerpen.

7.4 Punt- en intervalschatter

Indien we onder *n* individuele binomiaal experimenten *k* successen waarnemen, wordt de onbekende parameter *p* geschat door

$$\hat{p} = k / n$$

Als we het aantal successen opvatten als een stochastische variabele *K* (het aantal successen in een serie van *n* binomiaal experimenten is aan toeval onderhevig) dan is de **maximum likelihood schatter** $\hat{P} = K / n$ een **zuivere** (punt)**schatter** van *p* met variantie pq / n, d.w.z.

$$E(\hat{P}) = p, \quad \text{en} \qquad Var(\hat{P}) = \frac{1}{n^2} Var(K) = \frac{npq}{n^2} = \frac{pq}{n}$$

De kansverdeling van het steekproefgemiddelde $\hat{P} = K / n$ is exact gelijk aan

$$P(K = k) = P\left(\hat{P} = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0{,}1{,}2{,}...$$

M.a.w. het steekproefgemiddelde K/n neemt de waarde k/n aan met kans $\binom{n}{k}p^k(1-p)^{n-k}$. Voor 'grote' waarden van *n* (d.w.z. het interval $\hat{p} \pm 3 \times \sqrt{\hat{p}\hat{q}/n}$ ligt geheel in [0; 1]) is

$$\hat{Var}(\hat{P}) = \hat{\sigma}^2_{\hat{P}} = \hat{p}\hat{q}/n$$

een goede schatter van $Var(\hat{P})$ en heeft $\hat{P}$ bij benadering een normale verdeling. Dit stelt ons in staat om een **betrouwbaarheidsinterval** voor *p* op te stellen (ook wel **intervalschatting** van *p* genoemd). We zeggen dat het interval met als grenzen

$$\hat{p} \pm z_{1-\alpha/2} \times \sqrt{\hat{p} \times \hat{q}/n}$$

met een **betrouwbaarheid** van $100 \times (1-\alpha)\%$ de onbekende waarde *p* bevat.

Voorbeeld 7.6 Vervolg van Voorbeeld 7.5. Gegeven $n = 200$ en $\hat{p} = 40/200 = 0{,}2$, is een 90% betrouwbaarheidsinterval voor *p*: $0{,}2 \pm 1{,}645 \times \sqrt{0{,}2 \times 0{,}8/200}$, ofwel $[0{,}153;\ 0{,}452]$. Ook deze analyse suggereert $p > 0{,}15$.

8. Poisson verdelingen Pois(λ)

8.1 Kansverdeling, parameter, verwachtingswaarde en variantie

De Poisson verdeling wordt vaak gebruikt als model voor het aantal incidenten van een zeker soort in een zeker tijdsinterval, zoals bijvoorbeeld het aantal klanten dat zich per uur bij een bepaald loket meldt. De kans dat zo'n Poisson verdeelde variabele Y de waarde k aanneemt wordt gegeven door

$$P(Y=k)=\frac{e^{-\lambda}\lambda^{k}}{k!};\lambda>0,\quad k=0,1,2,...$$

Een **Poisson verdeling** wordt dus gekenmerkt door de **parameter** λ die elke waarde groter dan 0 mag hebben. Als de kansverdeling van een stochastische variabele Y een Poisson verdeling is dan schrijven we

$$Y\sim \text{Pois}(\lambda)$$

Belangrijke karakteristieken van zo'n Poisson variabele zijn

verwachtingswaarde	$\mu = \lambda$
variantie	$\sigma^2 = \lambda$
standaardafwijking	$\sigma = \sqrt{\lambda}$

Als de stochastische variabelen $X_1,...,X_n$ onderling onafhankelijk zijn met $X_i \sim \text{Pois}(\lambda_i)$, $i = 1,...,n$, dan heeft de som $Y = \sum_{i=1}^{n} X_i$ weer een Poisson verdeling en wel met parameter $\lambda = \sum_{i=1}^{n} \lambda_i$. Een bijzonder geval hiervan is de steekproefsituatie $X_i \sim \text{Pois}(\lambda)$, $i = 1,...,n$, waarbij $\sum_{i=1}^{n} X_i \sim$ Pois($n\lambda$).

8.2 Overschrijdingskansen

Als we beschikken over een waargenomen waarde, *k*, van een Poisson verdeelde variabele, *Y*, met onbekende parameter, λ, dan is (afhankelijk van de situatie) het toetsen van een drietal nul- en alternatieve hypothesen interessant. Onderstaande tabel vermeldt deze met de bijbehorende overschrijdingskansen; hierbij is $Y_0 \sim \text{Pois}(\lambda_0)$.

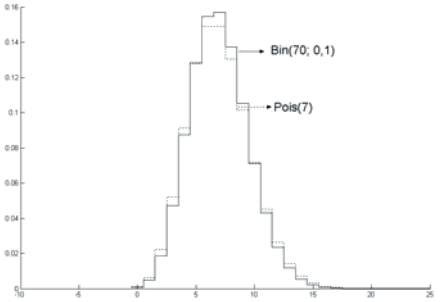
H_0	H_1	Overschrijdingskans
$\lambda \le \lambda_0$, of $\lambda = \lambda_0$	$\lambda > \lambda_0$	$P(Y_0 \ge k) = \sum_{i=k}^{\infty} e^{-\lambda_0} \lambda_0^i / i!$
$\lambda \ge \lambda_0$, of: $\lambda = \lambda_0$	$\lambda < \lambda_0$	$P(Y_0 \le k) = \sum_{i=0}^{k} e^{-\lambda_0} \lambda_0^i / i!$
$\lambda = \lambda_0$	$\lambda \ne \lambda_0$	$\sum_{i \in A} e^{-\lambda_0} \lambda_0^i / i!$, $A = \{j \mid P(Y_0 = j) \le P(Y_0 = k)\}$

Een nulhypothese wordt verworpen met onbetrouwbaarheidsdrempel α, indien de overschrijdingskans kleiner is dan α.

8.3 Het benaderen van een binomiaal verdeling door een Poisson verdeling

De kansverdeling van een binomiaal verdeelde stochast *X* met parameters *n* en *p* is voor *n* relatief groot en tegelijkertijd *p* relatief klein, in **benadering** gelijk aan de kansverdeling van een Poisson variabele *Y* met parameter $\lambda = np$. Als **vuistregel** kan men aannemen dat voldaan moet zijn aan $\lambda = np \le 7$ en $p \le 0{,}1$ voor een goede benadering. Het voordeel is natuurlijk dat een Poisson variabele gekenmerkt wordt door slechts 1 parameter, terwijl een binomiaal verdeling 2 parameters heeft. Figuur 8.1 illustreert de benadering van een binomiaal verdeling met parameters $n = 70$ en $p = 0{,}1$ door een Poisson verdeling met parameter $\lambda = np = 7$.

Figuur 8.1:
Poisson benadering van een binomiaal verdeling



Voorbeeld 8.1 Een verzekeringsmaatschappij heeft een groot aantal levensverzekeringen in haar portefeuille voor personen van allerlei leeftijden, en de kans dat een enkele polis tot uitkering komt gedurende een jaar is zeer klein.

8.4 Het benaderen van een Poisson verdeling door een normale verdeling

De kansverdeling van een Poisson verdeelde stochast X met parameter λ, gaat voor groter wordende waarden van λ steeds meer lijken op de kansverdeling van een normaalverdeelde variabele met gemiddelde $\mu = \lambda$ en variantie $\sigma^2 = \lambda$; dus

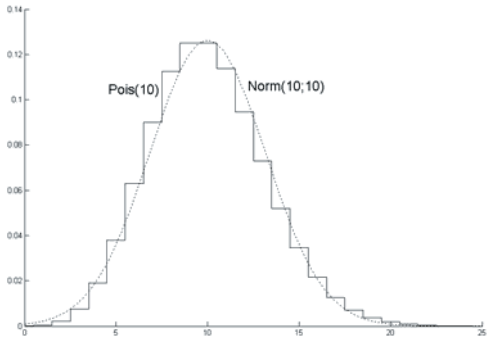
$$P(X \le k) \approx \Phi\left(\frac{k+0{,}5-\lambda}{\sqrt{\lambda}}\right), \text{ en} \qquad P(X \ge k) \approx 1-\Phi\left(\frac{k-0{,}5-\lambda}{\sqrt{\lambda}}\right)$$

met als gevolg

$$P(X=k) = P(X \le k) - P(X \le k-1) \approx \Phi\left(\frac{k+0{,}5-\lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{k-0{,}5-\lambda}{\sqrt{\lambda}}\right)$$

Als **vuistregel** dient voldaan te zijn aan $\lambda \ge 7$ voor een goede benadering. Figuur 8.2 illustreert de benadering van een Poisson verdeling met parameter $\lambda = np = 10$ door een normale verdeling met parameters $\mu = \sigma^2 = 10$.

Figuur 8.2:
Normale benadering van een Poisson verdeling



8.5 Punt- en intervalschatter

Als we beschikken over een aselecte steekproef $k_1, k_1, ..., k_n$ uit een Poisson verdeling met onbekende parameter λ, dan kan deze geschat worden door

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} k_i$$

Beschouwen we de aselecte steekproef als een serie onderling onafhankelijke gelijkverdeelde Poisson variabelen $K_1, ..., K_n$ met onbekende parameter λ, dan is de kansverdeling van $\sum_{i=1}^{n} K_i$ exact gegeven door (zie §8.1)

$$P\left(\sum_{i=1}^{n} K_i = k\right) = P\left(\frac{1}{n}\sum_{i=1}^{n} K_i = \frac{k}{n}\right) = \frac{e^{-n\lambda}(n\lambda)^k}{k!}, \quad k = 0,1,2,...$$

Het tweede '='-teken laat zien dat dit tegelijkertijd de steekproefverdeling is van het steekproefgemiddelde

$$\hat{\Lambda} = \frac{1}{n}\sum_{i=1}^{n} K_i$$

waarbij

$$E(\hat{\Lambda}) = E\left(\frac{1}{n}\sum_{i=1}^{n} K_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} K_i\right) = \frac{n\lambda}{n} = \lambda\,,$$
$$Var(\hat{\Lambda}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} K_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} K_i\right) = \frac{n\lambda}{n^2} = \lambda / n\,.$$

$\hat{\Lambda}$ is dus een **zuivere** (punt)**schatter** van λ met variantie λ / n. Voor grote waarden van $n\lambda$ (zie §8.4) geldt dat $\hat{\Lambda}$ bij benadering een normale verdeling heeft met gemiddelde λ en variantie λ / n heeft. Zo vinden we een **intervalschatting** voor λ. Het interval met als grenzen

$$\hat{\lambda} \pm z_{1-\alpha/2} \times \sqrt{\hat{\lambda} / n}$$

bevat de onbekende parameter λ met een betrouwbaarheid van $100 \times (1-\alpha)\%$.

9. Geometrische verdelingen Geo(p) en Negatief Binomiaal verdeling NB(r,p)

9.1 Geo(p): Kansverdeling, parameter, verwachtingswaarde en variantie

De kansen van een geometrisch verdeelde stochast nemen waarden aan die termen zijn van een meetkundige rij, vandaar de naam. De **modus** van een geometrische stochast, Y, is derhalve noodzakelijkerwijs 0. De **kansverdeling** wordt bepaald door 1 **parameter**, p:

$$P(Y = k) = p(1-p)^k, \quad 0 < p < 1; \; k = 0,1,2,...$$

Een geometrische variabele wordt vaak geïnterpreteerd als een soort 'wachttijd': als gedurende achtereenvolgende tijdseenheden de kans dat een bepaalde gebeurtenis optreedt (of: de kans op 'succes') gelijk is aan p (onafhankelijk van wat in andere tijdseenheden gebeurt), dan is de kans dat we k tijdseenheden moeten wachten *voordat* zo'n gebeurtenis optreedt gelijk aan $p(1-p)^k$. Als Y een geometrische verdeling heeft met parameter p, dan schrijven we

$$Y \sim \text{Geo}(p)$$

Belangrijke karakteristieken van zo'n geometrische variabele zijn

verwachtingswaarde	$\mu = (1-p)/p$
variantie	$\sigma^2 = (1-p)/p^2$
standaardafwijking	$\sigma = \sqrt{1-p}/p$

9.2 Cumulatieve geometrische kansen

Een bijzondere eigenschap van de geometrische verdeling is dat voor de **cumulatieve kansen** een concrete uitdrukking bestaat:

$$P(Y \geq k) = (1-p)^k$$

De gemiddelde wachttijd tot direct *na* het optreden van de eerste gebeurtenis gelijk is aan $E(Y+1) = 1/p$.

Voorbeeld 9.1 Stel je speelt (met 1 lot) iedere week mee met de Nederlandse Lotto. De kans dat je 6 getallen allemaal correct zijn, is

$$p = \binom{45}{6}^{-1} = \frac{6\times5\times4\times3\times2}{45\times44\times43\times42\times41\times40} = 0{,}000000123$$

De gemiddelde tijd tot het winnen van de hoofdprijs is $1/p = 8347680$ weken, ofwel ruim 1605 eeuwen.

9.3 NB(r,p): Kansverdeling, parameters, verwachtingswaarde en variantie

We behandelen de **negatief binomiaalverdeling** in hetzelfde hoofdstuk als de **geometrische verdeling** omdat de geometrische verdeling een speciaal geval is van de negatief binomiaalverdeling, en omdat de som van een aantal gelijkverdeelde, van elkaar onafhankelijke, geometrische variabelen een negatief binomiaalverdeling heeft. De kansverdeling van een negatief binomiaal verdeelde variabele

$$Y \sim \text{NB}(r, p)$$

wordt bepaald door 2 **parameters**, r en p:

$$P(Y=k)=\binom{r+k-1}{k}p^r(1-p)^k;\quad 0<p\le 1,\quad r=1,2,3,...,\quad k=0,1,2,...$$

Belangrijke karakteristieken van zo'n negatief binomiaal verdeelde variabele zijn

verwachtingswaarde	$\mu = r(1-p)/p$
variantie	$\sigma^2 = r(1-p)/p^2$
standaardafwijking	$\sigma = \sqrt{r(1-p)}/p$

Indien $r=1$ dan ontstaat het speciale geval van een geometrische variabele.

Voorbeeld 9.2 Voor een bepaald onderzoek heeft men mensen nodig die een bepaalde afwijking hebben in hun DNA-materiaal. De bedoelde afwijking kan worden vastgesteld door een relatief eenvoudig laboratoriumonderzoek en komt in Nederland voor met een frequentie van 1 op 500. Beschouw de uitslag van zo'n laboratoriumonderzoek als een 'failure' als de onderzochte persoon de bedoelde afwijking niet bezit. Het aantal failures, Y, voordat men uiteindelijk 10 mensen met de bedoelde eigenschap gevonden heeft, is een negatief binomiaalverdeelde variabele met parameters $r=10$ en $p=1/500$. Gemiddeld dient men dus $\mu=10\times(1-1/500)\times 500=4990$ mensen te onderzoeken voordat men er 10 heeft gevonden met de bedoelde afwijking.

Indien men beschikt over een aantal (onafhankelijke) observaties van een negatief binomiaal verdeelde variabele, $k_1,...,k_n$, dan kunnen de parameters r en p geschat worden m.b.v. het **steekproefgemiddelde** $\hat{\mu}=\bar{k}=\sum_{i=1}^{n}k_i/n$ en de **steekproefvariantie** $\hat{\sigma}^2=s^2=\sum(k_i-\bar{k})^2/(n-1)$:

$$\begin{cases}\hat{r} &= \hat{\mu}^2/(\hat{\sigma}^2-\hat{\mu})\\ \hat{p} &= \hat{\mu}/\hat{\sigma}^2\end{cases}$$

10. Hypergeometrische verdelingen HG(n,N,S)

10.1 Kansverdeling, parameters, verwachtingswaarde en variantie

Bij een binomiale verdeling wordt verondersteld dat de kans op succes voor elk binomiaal experiment gelijk is. Indien een steekproef wordt getrokken **met teruglegging** dan is hier altijd aan voldaan. Indien **zonder teruglegging** een steekproef uit een *grote* populatie wordt getrokken, blijft dit in benadering waar: de samenstelling van een populatie bestaande uit 100.000 objecten veranderd bijvoorbeeld nauwelijks als we daaruit een steekproef trekken van 30. Vandaar dat de binomiaal verdeling in veel gevallen toepasbaar is, ook als het gaat om een steekproef zonder teruglegging. Beschouw nu de situatie dat we een steekproef trekken uit relatief *kleine* populatie, bijvoorbeeld, we willen 5 mensen voor een project selecteren uit een groep van 20 mensen bestaande uit 12 mannen en 8 vrouwen, die allen gelijk gekwalificeerd zijn. Wat is nu de kans dat de geselecteerde groep bestaat uit 3 mannen en 2 vrouwen? Voor dit probleem hebben we de hypergeometrische verdeling nodig. Een stochastische variabele Y heeft een **hypergeometrische verdeling** als de kans op k 'successen' wordt gegeven door

$$P(Y=k)=\frac{\binom{S}{k}\binom{N-S}{n-k}}{\binom{N}{n}}=\frac{S!}{k!\times(S-k)!}\times\frac{(N-S)!}{(n-k)!\times(N-S-n+k)!}\Big/\frac{N!}{n!\times(N-n)!},\ 0\le S\le N, k=0,1,...,n$$

Hierin is N het totaal aantal objecten waarvan er S een bepaalde eigenschap hebben ('succes'); n is de grootte van de steekproef die zonder teruglegging getrokken wordt, en Y is het aantal successen in die steekproef. Er zijn dus 3 **parameters**: n, N en S, en de belangrijkste karakteristieken van een hypergeometrische variabele $Y\sim \mathrm{HG}(n,N,S)$ zijn

verwachtingswaarde	$\mu = nS/N$
variantie	$\sigma^2=\frac{nS(N-S)(N-n)}{N^2(N-1)}$
standaardafwijking	$\sigma=\frac{1}{N}\sqrt{\frac{nS(N-S)(N-n)}{(N-1)}}$

10.2 Voorbeelden hypergeometrische verdeling

Voorbeeld 10.1 De kans p dat in het voorbeeld van §10.1 de geselecteerde groep bestaat uit 3 mannen en 2 vrouwen is:

$$p=\frac{\binom{8}{2}\binom{12}{3}}{\binom{20}{5}}=\frac{(8\times7)\times(12\times11\times10)\times(5\times4\times3\times2)}{2\times6\times(20\times19\times18\times17\times16)}\approx 0{,}397$$

Voorbeeld 10.2 Beschouw opnieuw Voorbeeld 9.1. De kans p dat precies 5 van de 6 getallen goed zijn, is

$$p = \frac{\binom{6}{5}\binom{45-6}{6-5}}{\binom{45}{6}} = \frac{6\times 39\times 6\times 5\times 4\times 3\times 2}{45\times 44\times 43\times 42\times 41\times 40} \approx 0{,}0000287$$

zodat gemiddeld pas na ruim 6,7 eeuwen een prijs behorende bij '5 goed' te incasseren valt. Als je doet alsof 'met teruglegging' van toepassing is, dan zou je de *onjuiste* uitkomst vinden

$$p = \binom{6}{5}\left(\frac{6}{45}\right)^5\left(\frac{39}{45}\right)^1 \approx 0{,}00022$$

dus ruim een factor 7 te groot.

10.3 Het benaderen van een hypergeometrische verdeling

Indien de parameters van een hypergeometrische verdeling voldoen aan

$$\begin{cases} n/N < 0{,}1 \\ S/N < 0{,}1 \end{cases}$$

dan kan deze benaderd worden door een binomiaal verdeling met parameters $n = n$ en $p = S/N$ en door een Poisson verdeling met parameter $\lambda = nS/N$.

Indien de parameters van een hypergeometrische verdeling voldoen aan

$$\begin{cases} nS/N > 5 \\ n - nS/N > 5 \end{cases}$$

dan kan deze benaderd worden door een normale verdeling met parameters $\mu = nS/N$ en $\sigma^2 = \dfrac{nS(N-S)(N-n)}{N^2(N-1)}$

11. Multinomiaal verdelingen Mult(n, p1, … , pr)

11.1 Kansverdeling, parameters, verwachtingswaarde en variantie

De **multinomiaal** verdeling is een generalisatie van de **binomiaal** verdeling (zie ook §1.5 en Hoofdstuk 7). Een multinomiaal verdeelde variabele Y is een stochastische **vector** die waarden kan aannemen uit de verzameling

$$\left\{(k_1,...,k_r)\,|\,k_i \in \{0,1,...,n\};\ \sum_{i=1}^{r} k_i = n\right\}$$

De **kansverdeling** van een multinomiaal verdeelde variabele $Y \sim \text{Mult}(n, p_1,..., p_r)$ wordt gegeven door

$$P\big(Y=(k_1,...,k_r)\big)=\binom{n}{k_1\ ...\ k_r}\times p_1^{k_1}\times ...\times p_r^{k_r}$$

Er zijn dus $r+1$ **parameters**: $n, p_1,..., p_r$. De uitkomsten van binomiaal experimenten worden geclassificeerd als 'succes' of 'failure', waarbij p de kans is op 'succes' en $1-p$ de kans op 'failure'. Bij **multinomiaal experimenten** zijn er *meer* dan 2 mogelijke uitkomsten; uitkomst *i* treedt op met kans p_i. Als we een multinomiaal verdeelde variabele voorstellen als een **stochastische vector** $Y=(K_1,...,K_r)$ dan heeft component K_i als **verwachtingswaarde** $\mu_i = E(K_i)=n\times p_i$ en **variantie** $\sigma_i = Var(K_i)=n\times p_i\times(1-p_i)$.

Voorbeeld 11.1 We gooien met 6 dobbelstenen. De kans, *p*, dat de worp precies (k1=) 3 zessen en (k2=) 2 vijven bevat (en dus precies (k1=) 1 resultaat uit {1,2,3,4}), kan worden bepaald m.b.v. de multinomiaal verdeling met de parameters n=6, p1=p2=1/6, p3=4/6:

$$p=\binom{6}{3\ \ 2\ \ 1}\left(\frac{1}{6}\right)^3\left(\frac{1}{6}\right)^2\left(\frac{4}{6}\right)^1\approx 0{,}005$$

12. Uniforme (of rechthoekige) verdelingen U(a,b)

12.1 Kansdichtheidsfunctie, cumulatieve verdelingsfunctie, parameters, verwachtingswaarde en variantie

De eenvoudigste continue kansverdeling is de **uniforme verdeling**. De praktische toepassingen zijn echter beperkt. Een variabele, X, heeft een uniforme verdeling met **parameters** a en b als de kansdichtheidsfunctie gegeven is door

$$f(x) = \frac{1}{b-a} \qquad \text{voor } a \le x \le b$$

$$f(x) = 0 \qquad \text{voor } x < a \text{ of } x > b$$

De uniforme verdeling wordt ook wel **rechthoekige verdeling** genoemd vanwege de rechthoekige vorm van de kansdichtheidsfunctie $f(x)$. De **cumulatieve verdelingsfunctie** is

$$F(x) = 0 \qquad \text{voor } x < a$$

$$F(x) = \frac{b-x}{b-a} \qquad \text{voor } a \le x \le b$$

$$F(x) = 1 \qquad \text{voor } x > b$$

Belangrijke karakteristieken zijn

verwachtingswaarde	$\mu = (a+b)/2$
variantie	$\sigma^2 = (b-a)^2/12$
standaardafwijking	$\sigma = (b-a)/\sqrt{12}$

Voorbeeld 12.1 Als men geïnteresseerd is in de positie van het ventiel in het voorwiel na een lange fietstocht, dan zou de uniforme verdeling op het interval $[0;2\pi]$ een goed model zijn. Het ventiel helemaal bovenaan correspondeert dan bijvoorbeeld met 0, helemaal onderaan met π.

13. Exponentiële verdeling Exp(λ)

13.1 Kansdichtheidsfunctie, cumulatieve verdelingsfunctie, parameter, verwachtingswaarde en variantie

De exponentiële verdeling is een **continue verdeling** die bijvoorbeeld zijn toepassing vindt als een model voor de levensduur van allerlei zaken. Er is een belangrijke relatie tussen de **Poisson verdeling** en de **exponentiële verdeling**: indien de tijd tussen twee opeenvolgende incidenten een exponentiële verdeling heeft dan kan worden aangetoond dat het aantal incidenten in een zeker tijdsinterval een Poisson verdeling heeft. Ook omgekeerd geldt dat, wanneer het aantal incidenten in een zeker tijdsinterval een Poisson verdeling heeft, de tussentijden tussen opeenvolgende incidenten exponentieel verdeeld is. Een exponentieel verdeelde variabele $X \sim \text{Exp}(\lambda)$ heeft 1 **parameter**, $\lambda > 0$. De **kansdichtheidsfunctie** en **cumulatieve verdelingsfunctie** zijn, respectievelijk

$$f(x) = \lambda \exp(-\lambda x)$$

$$F(x) = 1 - \exp(-\lambda x)$$

Merk op dat $\exp(x)$ en e^x verschillende notaties zijn voor dezelfde (exponentiële) functie.

Belangrijke **karakteristieken** zijn

verwachtingswaarde	$\mu = 1/\lambda$
variantie	$\sigma^2 = 1/\lambda^2$
standaardafwijking	$\sigma = 1/\lambda$

Voorbeeld 13.1 Bij de productie van steenwol worden enorme ovens gebruikt waarin de gebruikte steensoorten, met allerlei toevoegingen, worden gesmolten en gesponnen tot wol. De productie gaat dag en nacht door. Indien het productieproces stagneert, koelt de oven af en kan het proces slechts met grote kosten weer opgestart worden. Stel dat de tijd tussen twee productiestoringen een exponentiële verdeling heeft met parameter $\lambda = 0.05\,\text{dag}^{-1}$, zodat gemiddeld na 20 dagen een productiestoring optreedt. De kans dat het proces gedurende minstens 30 dagen onafgebroken blijft werken, wordt gegeven door

$$1 - F(30) = \exp(-0.05 \times 30) \approx 0.22$$

Het gemiddeld aantal keren per jaar (=365 dagen) dat er een productiestoring optreedt is $365 \times \lambda = 18{,}25$ keer. Het aantal productiestoringen per jaar, Y, heeft een Poisson verdeling:

$$Y \sim \text{Pois}(18{,}25)$$

14. Normale verdeling N(μ; σ^2)

14.1 Kansdichtheidsfunctie, cumulatieve verdelingsfunctie, parameters, verwachtingswaarde en variantie

De **normale verdeling** is een van de belangrijkste continue kansverdelingen, zo niet de belangrijkste. Het belang ervan wordt vooral door de **Centrale Limiet Stelling** aangetoond (zie §4.3). De kansverdeling van vele stochastische variabelen blijkt een normale verdeling te zijn, of althans zeer veel te lijken op een normale verdeling. Veel technieken die in de statistiek worden toegepast maken daarom gebruik van de normale verdeling. De **verwachtingswaarde** μ en de **variantie** σ^2 zijn de **parameters** van een normaal verdeelde variabele X; we schrijven $X \sim \mathrm{N}(\mu;\sigma^2)$. De **kansdichtheidsfunctie** wordt gegeven door

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

Deze vergelijking laat zien dat $f(x)$ symmetrisch is t.o.v. $x = \mu$. Voor de **cumulatieve verdelingsfunctie** bestaat geen expliciete functie, maar is per definitie gegeven door

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \mathrm{d}x$$

Een lineaire functie $Y = aX + b$ ($a \neq 0$) van een normaal verdeelde variabele $X \sim \mathrm{N}(\mu;\sigma^2)$ is ook weer normaal verdeeld: $Y \sim \mathrm{N}\left(a\mu + b, (a\sigma)^2\right)$. De som van een aantal onafhankelijke normaal verdeelde variabelen $Y_i \sim \mathrm{N}\left(\mu_i;\sigma_i^2\right)$ heeft ook weer een normale verdeling: $\sum Y_i \sim \mathrm{N}\left(\sum \mu_i;\sum \sigma_i^2\right)$.

14.2 De standaardnormale verdeling

Een veel gebruikte normale verdeling is $\mathrm{N}(0;1)$, de zogenaamde **standaardnormale verdeling** met verwachtingswaarde $\mu = 0$ en variantie $\sigma^2 = 1$. De corresponderende **kansdichtheidsfunctie**, $\varphi(x)$, is

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

De **cumulatieve verdelingsfunctie** noteren we met $\Phi(x)$. Merk op dat geldt

$$\varphi(x) = \varphi(-x)$$

$$\Phi(x) = 1 - \Phi(-x)$$

We zullen, voor een gemakkelijke herkenning, een standaardnormaal verdeelde variabele vaak aangeven met de letter *Z* (en afzonderlijke waarden met *z*). Eén van de eigenschappen van de normale verdeling is dat een kansprobleem waar een normale verdeling mee gemoeid is, ook kan worden opgelost met de standaardnormale verdeling. Zo geldt voor bovenvermelde functie $F(x)$, de cumulatieve verdelingsfunctie van $X \sim \mathrm{N}(\mu;\sigma^2)$

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

en, voor de kans dat *X* een waarde aanneemt tussen *a* en *b*:

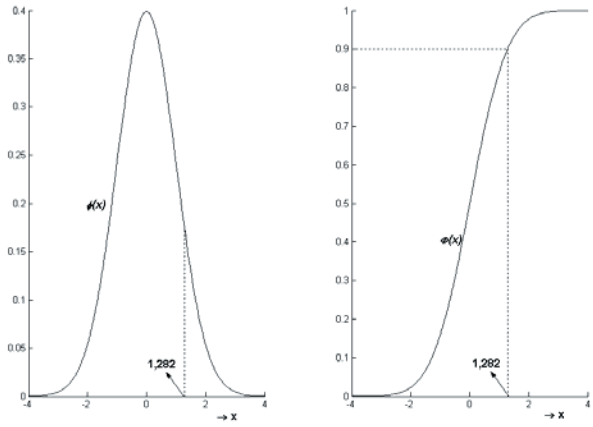
$$P(a < X < b) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Indien geen statistische software voorhanden is, gebruikt men tabellen. Voor de normale verdeling hebben we slechts één tabel nodig, die van de standaardnormale verdeling (zie Appendix C1). Zie Tabel 14.1 voor enkele waarden uit de tabel voor $\Phi(x)$ die vaak terugkomen bij het toetsen van hypothesen en het berekenen van betrouwbaarheidsintervallen; Figuur 14.1 geeft de grafiek van de kansdichtheids- en verdelingsfunctie van de standaardnormale verdeling en de punten $\left[1{,}282;\ \varphi(1{,}282)\right]$ en $\left[1{,}282;\ \Phi(1{,}282)\right]$ ter illustratie.

Tabel 14.1:
Enkele belangrijke waarden van de cumulatieve standaardnormale verdeling.

z	$\Phi(z) = P(Z < z)$	$P(-z < Z < z) = P(\mid Z \mid < z) = 2 \times (\Phi(z) - 0{,}5)$
1,000	0,841	0,683
1,282	0,900	0,800
1,645	0,950	0,900
1,960	0,975	0,950
2,326	0,990	0,980
3,090	0,999	0,998

Figuur 14.1:
De standaardnormale verdeling.



Zo kan ook bepaald worden dat voor een algemene normaal verdeelde variabele $X \sim \mathrm{N}(\mu; \sigma^2)$ geldt

$$P(\mu - \sigma < X < \mu + \sigma) = P(-1 < Z < 1) \approx 0{,}68$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < Z < 2) \approx 0{,}95$$

De waarde van z die voldoet aan $\Phi(z) = 1 - \alpha$ wordt meestal genoteerd als z_α zodat $\Phi(z_\alpha) = 1 - \alpha$ en $\Phi(z_{1-\alpha}) = \alpha$. In het algemeen geldt dus

$$P\left(\mu - z_{\alpha/2}\sigma < X < \mu + z_{\alpha/2}\sigma\right) = P\left(\left|\frac{X - \mu}{\sigma}\right| < z_{\alpha/2}\right) = P(\left|Z\right| < z_{\alpha/2}) = 1 - \alpha$$

14.3 Punt- en intervalschatter voor μ, puntschatter voor σ²

Als we beschikken over een aselecte steekproef $(x_1, ..., x_n)$ uit een normale verdeling met onbekende parameters μ en σ^2, dan is het **steekproefgemiddelde**

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

de gebruikelijke **schatting** voor μ. Als we de aselecte steekproef beschouwen als een **stochastische vector** $(X_1,...,X_n)$ dan is

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

een **zuivere schatter** (de maximum likelihood schatter) voor μ, d.w.z. $E(\bar{X}) = \mu$. Voor de variantie van $\bar{X}$ geldt $Var(\bar{X}) = \sigma^2 / n$, en voor de verdeling van $\bar{X}$, de zogenaamde **steekproefverdeling** van $\bar{X}$, geldt $\bar{X} \sim N(\mu; \sigma^2 / n)$. De **steekproefvariantie**

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

is een **zuivere schatter** van σ^2 (en $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$ een **schatting** van σ^2). De wortel uit de steekproefvariantie is natuurlijk de **steekproefstandaardafwijking**.

Een **intervalschatter** (of: **betrouwbaarheidsinterval**) voor μ op basis van de aselecte steekproef $(X_1,...,X_n)$ kan worden bepaald afhankelijk van de situatie dat σ bekend is of onbekend. Indien σ *bekend* is, geldt

$$P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right) = P(\mu - z_{\alpha/2}\sigma/\sqrt{n} < \overline{X} < \mu + z_{\alpha/2}\sigma/\sqrt{n}) = 1-\alpha$$

zodat

$$P(\overline{X} - z_{\alpha/2}\sigma/\sqrt{n} < \mu < \overline{X} + z_{\alpha/2}\sigma/\sqrt{n}) = 1-\alpha$$

We noemen

$$(\overline{X} - z_{\alpha/2}\sigma/\sqrt{n}\,;\, \overline{X} + z_{\alpha/2}\sigma/\sqrt{n})$$

een $100(1-\alpha)\%$ -**betrouwbaarheidsinterval** voor μ; dat interval bevat met een **betrouwbaarheid** van $100(1-\alpha)\%$ de onbekende waarde μ.

Indien σ *onbekend* is, dient deze ook uit de steekproef geschat te worden. We vervangen dan σ door de steekproefstandaardafwijking, *s*, maar moeten tegelijkertijd $z_{\alpha/2}$ vervangen door de corresponderende waarde, $t_{n-1;\,\alpha/2}$, uit de **Student's *t*-verdeling** (zie §15.2 en Appendix C3) met $n-1$ vrijheidsgraden om te compenseren voor het feit dat *s* slechts een schatting is van σ; het **betrouwbaarheidsinterval** wordt daardoor iets groter (geeft dus minder informatie over μ):

$$(\overline{X} - t_{n-1;\,\alpha/2}s/\sqrt{n}\,;\, \overline{X} + t_{n-1;\,\alpha/2}s/\sqrt{n})$$

We maken bij de constructie van dit interval gebruik van het feit dat $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ een Student's *t*-verdeling heeft.

14.4 Intervalschatter voor σ^2

Ook voor de constructie van een betrouwbaarheidsinterval voor σ^2 kunnen we twee gevallen onderscheiden, namelijk of μ bekend is of onbekend. Gezien het feit dat μ doorgaans onbekend is beperken we ons hiertoe. In dat geval heeft

$$\frac{(n-1)S^2}{\sigma^2}$$

een zogenaamde χ^2-verdeling met $n-1$ vrijheidsgraden (zie §15.1 en Appendix C2). Het gebruikelijke $100(1-\alpha)\%$ -**betrouwbaarheidsinterval** voor σ^2 wordt gegeven door

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}} ; \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \right)$$

15. Verdelingen gerelateerd aan de normale verdeling

15.1 χ2 –verdelingen ("chi-kwadraat")

Als $Z_i \sim \mathrm{N}(0;1)$, $i = 1,...,n$, van elkaar onafhankelijke standaardnormaal verdeelde variabelen zijn dan geldt dat de som

$$Y = \sum_{i=1}^{n} Z_i^2$$

een χ^2-verdeling (spreek uit: "**chi-kwadraat** verdeling") heeft met *n* **vrijheidsgraden** (notatie: $Y \sim \chi_n^2$). De (enige) **parameter** van zo'n verdeling is dus *n* ($n = 1,2,3,...$), het aantal vrijheidsgraden, vaak ook aangegeven met *df* ("***d*egrees of *f*reedom**"). **Verwachtingswaarde** en **variantie** zijn

$$E(Y) = \mu_Y = n$$

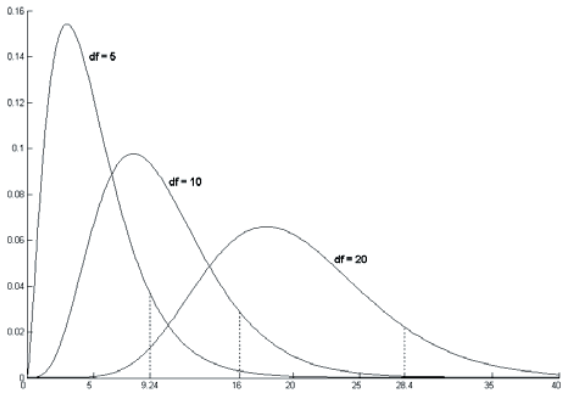
$$Var(Y) = \sigma_Y^2 = 2n$$

De **kansdichtheidsfunctie** wordt gegeven door

$$f(x) = c_n x^{n/2-1} e^{-x/2}, \ (0 < x < \infty)$$

waarbij c_n een constante is die alleen van *n* afhangt. De **cumulatieve verdelingsfunctie** is voor verschillende waarden van *n* getabelleerd in Appendix C2. De χ^2 -verdeling wordt bijvoorbeeld toegepast in §16.3. Figuur 15.1 geeft de grafiek van de kansdichtheidsfuncties van enkele χ^2 -verdelingen; bovendien worden de 90%-punten aangegeven (vgl. Appendix C2).

Figuur 15.1:
χ^2 -verdelingen met 5, 10 en 20 vrijheidsgraden



15.2 Student's t –verdelingen

Indien $X \sim \mathrm{N}(\mu;\sigma^2)$ en $(X_1,...,X_n)$ een aselecte steekproef is uit deze verdeling, dan geldt dat

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

een zogenaamde **Student's *t*-verdeling** heeft met $n-1$ **vrijheidsgraden** (notatie: $T \sim t_{n-1}$). Hierbij is $\overline{X}$ het **steekproefgemiddelde** en *S* de **steekproefstandaardafwijking** (zie ook §14.3).
Verwachtingswaarde en **variantie** zijn respectievelijk

$$E(T) = \mu_T = 0\,,\ (n>1)$$

$$Var(T) = \sigma_T^2 = \frac{n}{n-2},\ \ (n>2)$$

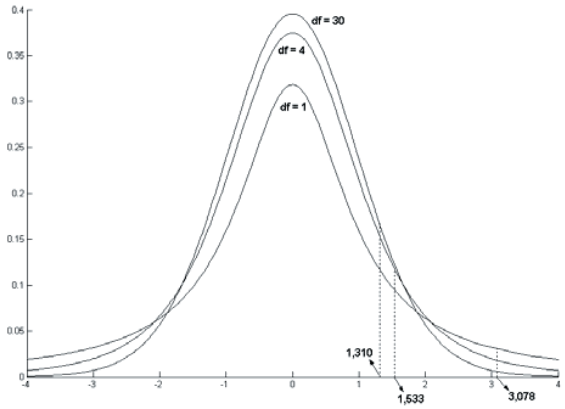
De **kansdichtheidsfunctie** wordt gegeven door

$$f(x) = \frac{c_n}{(1+x^2/n)^{(n+1)/2}},\ (-\infty < x < \infty)$$

waarbij c_n een constante is die alleen van *n* afhangt. De **cumulatieve verdelingsfunctie** is voor verschillende waarden van *n* getabelleerd in Appendix C3. Belangrijke eigenschap van de Student's

t-verdeling is dat deze steeds meer op de standaardnormale verdeling gaat lijken naarmate het aantal vrijheidsgraden groter wordt. Figuur 15.2 geeft de grafiek van de *t*-verdelingen met 1, 4 en 30 vrijheidsgraden en de corresponderende 90%-punten.

Figuur 51.2:
Student's *t*-verdelingen met 1, 4 en 30 vrijheidsgraden.



15.3 Fisher's F –verdelingen

Indien Y_1 en Y_2 van elkaar onafhankelijke χ^2-verdeelde variabelen zijn met respectievelijk n_1 en n_2 vrijheidsgraden, dan heeft

$$F = \frac{Y_1 / n_1}{Y_2 / n_2}$$

een zogenaamde ***F*-verdeling** met n_1 en n_2 **vrijheidsgraden** (notatie: $F \sim F_{n_2}^{n_1}$; n_1 noemen we ook het aantal vrijheidsgraden van de *teller*, en n_2 van de *noemer*). In het bijzonder geldt dat wanneer $(U_1,...,U_{m+1})$ en $(V_1,...,V_{n+1})$ onafhankelijke aselecte steekproeven zijn uit de normale verdelingen, $\mathrm{N}(\mu_U;\sigma_U^2)$ en $\mathrm{N}(\mu_V;\sigma_V^2)$, met steekproefvarianties, S_U^2 en S_V^2, het quotiënt

$$Q = \frac{S_U^2 / \sigma_U^2}{S_V^2 / \sigma_V^2}$$

een *F*-verdeling heeft met *m* en *n* vrijheidsgraden, dus $Q \sim F_n^m$. **Verwachtingswaarde** en **variantie** zijn respectievelijk

$$E(Q) = \mu_Q = \frac{n}{n-2},\ (n > 2)$$

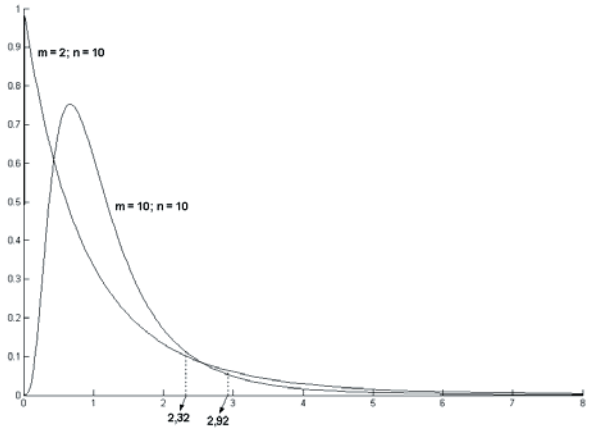
$$Var(Q) = \sigma_Q^2 = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)},\ \ (n > 4)$$

Merk op dat de verwachtingswaarde alleen van *n* afhangt. De **kansdichtheidsfunctie** is

$$f(x) = c_{m,n} \frac{x^{(m-2)/2}}{(n+mx)^{(m+n)/2}}, \ (0 < x < \infty)$$

waarbij $c_{m,n}$ een constante is die alleen van *m* en *n* afhangt. De **cumulatieve verdelingsfunctie** is voor verschillende waarden van *n* getabelleerd in Appendix C4. Figuur 15.3 geeft de grafiek van de *F*-verdelingen met respectievelijk $m = 2$, $n = 10$, en $m = 10$, $n = 10$, en de corresponderende 90%-punten.

Figuur 15.3:
Fisher's *F*-verdelingen met $m = 2$, $n = 10$, en $m = 10$, $n = 10$ vrijheidsgraden.



16. Op de normaalverdeling gebaseerde toetsen en betrouwbaarheidsintervallen

16.1 1 steekproef, σ bekend en/of n groot; H_0: μ = μ_0

Als we beschikken over een aselecte steekproef $X_1,...,X_n$ uit een normaalverdeelde populatie met onbekende verwachtingswaarde μ en bekende standaardafwijking σ, dan geldt $\overline{X} = \sum_{i=1}^{n} X_i / n$ $\sim \mathrm{N}\left(\mu; \sigma^2 / n\right)$. Stel dat in een concreet experiment $\overline{X} = \overline{x}$. In Tabel 16.1 zijn voor het toetsen van het gebruikelijke drietal nul- en alternatieve hypothesen m.b.t. de onbekende parameter, μ, de bijbehorende **overschrijdingskansen** vermeld. Als H_0 waar is, d.w.z. $\mu = \mu_0$, dan heeft de **toetsingsgrootheid** $Z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$ een standaardnormaal verdeling. *Z* wordt verkregen door $\overline{X}$ te **standaardiseren** onder de veronderstelling dat $\mu = \mu_0$. De **waarde** van de toetsingsgrootheid *Z* in de concrete situatie $\overline{X} = \overline{x}$ noteren we met $z = \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$.

Tabel 16.1:
Het bepalen van de overschrijdingskans van de waargenomen waarde van de toetsingsgrootheid.

H_0	H_1	**Overschrijdingskans**
$\mu \le \mu_0$, of: $\mu = \mu_0$	$\mu > \mu_0$	$P\left(\overline{X} > \overline{x} \,\middle|\, \mu = \mu_0\right) = P\left(Z > z\right) = 1 - \Phi(z)$
$\mu \ge \mu_0$, of: $\mu = \mu_0$	$\mu < \mu_0$	$P\left(\overline{X} < \overline{x} \,\middle|\, \mu = \mu_0\right) = P\left(Z < z\right) = \Phi(z)$
$\mu = \mu_0$	$\mu \ne \mu_0$	$P\left(\lvert\overline{X}\rvert > \lvert\overline{x}\rvert \,\middle|\, \mu = \mu_0\right) = P\left(\lvert Z\rvert > \lvert z\rvert\right) = 2\left(1 - \Phi\left(\lvert z\rvert\right)\right)$

Een nulhypothese wordt verworpen met **onbetrouwbaarheidsdrempel** α, indien de overschrijdingskans kleiner is dan α. Men komt tot dezelfde conclusie indien men verifieert of de waarde van de toetsingsgrootheid in het **verwerpingsgebied** ligt. In dit geval kan het verwerpingsgebied worden gekarakteriseerd door de **kritieke grens**, vergelijk Tabel 16.2.

Tabel 16.2:
Het verwerpen van de nulhypothese met onbetrouwbaarheidsdrempel α door de waarde van de toetsingsgrootheid te vergelijken met de kritieke grens.

H_0	H_1	**Kritieke grens**	Verwerp H_0 als
$\mu \le \mu_0$, of: $\mu = \mu_0$	$\mu > \mu_0$	$z_\alpha = \Phi^{-1}(1-\alpha)$	$z > z_\alpha$
$\mu \ge \mu_0$, of: $\mu = \mu_0$	$\mu < \mu_0$	$-z_\alpha = \Phi^{-1}(\alpha)$	$z < -z_\alpha$
$\mu = \mu_0$	$\mu \ne \mu_0$	$z_{\alpha/2} = \Phi^{-1}\left(1-\alpha/2\right)$	$\lvert z\rvert > z_{\alpha/2}$

Indien σ *onbekend* is maar de steekproefomvang, *n*, is groot genoeg, dan kunnen, nadat σ is vervangen door de steekproefstandaardafwijking, *s*, dezelfde regels gebruikt worden. Vanwege de **Centrale Limiet Stelling** mag de steekproef zelfs afkomstig zijn uit een andere verdeling dan de normaalverdeling. De onnauwkeurigheid die ontstaat door aan te nemen dat $\overline{X}$ een normaalverdeling heeft, is kleiner naarmate *n* groter is, maar over het algemeen is de gevolgde procedure voor $n > 30$ (**vuistregel**) nauwkeurig genoeg. De kritieke grenzen kunnen worden bepaald met speciale software (zoals EXCEL) of in een tabel worden opgezocht (zie Appendix C1). Voor het bepalen van een **betrouwbaarheidsinterval** voor μ wordt verwezen naar §14.3.

Voorbeeld 16.1: Een aselecte steekproef van 50 waarnemingen levert de volgende steekproefgrootheden op: $\bar{x} = 25$ en $s^2 = 55$. We willen weten of er op grond van deze steekproef reden is om aan te nemen dat het gemiddelde, μ, van de populatie kleiner is dan 27. We toetsen dus de hypothese $H_0 : \mu = 27$ vs. $H_1 : \mu < 27$. De toetsingsgrootheid is $Z = \left(\overline{X} - 27\right) / \sqrt{55/50}$; de **waarde** van de toetsingsgrootheid is $z = (25 - 27) / \sqrt{55/50} \approx -1{,}91$. Stel dat we als **fout van de eerste soort** $\alpha = 0.05$ accepteren. Omdat de **overschrijdingskans** gelijk is aan $\Phi(-1{,}91) \approx 0{,}028$, dus kleiner dan α, verwerpen we H_0. Tot dezelfde conclusie komen we als we vaststellen dat de waarde van de toetsingsgrootheid kleiner is dan de **kritieke grens** $-1{,}645 = \Phi^{-1}(0{,}05)$.

16.2 1 steekproef, σ onbekend en n klein; H_0: μ = μ_0

Indien we opnieuw de situatie van §16.1 beschouwen, maar nu voor relatief kleine steekproeven (**vuistregel**: $n \le 30$) en voor σ onbekend, dan kunnen we vergelijkbare, op normaliteit gebaseerde, toetsen uitvoeren onder de voorwaarde dat de aselecte steekproef $X_1,...,X_n$ uit een normaalverdeelde populatie komt (in de praktijk volstaat dat $\overline{X}$ *in benadering* een normaalverdeling heeft). De steekproefstandaardafwijking *S* wordt gebruikt als schatter van de onbekende σ. De **toetsingsgrootheid**, $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$, heeft een **Student's *t*-verdeling** met $n-1$ **vrijheidsgraden** (vergelijk §15.2) en deze verdeling dient nu bij het bepalen van de overschrijdingskansen en kritieke waarden gebruikt te worden in plaats van de standaardnormaal verdeling. Als we de waarde van T in een concrete situatie noteren met $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$, dan zijn de volgende tabellen analoog aan de tabellen in de vorige paragraaf.

Tabel 16.3:
Het bepalen van de overschrijdingskans van de waargenomen waarde van de toetsingsgrootheid indien *n* relatief klein en $\overline{X}$ een normaalverdeling heeft.

H_0	H_1	**Overschrijdingskans**
$\mu \le \mu_0$, of: $\mu = \mu_0$	$\mu > \mu_0$	$P(\overline{X} > \overline{x} \mid \mu = \mu_0) = P(T > t)$
$\mu \ge \mu_0$, of: $\mu = \mu_0$	$\mu < \mu_0$	$P(\overline{X} < \overline{x} \mid \mu = \mu_0) = P(T < t)$
$\mu = \mu_0$	$\mu \ne \mu_0$	$P(\lvert\overline{X}\rvert > \lvert\overline{x}\rvert \mid \mu = \mu_0) = P(\lvert T\rvert > \lvert t\rvert)$

Tabel 16.4: Het verwerpen van de nulhypothese met onbetrouwbaarheidsdrempel α door de waarde van de toetsingsgrootheid te vergelijken met de kritieke grens.

H_0	H_1	**Kritieke grens**	Verwerp H_0 als
$\mu \le \mu_0$, of: $\mu = \mu_0$	$\mu > \mu_0$	$t_{n-1;\alpha}$	$t > t_{n-1;\alpha}$
$\mu \ge \mu_0$, of: $\mu = \mu_0$	$\mu < \mu_0$	$-t_{n-1;\alpha}$	$t < -t_{n-1;\alpha}$
$\mu = \mu_0$	$\mu \ne \mu_0$	$t_{n-1;\alpha/2}$	$\lvert t\rvert > t_{n-1;\alpha/2}$

De kritieke grenzen kunnen worden bepaald met speciale software (zoals EXCEL) of in een tabel worden opgezocht (zie Appendix C3). Een **betrouwbaarheidsinterval** voor μ kan worden gevonden met de in §14.3 beschreven methode voor onbekende σ.

Voorbeeld 16.2: We willen nagaan of het eindexamen Wiskunde A in een plattelandsgemeente significant slechter is gemaakt dan het landelijk gemiddelde en accepteren daarbij een onbetrouwbaarheidsdrempel $\alpha = 0.05$. Daartoe worden de punten voor Wiskunde A verzameld van een aselecte steekproef van 15 eindexamenleerlingen in het desbetreffende gebied. Het landelijk gemiddelde is gelijk aan 7,7. De steekproefwaarnemingen zijn

$$\{7{,}5;\ 6{,}1;\ 5{,}4;\ 4{,}2;\ 8{,}9;\ 6{,}7;\ 9{,}0;\ 6{,}9;\ 7{,}2;\ 8{,}4;\ 8{,}3;\ 6{,}4;\ 6{,}9;\ 7{,}7;\ 8{,}3\}$$

zodat $\bar{x} \approx 7{,}19$ en $s \approx 1{,}33$. We toetsen de hypothese $H_0 : \mu = 7{,}7$ vs. $H_1 : \mu < 7{,}7$. De **toetsingsgrootheid** is $T = (\overline{X} - 7{,}7)/(S/\sqrt{15})$ en heeft onder H_0 een t_{14}-verdeling; de **waarde** van de toetsingsgrootheid is $t = (7{,}19 - 7{,}7)/(1{,}33/\sqrt{15} \approx -1{,}47$. Omdat de **overschrijdingskans** gelijk is aan $P(T < -1{,}47) \approx 0{,}08$, dus groter dan α, verwerpen we H_0 *niet*. Tot dezelfde conclusie komen we als we vaststellen dat de waarde van de toetsingsgrootheid groter is dan de **kritieke grens** $-t_{14;0,05} = -1{,}76$. Op basis van dit onderzoek is er dus geen reden om aan te nemen dat in de desbetreffende plattelandsgemeente gemiddeld lagere punten worden gehaald voor wiskunde A dan het landelijk gemiddelde.

16.3 1 steekproef, onbekende verwachtingswaarde μ; H_0: $\sigma^2 = \sigma_0^2$

Als is gegeven een aselecte steekproef $X_1,...,X_n$ uit een normaalverdeelde populatie met onbekende verwachtingswaarde μ en onbekende standaardafwijking σ, dan gebruiken we de **schatter**

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

voor het schatten van de steekproefvariantie σ^2. Men kan bewijzen dat

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

een χ^2-verdeling heeft met $n-1$ vrijheidsgraden (vergelijk §15.1 en §14.4). $\mathrm{X}^2 = (n-1)S^2/\sigma_0^2$ wordt gebruikt als **toetsingsgrootheid** voor het toetsen van de nulhypothese $H_0 : \sigma^2 = \sigma_0^2$; onder H_0 geldt $\mathrm{X}^2 \sim \chi^2_{n-1}$. Indien we in een concrete situatie de **waarde** van X^2 noteren met χ^2, dan beschrijft de Tabel 16.5 de desbetreffende toetsingsprocedures (zie Appendix C2 voor de kritieke grenzen).

Tabel 16.5:
Het toetsen van de nulhypothese H_0: $\sigma^2 = \sigma_0^2$ indien μ onbekend.

H_0	H_1	**Overschrijdingskans**	Verwerp H_0 als
$\sigma^2 \leq \sigma_0^2$, of: $\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$P(\mathrm{X}^2 > \chi^2)$	$\chi^2 > \chi^2_{n-1;\alpha}$
$\sigma^2 \geq \sigma_0^2$, of: $\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$P(\mathrm{X}^2 < \chi^2)$	$\chi^2 < \chi^2_{n-1;1-\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$2\min\left[P(\mathrm{X}^2 < \chi^2), P(\mathrm{X}^2 > \chi^2)\right]$	$\chi^2 > \chi^2_{n-1;\alpha/2}$ of $\chi^2 < \chi^2_{n-1;1-\alpha/2}$

Voor een **betrouwbaarheidsinterval** voor σ^2 wordt verwezen naar §14.4.

Voorbeeld 16.3: Een kunstmestproducent wil nagaan of de kwaliteit van de produceerde kunstmest voldoet aan een bepaalde Europese norm met m.b.t. de variantie van de hoeveelheid verontreinigingen in de kunstmest. Men analyseert $n=18$ willekeurig gekozen zakken en bepaalt per zak de hoeveelheid verontreinigingen. De uitkomst is een steekproefvariantie $s^2=6{,}4.$ De norm is $\sigma^2 \le 5$. De **onderzoekshypothese** luidt derhalve: $H_1:\sigma^2>5$. De waarde van de **toetsingsgrootheid** is $\chi^2=17\times 6{,}4/5\approx 21{,}76$. De **overschrijdingskans** is $P\left(\mathrm{X}^2>21{,}76\right)\approx 0{,}19$. Stel dat we een **onbetrouwbaarheidsdrempel** van $\alpha=0{,}1$ hanteren. De **kritieke grens** is dan $\chi^2_{17;0{,}1}\approx 24{,}8$. We verwerpen we de nulhypothese dus niet (immers $0{,}19>\alpha$ en $21{,}76<24{,}8$): op grond van dit onderzoek is er geen reden om aan te nemen dat de geproduceerde kunstmest niet aan de bedoelde norm voldoet.

16.4 2-steekproeven toets, bekende varianties σ_x^2 en σ_y^2; H_0: $\mu_x - \mu_y = d_0$

Als we de beschikking hebben over twee van elkaar onafhankelijke aselecte steekproeven $X_1,...,X_{n_x}$ en $Y_1,...,Y_{n_y}$ uit de normaalverdeelde populaties $N(\mu_x;\sigma_x^2)$ en $N(\mu_y;\sigma_y^2)$, respectievelijk, dan geldt

$$\overline{X}-\overline{Y}\sim N(\mu_x-\mu_y;\sigma_x^2/n_x+\sigma_y^2/n_y)$$

Stel nu dat verwachtingswaarden μ_x en μ_y *onbekend* zijn, en varianties σ_x^2 en σ_y^2 *bekend.* Voor het toetsen van $H_0:\mu_x-\mu_y=d_0$ (vaak zal $d_0=0$) gebruiken we de **toetsingsgrootheid**

$$Z=\frac{\overline{X}-\overline{Y}-d_0}{\sqrt{\sigma_x^2/n_x+\sigma_y^2/n_y}}$$

die onder H_0 een standaardnormale verdeling heeft. In een concrete situatie neemt de toetsingsgrootheid de **waarde**

$$z=\frac{\overline{x}-\overline{y}-d_0}{\sqrt{\sigma_x^2/n_x+\sigma_y^2/n_y}}$$

aan. Tabel 16.6 beschrijft de toetsingsprocedures voor de drie gebruikelijke gevallen.

Tabel 16.6:
Het toetsen van de nulhypothese H_0: $\mu_x - \mu_y = d_0$ bij bekende varianties en/of grote steekproefomvangen.

H_0	H_1	**Overschrijdingskans**	Verwerp H_0 als
$\mu_x - \mu_y \le d_0$, of: $\mu_x - \mu_y = d_0$	$\mu_x - \mu_y > d_0$	$P(Z > z) = 1 - \Phi(z)$	$z > z_\alpha$
$\mu_x - \mu_y \ge d_0$, of: $\mu_x - \mu_y = d_0$	$\mu_x - \mu_y < d_0$	$P(Z < z \mid d = d_0) = \Phi(z)$	$z < -z_\alpha$
$\mu_x - \mu_y = d_0$	$\mu_x - \mu_y \ne d_0$	$P(\lvert Z\rvert > \lvert z\rvert) = 2(1 - \Phi(\lvert z\rvert))$	$\lvert z\rvert > z_{\alpha/2}$

Merk op dat dezelfde toetsprocedure mag worden toegepast indien de steekproeven $X_1,...,X_{n_x}$ en $Y_1,...,Y_{n_y}$ *niet* uit normaalverdeelde populaties komen, mits de steekproefomvangen n_x en n_y groot genoeg zijn. In dat geval mogen zelfs de varianties σ_x^2 en σ_y^2 onbekend zijn en vervangen worden door de **steekproefvarianties** s_x^2 en s_y^2. Hoe minder de kansverdelingen van $\overline{X}$ en $\overline{Y}$ op normaalverdelingen lijken, des te groter moeten n_x en n_y zijn voor een betrouwbare toets. Een $100(1-\alpha)\%$ **-betrouwbaarheidsinterval** voor $\mu_x - \mu_y$ kan worden verkregen met de methode uit §14.3 door gebruik te maken van de kansverdeling van *Z*:

$$(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y};\ \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y})$$

16.5 2-steekproeven toets, onbekende gelijke varianties $\sigma^2 = \sigma x^2 = \sigma y^2$; H_0: $\mu_x - \mu_y = d_0$

Indien de steekproeven $X_1,...,X_{n_x}$ en $Y_1,...,Y_{n_y}$ uit normaalverdeelde populaties komen, σ_x^2 en σ_y^2 onbekend zijn, maar de steekproefomvangen te klein zijn om de toetsprocedures van §16.4 te mogen toepassen, dan kunnen we onder de *veronderstelling* dat de onbekende varianties *gelijk* zijn (§16.8 biedt een toets hiervoor), de volgende procedure toepassen voor het toetsen van $H_0 : \mu_x - \mu_x = d_0$. De onbekende variantie $\sigma^2 = \sigma_x^2 = \sigma_y^2$ wordt geschat door de afzonderlijke steekproefvarianties S_x^2 en S_y^2 te combineren tot de **'gepoolde' steekproefvariantie**

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$

Voor het toetsen van $H_0 : \mu_x - \mu_y = d_0$ gebruiken we de **toetsingsgrootheid**

$$T = \frac{\overline{X} - \overline{Y} - d_0}{S_p\sqrt{1/n_x + 1/n_y}}$$

die, als H_0 waar is, een Student's *t*-verdeling met $n_x + n_y - 2$ **vrijheidsgraden** heeft. Stel dat in een concrete situatie de toetsingsgrootheid de **waarde**

$$t = \frac{\bar{x} - \bar{y} - d_0}{s_p \sqrt{1/n_x + 1/n_y}}$$

aanneemt; Tabel 16.7 beschrijft de gebruikelijke toetsingsprocedures.

Tabel 16.7:
Het toetsen van de nulhypothese H_0: $\mu_x - \mu_y = d_0$ bij gelijke, onbekende, varianties.

H_0	H_1	**Overschrijdingskans**	Verwerp H_0 als
$\mu_x - \mu_y \leq d_0$, of: $\mu_x - \mu_y = d_0$	$\mu_x - \mu_y > d_0$	$P(T > t)$	$t > t_{n_x+n_y-2;\alpha}$
$\mu_x - \mu_y \geq d_0$, of: $\mu_x - \mu_y = d_0$	$\mu_x - \mu_y < d_0$	$P(T < t)$	$t < -t_{n_x+n_y-2;\alpha}$
$\mu_x - \mu_y = d_0$	$\mu_x - \mu_y \neq d_0$	$P(\lvert T \rvert > \lvert t \rvert)$	$\lvert t \rvert > t_{n_x+n_y-2;\alpha/2}$

Een $100(1-\alpha)\%$ **-betrouwbaarheidsinterval** voor $\mu_x - \mu_y$ kan worden verkregen met de methode uit §14.3 door gebruik te maken van de kansverdeling van T:

$$(\overline{X} - \overline{Y} - t_{n-1;\alpha/2} S_p \sqrt{1/n_x + 1/n_y}\,;\, \overline{X} - \overline{Y} + t_{n-1;\alpha/2} S_p \sqrt{1/n_x + 1/n_y})$$

16.6 2-steekproeven toets, onbekende varianties σ_x^2 en σ_y^2; H_0: $\mu_x - \mu_y = d_0$

Indien de steekproeven $X_1,...,X_{n_x}$ en $Y_1,...,Y_{n_y}$ uit normaalverdeelde populaties komen, σ_x^2 en σ_y^2 onbekend zijn, de steekproefomvangen te klein zijn om de toetsprocedures van §16.4 te mogen toepassen, *en* we er van uit moeten gaan dat de onbekende varianties *verschillend* zijn, dienen de in deze paragraaf beschreven toetsingsprocedures te worden toegepast. De toetsingsgrootheid is vergelijkbaar met die van §16.4 waarin de varianties zijn vervangen door steekproefvarianties. Onder H_0 heeft de **toetsingsgrootheid**

$$T = \frac{\overline{X} - \overline{Y} - d_0}{\sqrt{S_x^2/n_x + S_y^2/n_y}}$$

nu echter een **Student *t*-verdeling** met (in benadering)

$$\nu = \frac{\left(s_x^2/n_x + s_y^2/n_y\right)^2}{\left(s_x^2/n_x\right)^2/(n_x-1) + \left(s_y^2/n_y\right)^2/(n_y-1)}$$

vrijheidsgraden (i.h.a. dient ν eerst te worden afgerond tot een geheel getal); dus $T \sim t_\nu$. Een $100(1-\alpha)\%$ **-betrouwbaarheidsinterval** voor $\mu_x - \mu_y$ kan worden verkregen met de methode uit §14.3 door gebruik te maken van de kansverdeling van T:

$$\left(\overline{X} - \overline{Y} - t_{\nu;\alpha/2}\sqrt{S_x^2/n_x + S_y^2/n_y}\,;\, \overline{X} - \overline{Y} + t_{\nu;\alpha/2}\sqrt{S_x^2/n_x + S_y^2/n_y}\right)$$

16.7 Gepaarde steekproeven toets; H_0: $\mu_x - \mu_y = d_0$

In veel onderzoekssituaties is het mogelijk om aan dezelfde (of vergelijkbare) populatie elementen twee observaties te doen, bijvoorbeeld het meten van de opbrengst van *n* **paren** percelen waarbij de percelen in een paar onderling qua grondsamenstelling e.d. goed te vergelijken zijn, maar met verschillende soorten kunstmest worden behandeld. Door de opbrengsten paarsgewijs van elkaar af te trekken, krijgen we een beeld van de kwaliteiten van de soorten kunstmest ten opzichte van elkaar zonder last te hebben van het feit dat verschillende paren percelen mogelijk een verschillende grondsamenstelling hebben waardoor de opbrengsten mede beïnvloed worden. We beschikken in deze situatie over de **gepaarde steekproeven** $(X_1,Y_1),...,(X_n,Y_n)$. Onder de veronderstelling dat de **verschillen**, $D_i = (X_i - Y_i), i = 1,...,n$, uit een **normaalverdeelde populatie** komen met onbekende verwachtingswaarde μ_d en variantie σ_d^2, kunnen we toetsingsprocedures opstellen voor $H_0 : \mu_x - \mu_y = d_0$, of, equivalent, voor $H_0 : \mu_d = d_0$. De **toetsingsgrootheid**

$$T = \frac{\overline{D} - d_0}{S_d / \sqrt{n}}$$

heeft een **Student's *t*-verdeling** met $n-1$ vrijheidsgraden; $\overline{D} = \overline{X} - \overline{Y}$ is het steekproefgemiddelde en S_d de steekproefstandaardafwijking van de gepaarde verschillen.

Als we in een concrete situatie de **waarde** van de toetsingsgrootheid noteren met

$$t = \frac{\overline{d} - d_0}{s_d / \sqrt{n}}$$

dan zijn de toetsingsprocedures hetzelfde als die in §16.2. Een $100(1-\alpha)\%$ - **betrouwbaarheidsinterval** voor $\mu_x - \mu_y$ kan worden verkregen met de methode uit §14.3 door gebruik te maken van de kansverdeling van *T*:

$$(\overline{X} - \overline{Y} - t_{n-1;\alpha/2} S_d \sqrt{n};\ \overline{X} - \overline{Y} + t_{n-1;\alpha/2} S_d \sqrt{n})$$

16.8 2 steekproeven, onbekende verwachtingswaarden μ_x en μ_y, H_0: $\sigma_x^2 = \sigma_y^2$

Om de toetsen van §16.5 te mogen toepassen moeten we veronderstellen dat $\sigma_x^2 = \sigma_y^2$. In de praktijk wordt dan eerst statistisch getoetst of deze veronderstelling redelijk is. Verwijzend naar §15.3 gebruiken we daarvoor de wetenschap dat

$$F = \frac{S_x^2 / \sigma_x^2}{S_y^2 / \sigma_y^2} \sim F_n^m$$

als S_x^2 en S_y^2 de steekproefvarianties zijn gebaseerd op twee onafhankelijke aselecte steekproeven met omvangen $m+1$ en $n+1$, respectievelijk. Als de nulhypothese $H_0 : \sigma_x^2 = \sigma_y^2$ waar is, heeft de **toetsingsgrootheid**

$$F = \frac{S_x^2}{S_y^2}$$

een ***F*-verdeling** met *m* vrijheidsgraden in de teller en *n* in de noemer. Tabel 16.8 beschrijft de toetsprocedures.

H_0	H_1	Verwerp H_0 als
$\sigma_x^2 \le \sigma_y^2$, of: $\sigma_x^2 = \sigma_y^2$	$\sigma_x^2 > \sigma_y^2$	$F > F_{n;\alpha}^m$
$\sigma_x^2 = \sigma_y^2$	$\sigma_x^2 \ne \sigma_y^2$	$F > F_{n;\alpha/2}^m$ of $F < F_{n;1-\alpha/2}^m$

Merk op dat $H_1 : \sigma_x^2 < \sigma_y^2$ identiek is aan $H_1 : \sigma_y^2 > \sigma_x^2$ en de bijbehorende toetsprocedure dus gevonden wordt door *x* en *y* (en *m* en *n*) te verwisselen en de tweede regel uit de tabel toe te passen; de toetsingsgrootheid is dan $F = \frac{S_y^2}{S_x^2}$.

17. Variantie analyse

17.1 Inleiding

De in §16.5 beschreven 2-steekproeven toets voor de gelijkheid van twee populatiegemiddelden (dus met $d_0 = 0$) kan worden gegeneraliseerd naar een toets voor de gelijkheid van $k > 2$ gemiddelden. We nemen aan dat we de beschikking hebben over *k* aselecte steekproeven, en dat steekproef *i* ($i = 1,...,k$) uit de **normale verdeling** $N(\mu_i;\sigma^2)$ afkomstig is. Net zoals in §16.5 nemen we dus aan dat de populatievarianties gelijk zijn. De volgende paragraaf beschrijft de details van deze toets, meestal aangeduid met de term (één-weg-) **variantie analyse**.

17.2 k-steekproeven toets, onbekende gelijke varianties; H_0: $\mu_1 =...= \mu_k$

Beschouw de stochastische variabelen $X_i \sim N(\mu_i;\sigma^2), i = 1,...,k,$ waarvan we aselecte steekproeven beschouwen met respectievelijk $n_i, i = 1,...,k,$ elementen. Net zoals in §16.5 hebben we voor de toetsingsgrootheid, behalve de steekproefomvangen, alleen nodig de steekproefgemiddelden en steekproefvarianties. Deze noteren we hier met respectievelijk $M_1,...,M_k$ en $S_1^2,...,S_k^2$. De som van alle steekproefomvangen noteren we met $n = \sum_{i=1}^{k} n_i$; het gemiddelde van alle steekproefgegevens bij elkaar noteren we met $\overline{M} = \sum_{i=1}^{k} n_i M_i / n$. De **nulhypothese**

$$H_0 : \mu_1 = ... = \mu_k$$

wordt getoetst door twee verschillende schatters (onder H_0) van de gemeenschappelijke populatievariantie σ^2 op elkaar te delen. Beschouw daartoe de volgende **kwadratensommen**: *KSt*, een schatter van de variabiliteit *tussen* de populaties (aannemend dat de gemiddelden gelijk zijn, dus H_0 waar is), en *KSb*, een schatter van de variabiliteit *binnen* de populaties.

$KSt = \sum_{i=1}^{k} n_i (M_i - \overline{M})^2$ ('*k*wadraten*s*om *t*ussen groepen', met $k-1$ vrijheidsgraden)

$KSb = \sum_{i=1}^{k} (n_i - 1) S_i^2$ ('*k*wadraten*s*om *b*innen groepen', met $n-k$ vrijheidsgraden)

De **toetsingsgrootheid** is nu

$$F = \frac{KSt/(k-1)}{KSb/(n-k)}$$

waarbij (onder H_0) zowel de teller als de noemer zuivere schatters zijn van σ^2. Als H_0 waar is, heeft *F* een *F*-verdeling met $k-1$ vrijheidsgraden van de teller en $n-k$ vrijheidsgraden van de noemer. In

een concrete situatie **verwerpen** we (met een onbetrouwbaarheidsdrempel α) de nulhypothese (ten gunste van de **alternatieve hypothese** dat minstens 2 populatiegemiddelden verschillend zijn) als de **waarde**, f, van de toetsingsgrootheid, F, voldoet aan

$$f > F^{k-1}_{n-k;\alpha}$$

Hierbij is $F^{k-1}_{n-k;\alpha}$ bepaald door $P\left(F > F^{k-1}_{n-k;\alpha}\right) = \alpha$ (zie Appendix C4).

17.3 Voorbeeld *k*-steekproeven toets

De ANWB wil het brandstofgebruik van ($k =$) 3 nieuwe concurrerende benzineauto's testen. Zij heeft daartoe de beschikking over 13 auto's van type A, 14 van type B, en 14 van type C. Om te voorkomen dat het rijgedrag van de chauffeurs de validiteit van de uitkomsten beïnvloedt, worden elk van 13+14+14=41 beschikbare chauffeurs willekeurig toegewezen aan een van de $n = 41$ auto's. Vervolgens rijden alle chauffeurs eenzelfde route, probeert men zo veel mogelijk storende factoren (zoals filevorming) uit te sluiten, en wordt na afloop vastgesteld hoeveel km per liter benzine is gereden. De resultaten zijn samengevat in onderstaande tabel.

															n_i	m_i	s_i^2
A	21,2	22,5	18,6	21,5	14,0	20,9	22,5	18,1	18,7	22,4	24,2	18,1	21,6		13	20,33	7,37
B	20,2	24,5	24,4	23,3	27,1	16,6	22,1	22,6	17,1	23,5	20,1	21,9	20,8	16,6	14	21,86	9,94
C	24,9	19,3	18,2	17,9	17,7	20,0	19,1	17,0	18,3	22,9	17,4	11,9	21,0	15,4	14	18,89	9,79

De laatste twee kolommen bevatten de gerealiseerde gemiddelde aantallen km's en de bijbehorende steekproefvarianties. Het gemiddelde van de gemiddelden per auto is:
$\overline{m} = (13 \times 20{,}33 + 14 \times 21{,}86 + 14 \times 18{,}89) / 41 \approx 20{,}36$. Vervolgens bepalen we de waarden van *KSt* en *KSb*, delen beide door het bijbehorende aantal vrijheidsgraden en bepalen tenslotte de waarde van de toetsingsgrootheid

$$f = \frac{\left[13 \times (20{,}33 - 20{,}36)^2 + 14 \times (21{,}86 - 20{,}36)^2 + 14 \times (18{,}89 - 20{,}36)^2\right]/(3-1)}{\left[12 \times 7{,}37 + 13 \times 9{,}94 + 13 \times 9{,}79\right]/(41-3)} \approx 3{,}40$$

Met bijvoorbeeld EXCEL is vast te stellen dat de overschrijdingskans $P(F_{39}^2 > 3{,}40)$ gelijk is aan 0,0438. Indien we $H_0 : \mu_A = \mu_B = \mu_C$ willen toetsen met een onbetrouwbaarheidsdrempel van $\alpha = 0{,}05$, dan kunnen we deze nulhypothese dus verwerpen. Met andere woorden: het rij-onderzoek toont aan dat de drie geteste autotypen gemiddeld genomen niet alle drie evenveel km's op een liter rijden; deze uitspraak is onderhevig aan een onbetrouwbaarheid van 5%.

18. χ^2 - 'Goodness-of-fit' toets

18.1 Toetsingsgrootheid en aantal vrijheidsgraden

De χ^2 - 'goodness-of-fit' toets wordt gebruikt om te toetsen of de waarden uit een aselecte steekproef afkomstig zouden kunnen zijn van een volledig gespecificeerde (theoretische) verdeling, of van een verdeling waarvan het type wel bekend is (binomiaal, normaal, Poisson, enzovoorts), maar de parameter(s) onbekend. De *n* steekproefwaarnemingen worden verondersteld te zijn ingedeeld in *k* klassen, waarbij het aantal waarnemingen in klasse *i* genoteerd wordt met O_i ('*O*bserved') en het – op basis van de veronderstelde theoretische verdeling – *verwachte* aantal waarnemingen in klasse *i* met E_i ('*E*xpected'). Uit het voorgaande volgt $n = \sum_{i=1}^{k} O_i$. De formule van de **toetsingsgrootheid** is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Deze toetsingsgrootheid heeft onder de aanname (**nulhypothese**) dat de waarnemingen inderdaad uit de veronderstelde verdeling komen een **χ^2-verdeling** met een aantal **vrijheidsgraden** (*df*) dat afhankelijk is van de situatie of de verdeling wel of niet volledig gespecificeerd is. Indien de verdeling *volledig gespecificeerd* is, is het aantal vrijheidsgraden gelijk aan $df = k - 1$; indien echter eerst *m* onbekende parameters moeten worden geschat uit de waarnemingen is het aantal vrijheidsgraden gelijk aan $df = k - m - 1$ (meestal $m = 1$ of $m = 2$). In het eerste geval dient *k* dus minimaal 2 te zijn, en in het tweede geval minimaal $m + 2$.

'Grote' waarden van de toetsingsgrootheid suggereren dat de verschillen tussen geobserveerde en verwachte aantallen waarnemingen in één of meer klassen zo groot zijn dat moet worden getwijfeld aan de (nul)hypothese dat de waarnemingen afkomstig zijn uit de veronderstelde verdeling. Het **kritieke gebied** wordt dus gevormd door alle waarden van de toetsingsgrootheid die groter zijn dan de **kritieke grens** $\chi^2_{df;\alpha}$, indien de onbetrouwbaarheidsdrempel α is.

Aangezien deze toetsingsgrootheid gebaseerd is op normaalbenaderingen (vergelijk §15.1), is het van belang dat alle E_i *minimaal 5* zijn (vergelijk §7.3).

18.2 Voorbeelden met volledig gespecificeerde theoretische verdelingen

Voorbeeld 18.1 Beschouw alle gezinnen met 5 kinderen in een bepaalde gemeenschap. Met betrekking tot de samenstelling zijn er 6 mogelijkheden: {5*j*; 4*j*+1*m*; 3*j*+2*m*; 2*j*+3*m*; 1*j*+4*m*; 5*m*}, waarbij *j* staat voor jongen en *m* voor meisje; noteer de 6 mogelijkheden respectievelijk met 0 t/m 5. Het onderzoek omvat 1022 gezinnen met 5 kinderen. De verdeling over de 6 mogelijkheden blijkt alsvolgt:

i	0	1	2	3	4	5	Totaal
O_i	58	149	305	303	162	45	1022

Men is geïnteresseerd in de vraag of de waargenomen aantallen in de 6 klassen afwijken van wat men zou verwachten op basis van de veronderstelling dat ieder nieuw kind (onafhankelijk van het geslacht van de overige kinderen in het gezin) met een kans van 0,5 een jongen is (en met kans 0,5 een meisje). Indien deze veronderstelling waar is, wordt de kans, π_i, op mogelijkheid i ($i = 0,...,5$) bepaald door een **binomiaalverdeling** met parameters $n = 5$ en $p = 0{,}5$. Dus $\pi_i = \binom{5}{i} 0{,}5^i 0{,}5^{5-i} = \frac{5!}{i!(5-i)!} 0{,}5^5$.

Hiermee kunnen we ook de verwachte aantallen per klasse bepalen: $E_i = 1022 \times \pi_i$, en vervolgens de termen $(O_i - E_i)^2 / E_i$; zie onderstaande tabel.

i	0	1	2	3	4	5	Totaal
π_i	0,03125	0,15625	0,31250	0,31250	0,15625	0,03125	1
E_i	31,9375	159,6875	319,3750	319,3750	159,6875	31,9375	1022
$(O_i - E_i)^2 / E_i$	21,2682	0,7153	0,6470	0,8396	0,0335	5,3426	**28,85**
r_i	4,69	-0,92	-0,97	-1,11	0,20	2,35	

De waarde van de toetsingsgrootheid is dus 28,85. Het aantal vrijheidsgraden is 5 ($=6-1$). Stel dat we bij het toetsen van de nulhypothese een onbetrouwbaarheidsdrempel van $\alpha = 0.01$ accepteren. Dan is de kritieke grens $\chi^2_{5;0,01} = 15{,}09$. Aangezien de waarde van de toetsingsgrootheid dus in het kritieke gebied ligt ($28{,}85 > 15{,}09$), verwerpen we de nulhypothese dat in de onderzochte gemeenschap ieder kind in een gezin met 5 kinderen met kans 0,5 een jongen (en met kans 0,5 een meisje) is, onafhankelijk van het geslacht van de overige kinderen in het gezin. Aangezien we H_0 verwerpen, is het interessant om te weten welke klassen de grootste bijdrage leveren aan de waarde van de toetsingsgrootheid. Dat kunnen we analyseren met behulp van de zogenaamde **gestandaardiseerde residuen**

$$r_i = \frac{O_i - n\pi_i}{\sqrt{n\pi_i(1-\pi_i)}} = \frac{O_i - E_i}{\sqrt{E_i(1-\pi_i)}}$$

De onderste regel van bovenstaande tabel laat de corresponderende waarden van r_i zien. Deze waarden noemen we extreem als ze bijvoorbeeld groter zijn dan 2 of kleiner dan -2 (vergelijk de standaardnormale verdeling). Het blijkt dan dat met name het aantal gezinnen waarin alle kinderen hetzelfde geslacht hebben groter mag worden genoemd dan verwacht.

Voorbeeld 18.2 De eerste regel in onderstaande tabel bevat de fractie stemmen op diverse partijen bij de Tweede-Kamerverkiezing in 2003; de tweede regel bevat de resultaten van een landelijke peiling van de politieke voorkeur in week 20 van 2004 in een steekproef van 1000 stemgerechtigden.

	CDA	PvdA	VVD	LPF	Groen Links	SP	D66	Overig	Totaal
TK2003	0,293	0,280	0,187	0,053	0,053	0,060	0,040	0,034	1
Peiling 2004	220	313	207	27	60	87	33	53	1000

We willen onderzoeken of deze peiling ertoe aanleiding geeft te veronderstellen dat de politieke voorkeur is veranderd na de Tweede-Kamerverkiezing van 2003. Daartoe toetsen we de nulhypothese dat de resultaten van de peiling bepaald worden door een **multinomiale verdeling** met kansen zoals in de eerste regel van de tabel. De waarde van de toetsingsgrootheid is

$$\chi^2 = \frac{(220-293)^2}{293} + \ldots + \frac{(53-34)^2}{34} \approx 61{,}9$$

De corresponderende overschrijdingskans $P(\chi^2_7 > 61{,}9) << 0{,}001$; dus voor iedere gebruikelijke onbetrouwbaarheidsdrempel kunnen we de nulhypothese verwerpen.

18.3 Voorbeeld met een theoretische verdeling met onbekende parameters

Voorbeeld 18.3 Een statistisch analist van een grote supermarktketen wil met variantie analyse (vgl. Hoofdstuk 17) onderzoeken of de gemiddelde bedragen die klanten spenderen in een aantal vergelijkbare winkels van elkaar verschillen. Daartoe wil zij op basis van een aselecte steekproef van 100 klanten van een winkel onderzoeken of mag worden verondersteld dat de door hen besteedde

bedragen uit een normale verdeling komen (hetgeen een vereiste is voor de daarna uit te voeren variantie analyse). De 100 (geordende) waarnemingen staan in de volgende tabel.

1,04	20,12	33,18	41,19	47,31	53,04	60,45	68,07	75,76	87,10
3,78	21,46	33,21	41,33	47,64	53,87	60,86	68,32	76,56	87,62
9,77	22,19	33,96	42,07	47,85	54,05	61,19	69,38	77,25	90,16
12,08	25,36	34,39	42,19	48,64	54,07	65,28	69,41	77,85	92,41
12,50	25,46	35,57	42,96	48,91	54,38	65,95	69,76	79,14	92,94
14,11	26,21	35,72	43,98	49,38	54,65	66,20	71,48	79,57	96,34
16,21	27,45	36,72	44,32	50,74	56,12	66,51	71,54	82,48	97,91
16,88	28,21	36,76	45,66	51,52	57,25	67,39	72,06	83,05	105,17
17,73	29,98	38,19	45,75	52,15	58,61	67,51	74,10	86,24	105,32
18,34	30,42	39,02	45,81	53,04	60,00	68,01	74,17	86,27	112,94

Steekproefgemiddelde en –standaardafwijking van de 100 waarnemingen zijn respectievelijk $\bar{x} = €53{,}54$ en $s = €24{,}80$. Aangezien de veronderstelde verdeling een continue verdeling is, is het nodig om eerst een klassenindeling te maken. De analist kiest voor 8 klassen en maakt daarvoor gebruik van de standaardnormale verdeling:

i				klassegrenzen	O_i	E_i	$(O_i - E_i)^2 / E_i$
1	$P(Z < -1{,}15)$	=	0,125	[0;25,02]	13	12,5	0,020
2	$P(-1{,}15 < Z < -0{,}67)$	=	0,126	[25,02;36,92]	15	12,6	0,442
3	$P(-0{,}67 < Z < -0{,}32)$	=	0,123	[36,92;45,61]	9	12,3	0,888
4	$P(-0{,}32 < Z < 0)$	=	0,126	[45,61;53,54]	14	12,6	0,167
5	$P(0 < Z < 0{,}32)$	=	0,126	[53,54;61,48]	12	12,6	0,024
6	$P(0{,}32 < Z < 0{,}67)$	=	0,123	[61,48;70,16]	12	12,3	0,008
7	$P(0{,}67 < Z < 1{,}15)$	=	0,126	[70,16;82,07]	11	12,6	0,212
8	$P(Z > 1{,}15)$	=	0,125	[82,07;∞]	14	12,5	0,178
Totaal			1,000		100	100	1,939

De klassengrenzen van de standaardnormale verdeling zijn zo gekozen dat de resulterende kansen ongeveer gelijk zijn. Deze grenzen worden getransformeerd naar grenzen die corresponderen met de geschatte normale verdeling $N(53{,}54;24{,}80^2)$; bijvoorbeeld $25{,}02 = \bar{x} - 1{,}15 \times s$. Vervolgens wordt geteld hoeveel van de 100 bestedingen in elk van de klassen vallen; de resultaten staan vermeld in de kolom O_i. De verwachtte aantallen worden gevonden door de kansen per klasse te vermenigvuldigen met 100. Ten slotte volgt dan de waarde van de toetsingsgrootheid $\chi^2 = 1{,}939$. Het bijbehorende aantal vrijheidsgraden is gelijk aan het aantal klassen, verminderd met 1, verminderd met het aantal geschatte parameters, dus: $df = 8 - 1 - 2 = 5$. Voor de overschrijdingskans geldt $P(\chi_5^2 > 1{,}939) \approx 0{,}86$. Er is derhalve geen enkele reden om de nulhypothese dat de waarnemingen uit een normale verdeling komen te verwerpen.

19. χ^2 - toets voor onafhankelijkheid

19.1 Contingentietabellen

In deze paragraaf bespreken we een toets waarmee enerzijds kan worden vastgesteld (behoudens een onbetrouwbaarheidsdrempel) of *twee* nominale (of ordinale) variabelen (zoals geslacht, inkomenscategorie, religie, enz.) van elkaar onafhankelijk zijn, of, anderzijds, kan worden vastgesteld of *twee of meer populaties* verschillend zijn met betrekking tot *een* bepaalde nominale (of ordinale) variabele. Daartoe is verondersteld dat N paren waarnemingen zijn geordend volgens een zogenaamde **kruistabel** of **contingentietabel** met m (het aantal categorieën van de eerste variabele) rijen en n (het aantal categorieën van de tweede variabele) kolommen. Schematisch ziet zo'n kruistabel er als volgt uit:

O_{11}	O_{12}	..	O_{1n}	r_1
O_{21}			..	
..			..	
O_{m1}	..	..	O_{mn}	r_m
c_1	..	..	c_n	N

Hierin is O_{ij} (in 'cel' (i,j)) het aantal (paren) waarnemingen waarbij de eerste variabele gelijk is aan categorie *i* en de tweede variabele gelijk aan categorie *j* (ofwel het aantal keren dat categorie *j* voorkomt in de steekproef uit populatie *i*). Verder geldt $r_i = \sum_{j=1}^{n} O_{ij}$, $c_j = \sum_{i=1}^{m} O_{ij}$, en $N = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$. Indien de twee variabelen van elkaar **onafhankelijk** zijn *verwachten* we een *geschat* aantal $E_{ij} = N \times \frac{r_i}{N} \times \frac{c_j}{N} = \frac{r_i c_j}{N}$ in cel (i,j). Dat leidt tot de toetsingsgrootheid

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - r_i c_j / N)^2}{r_i c_j / N}$$

Net als in Hoofdstuk 18 geldt ook hier de voorwaarde $E_{ij} \geq 5$ opdat de toetsingsgrootheid, onder de nulhypothese van onafhankelijkheid, bij benadering een **chi-kwadraat verdeling** heeft. Het aantal **vrijheidsgraden** is gegeven door

$$df = (m-1) \times (n-1)$$

19.2 Voorbeeld contingentietabellen

Voorbeeld 19.1 Onderstaande 2×3 kruistabel geeft de resultaten weer van een opinie onderzoek waarin 919 mannen en 947 vrouwen naar hun mening werden gevraagd over een actueel onderwerp.

	voor	tegen	geen mening	
man	570	230	119	919
vrouw	545	221	181	947
	1115	451	300	1866

De vraag of mannen en vrouwen een verschillende mening hebben ten aanzien van dit actuele onderwerp kan worden beantwoord door de bovenbeschreven χ^2-toetsingsgrootheid te bepalen:

$$\chi^2 = \frac{(570 - 919 \times 1115/1866)^2}{919 \times 1115/1866} + \ldots + \frac{(181 - 947 \times 300/1866)^2}{947 \times 300/1866} \approx 13{,}14$$

Stel dat we een onbetrouwbaarheidsdrempel van $\alpha = 0{,}01$ accepteren. Aangezien de overschrijdingskans gelijk is aan $P(\chi_2^2 > 13{,}14) \approx 0{,}0014 < \alpha$, kunnen we de nulhypothese dat het standpunt ten aanzien van het bewuste onderwerp onafhankelijk is van geslacht, verwerpen.

19.3 2x2-contingentietabellen

Voor een kruistabel met 2 rijen en 2 kolommen kan de in §19.1 beschreven toetsingsgrootheid ook geschreven worden als

$$\chi^2 = \left(\frac{O_{11}O_{22} - O_{21}O_{12}}{N} \right)^2 \left(\frac{N}{r_1 c_1} + \frac{N}{r_1 c_2} + \frac{N}{r_2 c_1} + \frac{N}{r_2 c_2} \right)$$

Aangezien het aantal bijbehorende vrijheidsgraden gelijk is aan 1, kan ook getoetst worden met

$$z = \sqrt{\chi^2} = \left(\frac{O_{11}O_{22} - O_{21}O_{12}}{N} \right) \sqrt{\frac{N}{r_1 c_1} + \frac{N}{r_1 c_2} + \frac{N}{r_2 c_1} + \frac{N}{r_2 c_2}}$$

De nulhypothese dat beide variabelen onafhankelijk zijn van elkaar, wordt dan – met een onbetrouwbaarheidsdrempel van α – verworpen indien

$$P(Z > z) < z_{\alpha/2}$$

(vgl. §14.2).

Een interessante alternatieve procedure voor deze situatie is de exacte toets van Fisher voor 2×2-tabellen, die in de volgende paragraaf wordt behandeld.

19.4 Fisher's exacte toets voor 2x2-contingentietabellen

Beschouw twee (oneindige) populaties waarvan de eenheden behoren tot een van de twee (elkaar uitsluitende) categorieën *S* en *F* ('Succes', 'Failure'). Uit elke populatie wordt een steekproef getrokken en een 2×2-kruistabel legt de frequenties vast van de aantallen '*S*' en '*F*':

	Categorie		
	S	*F*	Totaal
Steekproef 1	*a*	*b*	r_1
Steekproef 2	*c*	*d*	r_2
Totaal	c_1	c_2	N

We veronderstellen dat de tabel zodanig is gerangschikt dat $r_1 \geq r_2$ en $a/r_1 \geq c/r_2$. Merk op dat, bij gegeven marginale frequenties en gegeven *a*, de frequenties in de overige drie cellen vastliggen. We willen de (nul)hypothese toetsen dat de proportie '*S*' in beide populaties hetzelfde is. De exacte toets van Fisher voor 2×2-kruistabellen die voor dit onderzoek gebruikt kan worden, berust op het feit dat de kans op de door de tabel weergegeven realisatie onder *onafhankelijkheid* van beide variabelen (dus onder H_0), bij aselecte en onafhankelijke trekkingen van de steekproefwaarnemingen, en bij gegeven marginale frequenties r_1, r_2, c_1, c_2, bepaald is door de hypergeometrische kans (vgl. Hoofdstuk 10):

$$P(a) = \binom{r_1}{a}\binom{r_2}{c} \Big/ \binom{N}{c_1} = \frac{r_1!r_2!c_1!c_2!}{N!a!b!c!d!}$$

Indien de *(eenzijdige) alternatieve hypothese* luidt dat de proportie '*S*' in de eerste populatie groter is dan die in de tweede populatie, dan is de overschrijdingskans

$$\sum_{i=0}^{\min(b,c)} P(a,b,c,d,i) = \sum_{i=0}^{\min(b,c)} \frac{r_1!r_2!c_1!c_2!}{N!(a+i)!(b-i)!(c-i)!(d+i)!}$$

Er bestaan uitgebreide tabellen in de literatuur en in sommige tekstboeken van dit soort kansen. We volstaan hier met de formule die relatief eenvoudig te evalueren is m.b.v. bijvoorbeeld EXCEL.

Wil men de nulhypothese toetsen versus een *tweezijdige alternatieve hypothese*, dan onderscheiden we de situatie dat $r_1 = r_2$ en/of $c_1 = c_2$ enerzijds, en $r_1 \neq r_2$ én $c_1 \neq c_2$ anderzijds. In het eerste geval kunnen we de eerder bepaalde (eenzijdige) overschrijdingskans *verdubbelen* ten einde de tweezijdige overschrijdingskans te krijgen. In het tweede geval dienen we de bovenvermelde overschrijdingskans te vermeerderen met

$$\sum_{i=0}^{\min(x,r_2-c_1+x)} P(x, r_1-x, c_1-x, r_2-c_1+x, -i) =$$
$$\sum_{i=0}^{\min(x,r_2-c_1+x)} \frac{r_1!r_2!c_1!c_2!}{N!(x-i)!(r_1-x+i)!(c_1-x+i)!(r_2-c_1+x-i)!}$$

waarbij *x* bepaald wordt door de eis dat $\frac{a}{r_1} - \frac{c}{r_2} = \frac{c_1 - x}{r_2} - \frac{x}{r_1}$ ofwel $x = 2r_1c_1 / N - a$ (naar beneden afgerond indien *x* niet geheeltallig).

19.5 Voorbeeld Fisher's exacte toets

Voorbeeld 19.2 Stel dat onderstaande tabel de bekende gegevens weergeeft met betrekking tot het succes na behandeling van een bepaalde ernstige zeldzame ziekte. Onder het kopje '*S*' zijn de behandelde aantallen mannen en vrouwen te vinden die 5 jaar na de behandeling nog in leven zijn; onder het kopje '*F*' vinden we de aantallen die binnen 5 jaar na de behandeling zijn overleden.

	S	*F*	
mannen	8	3	11
vrouwen	5	11	16
	13	14	27

We willen onderzoeken of de overlevingskansen na behandeling voor mannen en vrouwen verschillend zijn (en toetsen dus tweezijdig). De eerste stap is transformatie van de tabel zodat $r_1 \geq r_2$ en $a/r_1 \geq c/r_2$:

	F	S	
vrouwen	11	5	16
mannen	3	8	11
	14	13	27

De tweede stap is het bepalen van x: $x = 2 \times 16 \times 14/27 - 11 = 5$ (naar beneden afgerond). De som van de afzonderlijke kansen op de tabellen (11,5,3,8), (12,4,2,9), (13,3,1,10), (14,2,0,11) en (5,11,9,2), (4,12,10,1), (3,13,11,0) zijn respectievelijk, $0{,}035931 + 0{,}004990 + 0{,}000307 + 0{,}000006 \approx 0{,}0412$, en $0{,}011977 + 0{,}000998 + 0{,}000028 \approx 0{,}013$. Als we dus tweezijdig toetsen met onbetrouwbaarheidsdrempel $\alpha = 0{,}05$, dan kunnen we de nulhypothese (net) niet verwerpen aangezien $0{,}0412 + 0{,}013 > 0{,}05$.

20. Verdelingsvrije toetsen

20.1 Inleiding

De normale verdeling speelt een belangrijke rol in de tot nu toe besproken (Hoofdstukken 16 t/m 19) toetsprocedures. Het gebruik van de normale verdeling is vaak ook gebaseerd op de Centrale Limiet Stelling. Er zijn echter nogal wat situaties waarin het onduidelijk is wat de onderliggende verdeling is van de waarnemingen, en waarin het ook onduidelijk is of de kansverdeling van de toetsingsgrootheid op voldoend nauwkeurige wijze kan worden benaderd door een normale verdeling (of een andere parametrische verdeling zoals de Chikwadraat-, Student-, of F-verdeling). Een aanpak voor dit soort situaties wordt geboden door de zogenaamde **verdelingsvrije** (of **niet-parametrische**) toetsen die in grote mate toepasbaar zijn onafhankelijk van de aard van de onderliggende verdeling van de waarnemingen. De meeste verdelingsvrije toetsen zijn gebaseerd op zogenaamde “order-statistics”, d.w.z. dat de waarnemingen worden vervangen door **rangnummers** die op hun beurt worden gebruikt in de toetsingsgrootheid. In dit hoofdstuk worden enkele van de meest gebruikte verdelingsvrije toetsen besproken.

20.2 De rangteken-toets van Wilcoxon (‘Wilcoxon Signed Rank Test’)

De rangteken-toets van Wilcoxon is bedoeld om te toetsen of een bepaalde steekproef afkomstig zou kunnen zijn uit een *symmetrisch* verdeelde populatie met mediaan gelijk aan 0, maar kan ook worden gebruikt in situaties waarin we beschikken over gepaarde waarnemingen en willen weten of beide onderliggende variabelen dezelfde mediaan hebben, ofwel dat de *verschillen* tussen de gepaarde waarnemingen afkomstig zou kunnen zijn uit een symmetrisch verdeelde populatie met mediaan 0. Een veelvoorkomende situatie waarin deze toets van toepassing is, is wanneer we beschikken over metingen aan dezelfde personen vòòr en na een bepaalde behandeling (bijvoorbeeld het gebruik van bloeddrukverlagende medicijnen). In geval van gepaarde waarnemingen $(X_1, Y_1),...,(X_n, Y_n)$ bepalen we de verschillen

$$D_i = X_i - Y_i, \quad i = 1,...,n$$

We veronderstellen dat alle waarden van *D* gelijk aan 0 weggelaten zijn (deze zijn niet bruikbaar voor de toets). We vervangen nu de *absolute* waarden van $D_i, i = 1,...,n$ door rangnummers $R_1,...,R_n$, waarbij de kleinste waarde $|D_i|$ rangnummer 1 krijgt en de grootste rangnummer *n*. In het geval dat 2 of meer opeenvolgende waarden van $|D_i|$ gelijk zijn (een situatie die niet te vaak mag voorkomen), worden *gemiddelde* rangnummers toegekend: bijvoorbeeld als de vier kleinste waarden van $|D_i|$ gelijk zijn, dan krijgen zij alle het rangnummer $(1+2+3+4)/4 = 2{,}5$.

De volgende stap is het sommeren van alle rangnummers behorende bij de positieve verschillen D_i en bij de negatieve verschillen:

$$W_+ = \sum_{D_i>0} R_i$$
$$W_- = \sum_{D_i<0} R_i$$

Voor elk van de drie mogelijke alternatieve hypothesen is de toetsingsgrootheid als volgt:

H_1 : de mediaan is niet gelijk aan 0 $\Rightarrow$

toetsingsgrootheid is $W = \min(W_+, W_-)$; bepaal de **kritieke grens** m.b.v. Tabel C5 na *halvering* van de onbetrouwbaarheidsdrempel α (dus voor $\alpha = 0{,}1$ zoekt men in Tabel C5 de kritieke grens bij $\alpha = 0{,}05$).

H_1 : de mediaan is kleiner dan 0 (de mediaan van X is kleiner dan die van Y) $\Rightarrow$

toetsingsgrootheid is $W = W_+$; bepaal de **kritieke grens** m.b.v. Tabel C5.

H_1 : de mediaan is groter dan 0 (de mediaan van X is groter dan die van Y) $\Rightarrow$

toetsingsgrootheid is $W = W_-$; bepaal de **kritieke grens** m.b.v. Tabel C5.

20.3 Voorbeeld rangteken-toets van Wilcoxon

Voorbeeld 20.1 Men verricht een onderzoek naar het effect van sport op de bloeddruk. Daartoe meet men van 10 willekeurige mannen die de laatste 5 jaren geen sport beoefend hebben de bloeddruk (om meetfouten zoveel mogelijk te voorkomen meet men de bloeddruk onder verschillende omstandigheden en neemt het gemiddelde). Vervolgens laat men de 10 mannen gedurende een maand deelnemen aan een bepaalde sporttraining en registreert na die maand opnieuw de (gemiddelde) bloeddruk. De onderzoeker wil met de rangteken-toets van Wilcoxon en met $\alpha = 0{,}05$ nagaan of sport de bloeddruk positief beïnvloedt (d.w.z. dat na een maand sporten de gemiddelde bloeddruk is gedaald). Onderstaande tabel vermeldt de meetgegevens en de transformatie naar rangnummers.

	1	2	3	4	5	6	7	8	9	10
1e meeting	140	125	110	130	170	165	135	140	155	145
2e meeting	137	137	102	104	172	125	140	110	140	126
verschil	3	-12	8	26	-2	40	-5	30	15	19
rangnummer	2	*5*	4	8	*1*	10	*3*	9	6	7

Gezien de onderzoekshypothese (H_1) is de waarde van de toetsingsgrootheid $W = W_-$ gelijk aan $5 + 1 + 3 = 9$. Volgens Tabel C5 is de kritieke grens in dit geval 10, zodat we bij een onbetrouwbaarheidsdrempel van 5%, aangezien de waarde van de toetsingsgrootheid kleiner is dan 10, de nulhypothese dat sport geen effect heeft op de bloeddruk kunnen verwerpen ten gunste van de alternatieve hypothese dat sportbeoefening bloeddrukverlagend werkt.

20.4 Wilcoxon's rangsom-toets

De rangsom-toets van Wilcoxon is bedoeld om m.b.v. twee onafhankelijke aselecte steekproeven (respectievelijk bestaande uit n en m waarnemingen) uit twee verschillende populaties te toetsen dat de twee populaties gelijk zijn, tegen de alternatieve hypothese dat ze niet gelijk zijn maar, in het bijzonder, dat de locaties (medianen) van de verdelingen verschillend zijn. Deze toets is ook bekend als de Mann-Whitney 2-steekproeven toets (die er equivalent mee is maar een ander toetsingsgrootheid hanteert) en is een (verdelingsvrije) alternatieve toets voor de 2-steekproeven t-toets (zie §16.5 en §16.6). De rangsom-toets van Wilcoxon heeft zeker de voorkeur boven de 2-steekproeven t-toets indien men er aan twijfelt of aan de aannames voor de t-toets wel voldaan is.

Voor de rangsom-toets van Wilcoxon worden alle $n+m$ waarnemingen samen als één rij geordend van klein naar groot en vervolgens vervangen door rangnummers (het kleinste getal krijgt rangnummer 1, het grootste rangnummer $n+m$). Vervolgens bepalen we de grootheid W_n die de som is van de rangnummers behorende bij de eerste steekproef (met n waarnemingen); met W_m noteren we dan de rangsom van de tweede steekproef. Om het rekenwerk te minimaliseren veronderstellen we dat $n \le m$; zo nodig dienen de twee steekproeven dus te worden verwisseld. Merk op dat de grootheid W_n onder de nulhypothese een verdeling heeft die *symmetrisch* is rond $\overline{W} = n \times (1+n+m)/2$ (n keer het gemiddelde rangnummer).

Indien nu de **alternatieve hypothese** luidt dat de locatie van de eerste verdeling *kleiner* is dan die van de tweede, dan is de **toetsingsgrootheid** de boven gedefinieerde grootheid W_n (de toets is **links-eenzijdig**). Indien de **alternatieve hypothese** luidt dat de locatie van de eerste verdeling *groter* is dan die van de tweede, dan is de **toetsingsgrootheid** $2\overline{W} - W_n$ (de toets is **rechts-eenzijdig**). Voor een **twee-zijdige alternatieve hypothese** (de locaties zijn *verschillend*), is de **toetsingsgrootheid** het minimum van W_n en $2\overline{W} - W_n$. Indien – afhankelijk van de alternatieve hypothese – de **waarde van de toetsingsgrootheid** is bepaald, kan met behulp van Tabel C6 in de appendix worden vastgesteld of de nulhypothese kan worden verworpen. H_0 wordt verworpen indien die waarde kleiner is dan of gelijk aan de getabelleerde waarde bij de gewenste onbetrouwbaarheidsdrempel α; als echter bij een twee-zijdige toets de onbetrouwbaarheidsdrempel bijvoorbeeld gelijk is aan 0,10 dient men in de tabel te kijken onder $\alpha = 0{,}05$, de helft dus, aangezien de getabelleerde waarden horen bij een eenzijdige toets.

Tot slot: Tabel C6 is zeer beperkt; in tekstboeken en verdere literatuur vindt men (veel) uitgebreidere tabellen. De kansverdeling van W_n onder de nulhypothese kan echter voor grotere waarden van *n* en *m* benaderd worden door een normaalverdeling. Voor $n = m \geq 25$ kan als toetsingsgrootheid worden genomen de (in benadering) standaardnormaal verdeelde grootheid

$$Z = \frac{W_n - \overline{W}}{\sigma_{W_n}}, \text{ met } \sigma_{W_n} = \sqrt{nm(1+n+m)/12}$$

Echter voor kleinere waarden van *n* en *m* (bijv. $n > 10$ en $m > 10$) kan de benadering al redelijk genoemd worden.

20.5 Voorbeeld rangsom-toets van Wilcoxon

Een farmaceutisch bedrijf wil een nieuw middel voor cholesterolverlaging op de markt brengen. In een pilot-onderzoek voor het vaststellen van de effectiviteit worden 21 mannen tussen 40 en 45 jaar met een vergelijkbaar verhoogd cholesterol niveau geselecteerd (met een verleden zonder het gebruik van cholesterolverlagende middelen) van wie er 11 (random gekozen) het nieuwe middel toegediend krijgen en de overige 10 het gangbare middel. Na gedurende enkele weken het desbetreffende middel geslikt te hebben, wordt bij alle onderzoekspersonen het cholesterolniveau (om precies te zijn, de totaal-cholesterol/HDL-cholesterolratio, een maat voor de kans op hart- en vaatziekten) vastgesteld, en het *verschil* (oude - nieuwe waarde) bepaald met het niveau bij de start van het onderzoek.

De resultaten staan in onderstaande tabel.

Steekproef	Resultaten na enkele weken slikken (reeds geordend per steekproef)										
I (nieuw)	-0.77	0.16	0.33	0.74	1.54	1.61	1.70	2.24	2.24	2.50	2.59
II (gangbaar)	-1.18	0.09	0.16	0.24	0.29	1.18	1.43	1.49	1.79	3.03	

We willen met een onbetrouwbaarheidsniveau van $\alpha = 0{,}05$ toetsen of op grond van deze resultaten de nulhypothese dat beide middelen gemiddeld hetzelfde effect hebben mag worden verworpen ten

gunste van de alternatieve hypothese dat het nieuwe middel beter is (d.w.z. gemiddeld grotere verschillen). Aangezien de eerste steekproef de meeste waarnemingen telt, verwisselen we beide steekproeven en kennen daarna rangnummers toe. Aangezien er sprake is van gelijken, kennen we waar nodig gemiddelde rangnummers toe (merk op dat in deze situatie de validiteit van de toets wel enigszins afneemt).

Steekproef	Resultaten na enkele weken slikken										
I (gangbaar)	-1.18	0.09	0.16	0.24	0.29	1.18	1.43	1.49	1.79	3.03	
II (nieuw)	-0.77	0.16	0.33	0.74	1.54	1.61	1.70	2.24	2.24	2.50	2.59

21. Lineaire regressie

21.1 Inleiding

Met (**enkelvoudige**) **lineaire regressie** kunnen we onderzoeken of er een *lineair* verband bestaat tussen een stochastische variabele Y en een variabele x, zoals (klassiek voorbeeld) de gemiddelde lengte van de volwassen kinderen uit een gezin (Y) en de gemiddelde lengte van hun ouders (x). Een lineaire relatie met Y als **afhankelijke variabele** en x als **onafhankelijke variabele** heeft de vorm $Y = \beta_0 + \beta_1 x$, waarin β_0 het zogenaamde **intercept** is en β_1 de **richtingscoëfficiënt**. Aangezien de afhankelijke variabele Y echter meestal mede bepaald wordt door (verstorende) effecten die niet kunnen worden verklaard door x, is een meer realistisch model

$$Y = \beta_0 + \beta_1 x + E$$

waarin E de stochastische variabele $Y - \beta_0 - \beta_1 x$ representeert die kan worden opgevat als een verstoring van de veronderstelde deterministische (lineaire) relatie $Y = \beta_0 + \beta_1 x$.

Indien men in de praktijk beschikt over een aselecte steekproef $(x_1, y_1),...,(x_n, y_n)$, wil men β_0 en β_1 in het **regressiemodel**

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1,...,n$$

zodanig schatten dat de fouttermen (**residuen**) $e_i, i = 1,...,n$ 'zo klein mogelijk' zijn. Om precies te zijn past men de zogenaamde **kleinste-kwadraten methode** toe om schattingen van de **regressie coëfficiënten** β_0 en β_1 te vinden: men kiest β_0 en β_1 zodanig dat de **fout-kwadraten som**

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} e_i^2$$

zo klein mogelijk is. Dit levert eenduidige schattingen $\hat{\beta}_0$ en $\hat{\beta}_1$ van β_0 en β_1 op. De lijn

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

noemen we de **kleinste-kwadratenlijn** of **regressielijn**. Een belangrijke eigenschap van de resulterende schattingen is

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} e_i = 0$$

m.a.w. de som van de (geschatte) residuen is 0. Het resultaat van deze berekeningen is dat we voor ieder (nieuw) element uit de populatie met de waarde *x* een *voorspelling* kunnen doen m.b.t. de waarde van de afhankelijke variabele *Y*: *gemiddeld* verwachten we de waarde $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. De kansverdeling van de storingsterm *E* speelt een belangrijke rol bij de kwaliteit van die voorspelling.

21.2 Het schatten van de regressiecoëfficiënten β_0 en β_1

Zonder de afleiding te geven vermelden we hier de formules voor de berekening van $\hat{\beta}_0$ en $\hat{\beta}_1$ gegeven de steekproefwaarnemingen $(x_1, y_1),...,(x_n, y_n)$:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \text{en} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Hierin zijn (vgl. §5.4)

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

grootheden die verband houden met de steekproefcovariantie van de paren waarnemingen (x_i, y_i) en de steekproefvariantie van x_i $(i = 1,...,n)$. De formule voor de schatting van β_0 is gebaseerd op het feit dat het 'zwaartepunt' $(\bar{x}, \bar{y})$ op de regressielijn ligt.

21.3 Aannames

Om op basis van statistiek conclusies te kunnen trekken uit deze regressieanalyse dient de kansverdeling van de stochastische storingsterm *E* aan de volgende voorwaarden te voldoen:

1. De verwachtingswaarde van *E* is 0, dus $E(E) = 0$, en dus $E(Y) = \beta_0 + \beta_1 x$;
2. De variantie van *E* is constant en dus niet afhankelijk van de waarde van *x*: $Var(E) = \sigma^2$;
3. De kansverdeling van *E* is een normaal verdeling, dus $E \sim N(0, \sigma^2)$;
4. Er is geen verband tussen *Y* en *E*: het mag bijvoorbeeld niet zo zijn dat de storingsterm positief is voor grote waarden van *Y* en negatief voor kleine.

De resultaten van de regressieanalyse bieden mogelijkheden om een indruk te krijgen van de validiteit van deze aannames. Bovendien kunnen we met die resultaten een schatting, s^2, bepalen van σ^2 en uitspraken doen over de nauwkeurigheid van de schattingen $\hat{\beta}_0$ en $\hat{\beta}_1$.

21.4 Het schatten van de variantie σ^2

Voor het schatten van σ^2 gebruiken we zoals eerder (vgl. §5.3) een gemiddelde kwadratensom:

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2}$$

Hierin is *SSE* een afkorting van '*S*um of *S*quared *E*rrors', SS_{yy} (ook wel de 'totale kwadratensom ') analoog gedefinieerd als SS_{xx}, en de noemer het aantal **vrijheidsgraden** (we hebben uit de data immers 2 parameters geschat). Merk op dat $\hat{\beta}_1 SS_{xy} = SS_{yy} - SSE$, m.a.w. het linkerlid staat voor de door de regressie *verklaarde* variabiliteit van de waarden $y_i, i = 1,...,n$. De wortel, *s*, van s^2 noemen we de **geschatte standaardfout van de regressie**. Een *globale* interpretatie van *s* is dat ongeveer 95% van de waarden $y_i, i = 1,...,n$ liggen tussen de lijnen $y = \hat{\beta}_0 + \hat{\beta}_1 x + 2s$ en $y = \hat{\beta}_0 + \hat{\beta}_1 x - 2s$.

21.5 Het toetsen van de hypothese $H_0 : \beta_1 = b_1$

Indien (in redelijke mate) voldaan is aan de aannames uit §21.3 mogen we er van uitgaan dat de schatter $\hat{\beta}_1$ een normale verdeling heeft met gemiddelde β_1 en variantie $\sigma^2_{\hat{\beta}_1} = \sigma^2 / SS_{xx}$ ($\hat{\beta}_1$ is dus een **zuivere schatter** van β_1). Aangezien σ meestal onbekend is, vervangen we deze door *s*:

$$s_{\hat{\beta}_1} = \hat{\sigma}_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

Een statistische toets aangaande β_1 volgt de principes van §16.1: de **toetsingsgrootheid**

$$T = \frac{\hat{\beta}_1 - b_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - b_1}{s / \sqrt{SS_{xx}}}$$

heeft een Student's *t*-verdeling met $(n-2)$ vrijheidsgraden (vgl.§21.4) en kan worden gebruikt voor het toetsen van de **nulhypothese**

$$H_0 : \beta_1 = b_1$$

Vaak wil men weten of het lineaire regressiemodel *nut* heeft, m.a.w. of inderdaad een niet triviale lineaire relatie bestaat tussen *Y* en *x*. In dat geval luidt de nulhypothese $H_0 : \beta_1 = 0$ (en is dus $b_1 = 0$). Het geval $b_1 \neq 0$ kan worden gebruikt om te toetsen of de richtingscoëfficiënt een bepaalde specifieke waarde heeft (merk op dat volgens het model de gemiddelde waarde van *Y* toeneemt met β_1 voor iedere toename met één eenheid van *x*). Onderstaande tabel vermeldt het **kritieke gebied** voor elk van de mogelijke **alternatieve hypothesen**. Hierin is *t* de **waarde** van bovenvermelde toetsingsgrootheid *T* en α de gehanteerde onbetrouwbaarheidsdrempel (vgl. Tabel 16.4).

H_0	H_1	Kritieke grens	Verwerp H_0 als
$\beta_1 \le b_1$ of: $\beta_1 = b_1$	$\beta_1 > b_1$	$t_{n-2;\alpha}$	$t > t_{n-2;\alpha}$
$\beta_1 \ge b_1$ of: $\beta_1 = b_1$	$\beta_1 < b_1$	$-t_{n-2;\alpha}$	$t < -t_{n-2;\alpha}$
$\beta_1 = b_1$	$\beta_1 \ne b_1$	$t_{n-2;\alpha/2}$	$\lvert t \rvert > t_{n-2;\alpha/2}$

Vanwege de omvangrijke berekeningen zal men regressie vrijwel altijd uitvoeren met statistische software. In de output van een regressie analyse wordt doorgaans de **overschrijdingskans**, *p*, gegeven voor het toetsen van $H_0 : \beta_1 = 0$ vs. de (tweezijdige) alternatieve hypothese $H_1 : \beta_1 \ne 0$. Men kan de waarde van *p* als volgt gebruiken voor het toetsen van $H_0 : \beta_1 = 0$ (afhankelijk van H_1)

H_1	Verwerp H_0 als
$\beta_1 \ne 0$	$p < \alpha$
$\beta_1 > 0$	$t > 0$ én $p/2 < \alpha$
$\beta_1 < 0$	$t < 0$ én $p/2 < \alpha$

21.6 Betrouwbaarheidsinterval voor β_1

Naast de **puntschatting**, $\hat{\beta}_1$, kunnen we ook een betrouwbaarheidsinterval voor β_1 opstellen (vgl. §14.3); een dergelijke intervalschatting geeft direct inzicht in het nut van het regressiemodel. De formule voor een $100(1-\alpha)\%$ -**betrouwbaarheidsinterval** voor β_1 is

$$(\hat{\beta}_1 - t_{n-2;\,\alpha/2} \times s_{\hat{\beta}_1},\, \hat{\beta}_1 + t_{n-2;\,\alpha/2} \times s_{\hat{\beta}_1}) \text{ ofwel } \hat{\beta}_1 \pm t_{n-2;\,\alpha/2} \times s_{\hat{\beta}_1}$$

De *interpretatie* van zo'n interval is dat we er voor $100(1-\alpha)\%$ *vertrouwen* in hebben dat het gegeven interval de *werkelijke* waarde van de richtingscoëfficiënt, β_1, bevat. Men kan ook een betrouwbaarheidsinterval voor β_0 opstellen; echter, omdat een dergelijk interval van weinig praktische betekenis is, vermelden we het hier niet.

21.7 De correlatie- en determinatiecoëfficiënt

Enkelvoudige lineaire regressie is een statistische methode voor het vaststellen van lineaire samenhang tussen twee variabelen. In §5.4 is reeds de **correlatiecoëfficiënt** als maat voor lineaire samenhang besproken. Uitgedrukt in de notatie van dit hoofdstuk is de formule

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

De **determinatiecoëfficiënt**, r^2 (vaak ook genoteerd als R^2), is eenvoudigweg het kwadraat van *r*. De volgende uitdrukkingen zijn equivalent voor r^2 (vergelijk eerder vermelde uitdrukkingen voor diverse grootheden):

$$r^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} = \frac{\hat{\beta}_1 SS_{xy}}{SS_{yy}} = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

r^2 is de proportie door de regressie verklaarde variantie van de variabiliteit van de waarden van de afhankelijke variabele *Y* (vgl. §21.4). Als maatstaf voor het *nut* van het regressiemodel is r^2 minder geschikt; daarvoor dient men $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ te toetsen (zie §21.5). Merk op dat het toetsen van $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$, waarbij ρ de populatiecorrelatiecoëfficiënt is, hiermee equivalent is.

21.8 Intervalschattingen voor een *individuele* waarneming, gegeven $x = x_p$, en voor de *gemiddelde* waarneming, gegeven $x = x_p$

Een belangrijke functie van de regressielijn is het voorspellen van de waarde van de afhankelijke variabele, *Y*, voor een gegeven waarde van de onafhankelijke variabele, *x* (zie ook §21.1). Bij het opstellen van **intervalschattingen** onderscheid men twee verschillende situaties. Indien men een intervalschatting wil voor de *gemiddelde* waarde van *Y*, gegeven een waarde, x_p, van *x*, dan is

$$\hat{y} \pm t_{n-2;\alpha/2} \times s \times \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

het gebruikelijke $100(1-\alpha)\%$ - **betrouwbaarheidsinterval** voor de *gemiddelde* waarde van *Y*, gegeven $x = x_p$. Indien men daarentegen een intervalschatting wil voor de *individuele* waarde van *Y* die we bij $x = x_p$ kunnen verwachten, dan is

$$\hat{y} \pm t_{n-2;\alpha/2} \times s \times \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

het gebruikelijke $100(1-\alpha)\%$ - **voorspellingsinterval** (altijd *groter* dan het corresponderende betrouwbaarheidsinterval). Het begrip *betrouwbaarheidinterval* wordt i.h.a. gebruikt voor een intervalschatting van een populatieparameter (zoals hier $E(Y \mid x_p)$); het begrip *voorspellingsinterval* wordt doorgaans gebruikt voor een intervalschatting van een individuele waarde van een stochastische variabele (zoals hier $(Y \mid x_p)$).

Steekproef	Rangnummers											som
I (gangbaar)	1	3	4.5	6	7	10	11	12	16	21		**91.5**
II (nieuw)	2	4.5	8	9	13	14	15	17.5	17.5	19	20	139.5

Aangezien we willen toetsen of er aanleiding is te veronderstellen dat het nieuwe middel beter is, is de toets links-eenzijdig (hoe kleiner de som van de rangnummers behorende bij steekproef I (gangbaar), des te meer reden is er om aan te nemen dat het nieuwe middel beter is). We vergelijken dus de **waarde van de toetsingsgrootheid**, 91.5, met de kritieke grens uit Tabel C6 bij $\alpha = 0{,}05$, 86, en concluderen dat we H_0 niet mogen verwerpen; er is op basis van dit experiment onvoldoende reden om aan te nemen dat het nieuwe middel beter is. Op grond van de resultaten zou het bedrijf echter kunnen besluiten tot een nieuw onderzoek, met meer proefpersonen. Toepassing van de standaardnormale benadering levert dezelfde conclusie op: de waarde van de toetsingsgrootheid is dan $z = (91{,}5 - 110)/14{,}2 \approx -1{,}30$; de bijbehorende overschrijdingskans is $\Phi(-1{,}30) \approx 0{,}096$, groter dan α (dus H_0 niet verwerpen).

A. Statistische termen: Engels-Nederlands

addition rule	somregel
alternative hypothesis	alternatieve hypothese (onderzoeks-)
analysis of variance (ANOVA)	variantie analyse
approximation	benadering
Bayes' rule	regel van Bayes
binomial	binomiaal
binomial coefficient	binomiaal-coëfficiënt
binomial distribution	binomiaal verdeling
binomial experiment	binomiaal-experiment
bivariate	bivariaat
Central Limit Theorem	Centrale Limiet Stelling (CLS)
Chebyshev's theorem	ongelijkheid van Chebyshev
chi-square distribution	chi-kwadraat verdeling
coefficient of correlation	correlatiecoëfficiënt
coefficient of determination	determinatiecoëfficiënt
coefficient of variation	variatiecoëfficiënt
combination	combinatie
combinatorics	combinatoriek

complement	complement
condition	voorwaarde
conditional probability	voorwaardelijke kans
confidence	betrouwbaarheid
confidence interval	betrouwbaarheidsinterval
confidence level (significance -)	onbetrouwbaarheidsdrempel
contingency tabel (crosstable)	kruistabel (contingentie -)
continuous	continu
correction for continuity	continuiteitscorrectie
covariance	covariantie
critical region (rejection -)	kritieke gebied (verwerpingsgebied)
critical value	kritieke waarde
cross table	kruistabel
degrees of freedom	vrijheidsgraden
density	dichtheid
dependent	afhankelijk
dependent variable	afhankelijke variabele
discrete	discreet
disjoint (mutually exclusive)	disjunct
dissection	dissectie
distribution (probability -)	verdeling (kans-)
distribution function (cumulative -), cdf	verdelingsfunctie (cumulatieve -)
error sum of squares	fout-kwadraten som
estimate	schatten, schatting
estimator	schatter
event	gebeurtenis
expectation, expected value	verwachting, verwachtingswaarde
exponential distribution	exponentiële verdeling
extreme value	extreme waarde, uitschieter, uitbijter
failure	mislukking
F-distribution	F-verdeling
geometric distribution	geometrische verdeling
hypergeometric distribution	hypergeometrische (kans)verdeling
hypothesis	hypothese
independence	onafhankelijkheid
independent	onafhankelijk
independent variable	onafhankelijke variabele (verklarende -)
intercept	intercept (regressieconstante)
interquartile range	interkwartielsafstand
intersection	doorsnede
interval variable	interval variabele
joint (probability) density (function)	gezamenlijke (kans)dichtheid(sfunctie)
joint probability distribution	gezamenlijke kansverdeling

k-factorial	k-faculteit
least squares line	kleinste-kwadraten lijn (regressielijn)
level of significance	onbetrouwbaarheidsdrempel
likelihood	waarschijnlijkheid
lineair regression (single/simple - -)	lineaire regressie (enkelvoudige - -)
linear relation	lineaire relatie
linear transformation	lineaire transformatie
location measure	locatiemaat
margin	marge (rand)
marginal distribution	marginale verdeling
maximum likelihood estimator	maximum likelihood schatter
mean	gemiddelde
measure of variability	spreidingsmaat
median	mediaan
mode	modus
multinomial	multinomiaal
multinomial experiment	multinomiaal experiment
multiplicative rule of probability	productregel (vermenigvuldigingsregel) voor kansen
negative binomial distribution	negatief binomiaal verdeling
nominal variable	nominale variabele
non-parametric test (distribution-free -)	niet-parametrische toets (verdelingsvrije -)
normal distribution (standard -)	normale (kans)verdeling (standaard- -)
null hypothesis	nulhypothese
one-sided alternative hypothesis	eenzijdige alternatieve hypothese
ordinal variable	ordinale variabele
outcome	uitkomst
outlier	uitschieter, uitbijter, extreme waarde
parameter	parameter
Pascal's triangle	driehoek van Pascal
permutation	permutatie
point estimator	puntschatter
Poisson distribution	Poisson verdeling
population	populatie
prediction interval	voorspellingsinterval
probability	kans
probability density function, density, pdf	dichtheid, kansdichtheid(sfunctie)
probability function	kansfunctie
probability space	kansruimte
probability theory	kansrekening
p-value	overschrijdingskans, p-waarde
quartile	kwartiel
random experiment	stochastisch experiment
random sample	aselecte steekproef
random variable (stochastic -)	toevalsvariabele, stochast, stochastische variabele

random vector	stochastische vector
range	bereik
ratio variable	ratio variabele
regression coefficients	regressie coëfficiënten
rejection of nulhypothesis	verwerping van nulhypothese
rejection region (critical)	kritieke gebied (verwerpingsgebied)
replacement (with/without)	teruglegging (met/zonder)
rule of thumb	vuistregel
sample	steekproef
sample correlation coefficient	steekproefcorrelatiecoëfficiënt
sample covariance	steekproefcovariantie
sample distribution	steekproefverdeling
sample mean	steekproefgemiddelde
sample size	steekproefgrootte (-omvang)
sample space	uitkomstenruimte
sample standard deviation	steekproefstandaardafwijking
sample variance	steekproefvariantie
set	verzameling
short-cut formula	rekenformule
significance level (confidence -)	onbetrouwbaarheidsdrempel
single	enkelvoudig

slope	richtingscoëfficiënt
standard deviation	standaardafwijking
standard error	standaard fout
standardise	standaardiseren
stochastic variable (random -)	toevalsvariabele, stochast, stochastische variabele
Student's t-distribution	Student's t-verdeling
subset	deelverzameling
success	succes
sum of squares	kwadratensom
test	toets(en)
test statistic	toetsingsgrootheid
two-sided alternative hypothesis	tweezijdige alternatieve hypothese
type I (II) error	fout van de eerste (tweede) soort
unbiased estimator	zuivere schatter
uniform distribution	uniforme (kans)verdeling, rechthoekige -
union	vereniging
value of test statistic	waarde van de toetsingsgrootheid
variance	variantie
variation	variatie
Venn diagram	Venndiagram
weak law of large numbers	zwakke wet van de grote aantallen
weighted mean	gewogen gemiddelde
Wilcoxon rank sum test	rangsom toets van Wilcoxon
Wilcoxon signed rank test	rangteken-toets van Wilcoxon

B. Overzicht discrete verdelingen

Type verdeling	Beschrijving	Puntmassaverdeling	Verwachtings-waarde	Variantie
Binomiaal verdeling $X \sim \text{Bin}(n,p)$	aantal successen onder n binomiaal experimenten (p is kans op succes)	$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$ $(k=0,...,n)$	$\mu_X = np$	$\sigma_X^2 = np(1-p)$
Poisson verdeling $X \sim \text{Pois}(\lambda)$	aantal 'incidenten' per tijdseenheid	$P(X=k) = e^{-\lambda}\lambda^k / k!$ $(k=0,1,2,...)$	$\mu_X = \lambda$	$\sigma_X^2 = \lambda$
Geometrische verdeling $X \sim \text{Geo}(p)$	'wachttijd' tot optreden volgende gebeurtenis	$P(X=k) = p(1-p)^k$ $(k=0,1,2,...)$	$\mu_X = \frac{1-p}{p}$	$\sigma_X^2 = \frac{1-p}{p^2}$
Negatief binomiaal verdeling $X \sim \text{NB}(r,p)$	som van aantal onderling onafhankelijke geometrisch verdeelde variabelen	$P(X=k) = \binom{r+k-1}{k} p^r (1-p)^k$ $(k=0,1,2,...)$	$\mu_X = \frac{r(1-p)}{p}$	$\sigma_X^2 = \frac{r(1-p)}{p^2}$
Hypergeometrische verdeling $X \sim \text{HG}(n,N,S)$	aantal 'successen' in een steekproef (zonder teruglegging) ter grootte n uit N objecten waarvan er S de gewenste eigenschap hebben	$P(X=k) = \frac{\binom{S}{k}\binom{N-S}{n-k}}{\binom{N}{n}}$ $(k=0,...,n)$	$\mu_X = nS/N$	$\sigma_X^2 = \frac{nS(N-S)(N-n)}{N^2(N-1)}$
$\mathbf{X} \sim \text{Mult}(n,p_1,...,p_r)$	aantallen in elk van r categoriën	$P(\mathbf{X}=(k_1,...,k_r)) = \binom{n}{k_1 \ldots k_r} p_1^{k_1} \times ... \times p_1^{k_1}$	-	-

C1. De cumulatieve standaardnormaal verdeling

De tabel vermeldt voor z-waarden variërend van 0,00 met stapjes van 0,01 tot en met 3,59 de waarde van de cumulatieve verdelingsfunctie $\Phi(z)$, Voor bijvoorbeeld $z = 1{,}03$ vinden we $\Phi(1{,}03) = 0{,}8485$ op het kruispunt van de rij met 1,0 in de marge en de kolom met 0,03 in de marge.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621

1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

C2. Cumulatieve Chi-kwadraat verdeling

De tabel vermeldt bij de cumulatieve kansen p (variërend van 0,005 tot 0,995) de waarde (χ^2) van een chi-kwadraat variabele met aantal vrijheidsgraden gelijk aan df (variërend van 1 tot 100) die voldoet aan $P(\chi^2_{df} < \chi^2) = p$. Men leest hieruit bijvoorbeeld dat $\chi^2_{17;0,1} \approx 24{,}8$ (vgl. **Voorbeeld 16.3**).

	p									
df	**0,005**	**0,010**	**0,025**	**0,050**	**0,100**	**0,900**	**0,950**	**0,975**	**0,990**	**0,995**
1	0,0000393	0,000157	0,000982	0,00393	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,020	0,051	0,103	0,211	4,61	5,99	7,38	9,21	10,6
3	0,0717	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3	12,8
4	0,207	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3	14,9
5	0,412	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1	16,7
6	0,676	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8	18,5
7	0,989	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5	20,3
8	1,34	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1	22,0
9	1,73	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7	23,6
10	2,16	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2	25,2
11	2,60	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7	26,8
12	3,07	3,57	4,40	5,23	6,30	18,5	21,0	23,3	26,2	28,3
13	3,57	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7	29,8
14	4,07	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1	31,3
15	4,60	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6	32,8
16	5,14	5,81	6,91	7,96	9,31	23,5	26,3	28,8	32,0	34,3
17	5,70	6,41	7,56	8,67	10,1	24,8	27,6	30,2	33,4	35,7
18	6,26	7,01	8,23	9,39	10,9	26,0	28,9	31,5	34,8	37,2
19	6,84	7,63	8,91	10,1	11,7	27,2	30,1	32,9	36,2	38,6
20	7,43	8,26	9,59	10,9	12,4	28,4	31,4	34,2	37,6	40,0
21	8,03	8,90	10,3	11,6	13,2	29,6	32,7	35,5	38,9	41,4
22	8,64	9,54	11,0	12,3	14,0	30,8	33,9	36,8	40,3	42,8
23	9,26	10,2	11,7	13,1	14,8	32,0	35,2	38,1	41,6	44,2
24	9,89	10,9	12,4	13,8	15,7	33,2	36,4	39,4	43,0	45,6
25	10,5	11,5	13,1	14,6	16,5	34,4	37,7	40,6	44,3	46,9

26	11,2	12,2	13,8	15,4	17,3	35,6	38,9	41,9	45,6	48,3
27	11,8	12,9	14,6	16,2	18,1	36,7	40,1	43,2	47,0	49,6
28	12,5	13,6	15,3	16,9	18,9	37,9	41,3	44,5	48,3	51,0
29	13,1	14,3	16,0	17,7	19,8	39,1	42,6	45,7	49,6	52,3
30	13,8	15,0	16,8	18,5	20,6	40,3	43,8	47,0	50,9	53,7
35	17,2	18,5	20,6	22,5	24,8	46,1	49,8	53,2	57,3	60,3
40	20,7	22,2	24,4	26,5	29,1	51,8	55,8	59,3	63,7	66,8
45	24,3	25,9	28,4	30,6	33,4	57,5	61,7	65,4	70,0	73,2
50	28,0	29,7	32,4	34,8	37,7	63,2	67,5	71,4	76,2	79,5
60	35,5	37,5	40,5	43,2	46,5	74,4	79,1	83,3	88,4	92,0
70	43,3	45,4	48,8	51,7	55,3	85,5	90,5	95,0	100	104
80	51,2	53,5	57,2	60,4	64,3	96,6	102	107	112	116
90	59,2	61,8	65,6	69,1	73,3	108	113	118	124	128
100	67,3	70,1	74,2	77,9	82,4	118	124	130	136	140

C3. Student verdelingen: waarden van $t_{df;\alpha}$

De tabel vermeldt de waarden van $t_{df;\alpha}$ voor variërende waarden van α en vrijheidsgraden *df* variërend van 1 tot 120; voor de laatste regel geldt een aantal vrijheidsgraden van oneindig, zodat de daarvermelde waarden overeenkomen met waarden uit de standaardnormaalverdeling. Zo is bijvoorbeeld $t_{14;0,05} \approx 1{,}76$ (vgl. **Voorbeeld 16.2**) en $t_{\infty;0,025} = z_{0,025} \approx 1{,}960$.

				α			
df	**0,1000**	**0,0500**	**0,0250**	**0,0100**	**0,0050**	**0,0010**	**0,0005**
1	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	1,638	2,353	3,182	4,541	5,841	10,214	12,92
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,894	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,689
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,660
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	1,303	1,684	2,021	2,423	2,704	3,307	3,551
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460
120	1,289	1,658	1,980	2,358	2,617	3,160	3,373
∞	*1,282*	*1,645*	*1,960*	*2,326*	*2,576*	*3,090*	*3,291*

C4. Cumulatieve F-verdeling

De tabel vermeldt bij de cumulatieve kansen p (variërend van 0,9 tot 0,995) de waarde (F) van een F-variabele met m vrijheidsgraden van de teller en n vrijheidsgraden van de noemer (beide variërend van 1 tot 120) die voldoet aan $P(F_n^m < F) = p$. Zo blijkt uit de tabel bijvoorbeeld dat $F_{30;0,05}^2 \approx 3{,}32$. Merk op dat met behulp van de relatie $F_{n;\alpha}^m = 1/F_{m;1-\alpha}^n$ ook F-waarden met kleine $(1-p)$-waarden (of: grote α-waarden) bepaald kunnen worden; dus, bijvoorbeeld: $F_{2;0,95}^{30} = 1/F_{30;0,05}^2 \approx 1/3{,}32 \approx 0{,}301$.

		m														
p	*n*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**20**	**30**	**60**	**120**
	1	39,9	49,5	53,6	55,8	57,2	58,2	58,9	59,4	59,9	60,2	61,2	61,7	62,3	62,8	63,1
	2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,42	9,44	9,46	9,47	9,48
	3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,20	5,18	5,17	5,15	5,14
	4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,87	3,84	3,82	3,79	3,78
	5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,24	3,21	3,17	3,14	3,12
	6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,87	2,84	2,80	2,76	2,74
	7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,63	2,59	2,56	2,51	2,49
0,900	**8**	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,46	2,42	2,38	2,34	2,32
	9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,34	2,30	2,25	2,21	2,18
	10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,24	2,20	2,16	2,11	2,08
	15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	1,97	1,92	1,87	1,82	1,79
	20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,84	1,79	1,74	1,68	1,64
	30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,72	1,67	1,61	1,54	1,50
	60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,60	1,54	1,48	1,40	1,35
	120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,55	1,48	1,41	1,32	1,26

p	*n*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**20**	**30**	**60**	**120**
	1	161	199	216	225	230	234	237	239	241	242	246	248	250	252	253
	2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5
	3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,62	8,57	8,55
	4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,75	5,69	5,66
	5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,50	4,43	4,40
	6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,81	3,74	3,70
	7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,38	3,30	3,27
0,950	**8**	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,08	3,01	2,97
	9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,86	2,79	2,75
	10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,70	2,62	2,58
	15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,25	2,16	2,11
	20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,04	1,95	1,90
	30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,84	1,74	1,68
	60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,65	1,53	1,47
	120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,75	1,66	1,55	1,43	1,35

		m														
p	***n***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**20**	**30**	**60**	**120**
	1	648	799	864	900	922	937	948	957	963	969	985	993	1001	1010	1014
	2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5
	3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,3	14,2	14,1	14,0	13,9
	4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,66	8,56	8,46	8,36	8,31
	5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,23	6,12	6,07
	6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,27	5,17	5,07	4,96	4,90
	7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,57	4,47	4,36	4,25	4,20
0,975	**8**	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,10	4,00	3,89	3,78	3,73
	9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,77	3,67	3,56	3,45	3,39
	10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42	3,31	3,20	3,14
	15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,86	2,76	2,64	2,52	2,46
	20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,57	2,46	2,35	2,22	2,16
	30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,31	2,20	2,07	1,94	1,87
	60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,06	1,94	1,82	1,67	1,58
	120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	1,94	1,82	1,69	1,53	1,43

p	***n***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**20**	**30**	**60**	**120**
	1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6157	6209	6260	6313	6340
	2	98,5	99,0	99,2	99,3	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4	99,5	99,5	99,5
	3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	26,9	26,7	26,5	26,3	26,2
	4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5	14,2	14,0	13,8	13,7	13,6
	5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	9,72	9,55	9,38	9,20	9,11
	6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,56	7,40	7,23	7,06	6,97
	7	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,31	6,16	5,99	5,82	5,74
0,990	**8**	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,52	5,36	5,20	5,03	4,95
	9	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	4,96	4,81	4,65	4,48	4,40
	10	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,56	4,41	4,25	4,08	4,00
	15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,52	3,37	3,21	3,05	2,96
	20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,09	2,94	2,78	2,61	2,52
	30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,70	2,55	2,39	2,21	2,11
	60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,35	2,20	2,03	1,84	1,73
	120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,19	2,03	1,86	1,66	1,53

p	***n***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**20**	**30**	**60**	**120**
0,995	**1**	16212	19997	21614	22501	23056	23440	23715	23924	24091	24222	24632	24837	25041	25254	25358
	2	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
	3	55,6	49,8	47,5	46,2	45,4	44,8	44,4	44,1	43,9	43,7	43,1	42,8	42,5	42,1	42,0
	4	31,3	26,3	24,3	23,2	22,5	22,0	21,6	21,4	21,1	21,0	20,4	20,2	19,9	19,6	19,5
	5	22,8	18,3	16,5	15,6	14,9	14,5	14,2	14,0	13,8	13,6	13,1	12,9	12,7	12,4	12,3
	6	18,6	14,5	12,9	12,0	11,5	11,1	10,8	10,6	10,4	10,3	9,81	9,59	9,36	9,12	9,00
	7	16,2	12,4	10,9	10,1	9,52	9,16	8,89	8,68	8,51	8,38	7,97	7,75	7,53	7,31	7,19
	8	14,7	11,0	9,60	8,81	8,30	7,95	7,69	7,50	7,34	7,21	6,81	6,61	6,40	6,18	6,06
	9	13,6	10,1	8,72	7,96	7,47	7,13	6,88	6,69	6,54	6,42	6,03	5,83	5,62	5,41	5,30
	10	12,8	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,97	5,85	5,47	5,27	5,07	4,86	4,75
	15	10,8	7,70	6,48	5,80	5,37	5,07	4,85	4,67	4,54	4,42	4,07	3,88	3,69	3,48	3,37

20	9,94	6,99	5,82	5,17	4,76	4,47	4,26	4,09	3,96	3,85	3,50	3,32	3,12	2,92	2,81
30	9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,45	3,34	3,01	2,82	2,63	2,42	2,30
60	8,49	5,79	4,73	4,14	3,76	3,49	3,29	3,13	3,01	2,90	2,57	2,39	2,19	1,96	1,83
120	8,18	5,54	4,50	3,92	3,55	3,28	3,09	2,93	2,81	2,71	2,37	2,19	1,98	1,75	1,61

C5. Wilcoxon's rangteken-toets

De tabel vermeldt voor vier waarden van de onbetrouwbaarheidsdrempel α en voor $5 \le n \le 50$ de grootste waarde, $w_{n;\alpha}$, van de toetsingsgrootheid W waarvoor onder de nulhypothese geldt: $P(W \le w_{n;\alpha}) < \alpha$. (Bij een tweezijdige alternatieve hypothese dient men de onbetrouwbaarheidsdrempel eerst te halveren, zie §20.2).

$w_{n;\alpha}$	α			
n	**0,005**	**0,010**	**0,025**	**0,050**
5	-	-	-	0
6	-	-	0	2
7	-	0	2	3
8	0	1	3	5
9	1	3	5	8
10	3	5	8	10
11	5	7	10	13
12	7	9	13	17
13	9	12	17	21
14	12	15	21	25
15	15	19	25	30
16	19	23	29	35
17	23	27	34	41
18	27	32	40	47
19	32	37	46	53
20	37	43	52	60
21	42	49	58	67
22	48	55	65	75
23	54	62	73	83
24	61	69	81	91
25	68	76	89	100
26	75	84	98	110
27	83	92	107	119

$w_{n;\alpha}$	α			
n	**0,005**	**0,010**	**0,025**	**0,050**
28	91	101	116	130
29	100	110	126	140
30	109	120	137	151
31	118	130	147	163
32	128	140	159	175
33	138	151	170	187
34	148	162	182	200
35	159	173	195	213
36	171	185	208	227
37	182	198	221	241
38	194	211	235	256
39	207	224	249	271
40	220	238	264	286
41	233	252	279	302
42	247	266	294	319
43	261	281	310	336
44	276	296	327	353
45	291	312	343	371
46	307	328	361	389
47	322	345	378	407
48	339	362	396	426
49	355	379	415	446
50	373	397	434	466

C6. Wilcoxon's rangsom-toets

De tabel vermeldt de links-eenzijdige kritieke waarden van de toetsingsgrootheid, W_n, van de rangsom-toets van Wilcoxon (zie §20.4). Verondersteld is dat $n \le m$; indien $n = m$ is verondersteld dat W_n de *kleinste* rangsom is. Voor een rechts-eenzijdige toets is de toetsingsgrootheid $2\overline{W} - W_n$. Voor een 2-zijdige toets is de toetsingsgrootheid het minimum van W_n en $2\overline{W} - W_n$; de getabelleerde waarde van α moet dan met 2 vermenigvuldigd worden. De nulhypothese wordt verworpen als de toetsingsgrootheid kleiner dan of gelijk is aan de getabelleerde kritieke waarde.

		α				
	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n = 2	5			3	4	16
	6			3	4	18
	7		-	3	4	20
	8		3	4	5	22
	9		3	4	5	24
	10		3	4	6	26
	11	-	3	4	6	28

		α				
	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n = 3	3			6	7	21
	4		-	6	7	24
	5		6	7	8	27
	6	-	7	8	9	30
	7	6	7	8	10	33
	8	6	8	9	11	36
	9	7	8	10	11	39
	10	7	9	10	12	42
	11	7	9	11	13	45

	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n = 4	4	-	10	11	13	36
	5	10	11	12	14	40
	6	11	12	13	15	44
	7	11	13	14	16	48
	8	12	14	15	17	52
	9	13	14	16	19	56
	10	13	15	17	20	60
	11	14	16	18	21	64

	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n = 5	5	16	17	19	20	55
	6	17	18	20	22	60
	7	18	20	21	23	65
	8	19	21	23	25	70
	9	20	22	24	27	75
	10	21	23	26	28	80
	11	22	24	27	30	85

	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n = 6	6	24	26	28	30	78
	7	25	27	29	32	84
	8	27	29	31	34	90
	9	28	31	33	36	96
	10	29	32	35	38	102
	11	30	34	37	40	108

	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n = 7	7	34	36	39	41	105
	8	35	38	41	44	112
	9	37	40	43	46	119
	10	39	42	45	49	126
	11	40	44	47	51	133

	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n = 8	8	45	49	51	55	136
	9	47	51	54	58	144
	10	49	53	56	60	152
	11	51	55	59	63	160

	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n = 9	9	59	62	66	70	171
	10	61	65	69	73	180
	11	63	68	72	76	189

	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n=10	10	74	78	82	87	210
	11	77	81	86	91	220

	m	0,01	0,025	0,05	0,10	$2\overline{W}$
n = 11	11	91	96	100	106	253

D. Notatie (selectie)

$P(A)$	de kans dat gebeurtenis A optreedt
$P(A \mid B)$	de kans dat gebeurtenis A optreedt, gegeven dat gebeurtenis B optreedt
$\mathbf{R}$	verzameling der reële getallen
μ_X, $E(X)$	verwachtingswaarde van de toevalsvariabele X
σ_X^2, $Var(X)$	variantie van de toevalsvariabele X
s^2	steekproefvariantie
$\rho_{X,Y}$	populatiecorrelatiecoëfficiënt tussen de toevalsvariabelen X en Y
r_{xy}	steekproefcorrelatiecoëfficiënt tussen de toevalsvariabelen X en Y, geschat op basis van de gepaarde waarnemingen $(x_1, y_1),...,(x_n, y_n)$
$\bar{x}_n$	steekproefgemiddelde van een steekproef, $x_1,..., x_n$
H_0	nulhypothese
H_1	alternatieve hypothese (onderzoeks-)
CLS	Centrale Limiet Stelling
$\mathrm{N}(\mu;\sigma^2)$	normale verdeling met parameters μ en σ^2
$\mathrm{Bin}(n, p)$	binomiaal verdeling met parameters n en p
$\mathrm{Pois}(\lambda)$	Poisson verdeling met parameter met parameter λ
$\mathrm{Geo}(p)$	geometrische verdeling met parameter p
$\mathrm{NB}(r, p)$	negatief binomiaal verdeling met parameters r en p
$\mathrm{HG}(n, N, S)$	hypergeometrische verdeling met parameters n, N, en S
$\mathrm{Mult}(n, p_1,..., p_r)$	multinomiale verdeling met parameters $n, p_1,..., p_r$
$\mathrm{U}(a, b)$	uniforme (rechthoekige) verdeling met parameters a en b
$\mathrm{Exp}(\lambda)$	exponentiële verdeling met parameter λ
α	onbetrouwbaarheidsdrempel, fout van de eerste soort
$\varphi(\cdot)$	kansdichtheidsfunctie van de standaardnormale verdeling
$\Phi(\cdot)$	cumulatieve verdelingsfunctie van de standaardnormale verdeling
z_α	$100(1-\alpha)$-percentiel van de standaardnormale verdeling $(=\Phi^{-1}(1-\alpha))$
χ_n^2	chi-kwadraatverdeling met n vrijheidsgraden
$\chi_{n;\alpha}^2$	$100(1-\alpha)$-percentiel van de chi-kwadraat verdeling met n vrijheidsgraden
t_n	Student's t-verdeling met n vrijheidsgraden
$t_{n;\alpha}$	$100(1-\alpha)$-percentiel van de Student verdeling met n vrijheidsgraden
F_n^m	Fisher's F-verdeling met m vrijheidsgraden van de teller en n van de noemer
$F_{n;\alpha}^m$	$100(1-\alpha)$-percentiel van de Fisher's F-verdeling met m vrijheidsgraden van de teller en n van de noemer

E. Index