

Stochastiek voor Informatici
Sara van de Geer
1993

Inhoud

0. Inleiding

1. Waarschijnlijkheidsrekening

1.1. Empirische wet van de grote aantallen

1.2. Gebeurtenissen

1.3. Axiomatische opzet

1.4. Combinatoriek

1.5. Stochastische onafhankelijkheid

1.6. Urnmodellen

1.7. Stochastische grootheden

1.7.1. Discreet verdeelde stochastische grootheden

1.7.2. Continu verdeelde stochastische grootheden

1.8. Simultane verdelingen

1.8.1. Simultane discrete verdelingen

1.8.2. Continue simultane verdelingen

1.8.3. Onafhankelijke stochastische grootheden

1.9. De verdeling van functies van stochastische grootheden

1.10. Verwachting en variantie

1.11. De wet van de grote aantallen

1.12. De centrale limietstelling

2. Mathematische statistiek

2.1. Inleiding

2.2. Schattingstheorie

2.3. De empirische verdelingsfunctie

2.4. Meest aannemelijke schatters

2.5. Regressieanalyse

2.6. Toetsingstheorie

2.7. Toets voor de alternatieve verdeling

2.8. Toetsen voor de normale verdeling

2.8.1. Toetsen van hypothesen over μ ; σ^2 bekend

2.8.2. Toetsen van hypothesen over μ ; σ^2 onbekend

2.9. Studenttoets voor paren

2.10. Twee-steekproeventoets

2.11. Toets voor het vergelijken van twee kansen

2.12. Verdelingsvrije methoden

2.12.1. Verdelingsvrije methoden voor paren van waarnemingen

2.12.2. Twee-steekproeventoets van Wilcoxon

0. Introductie

WAT IS EEN KANS? In het dagelijks taalgebruik komt men het begrip *kans* regelmatig tegen.

Vb. Het gebruik van veiligheidsgordels doet de kans op een ongeluk met dodelijke afloop afnemen.

Vb. De kans van slagen van een experiment is groter als de proefneming door deskundigen wordt verricht.

Het begrip kans komt gedeeltelijk overeen met *mogelijkheid*, en wordt soms geoperationaliseerd door *fractie*, *frequentie*, of *percentage*.

Vb. Van de mensen in Nederland tussen de 18 en 65 jaar heeft $x\%$ een baan.

Een interpretatie van het laatste voorbeeld is dat in Nederland de kans op een baan $x\%$ is. In de wiskunde wordt echter een veel abstracter begrip kans gehanteerd. Kans en fractie zijn in dit college over het algemeen essentieel verschillende concepten. Het idee is (zoals bij de meeste wiskundige theorieën) om een aantal z.g. axioma's op te stellen waaraan een kans moet voldoen, en wel zodanig dat de eigenschappen die volgen uit de axioma's ongeveer voldoen aan een intuïtief idee van kans.

WAARSCHIJNLIJKHEIDSREKENING. In het eerste deel van dit dictaat zullen we het begrip kans als abstractie introduceren, met inbegrip van diverse definities van speciale eigenschappen (b.v. onafhankelijkheid) en bepaalde interessante karakteristieken (verwachting, variantie e.d.).

STATISTIEK. In het tweede deel wordt de waarschijnlijkheidsrekening toegepast om uitspraken te kunnen doen over bepaalde praktische problemen. Voor dergelijke problemen kan men een *model* maken, met als doel het verkrijgen van inzicht in de eventuele structuur. Bovendien maakt dit het soms mogelijk voorspellingen te geven. Vroeger (eind vorige eeuw) hield de statistiek zich alleen maar bezig met het samenvatten en overzichtelijk representeren van gegevens (b.v. bevolkingsstatistieken weergeven d.m.v. gemiddelden en grafieken). Statistiek was dus niet meer dan een beschrijving van een toestand. De modelmatige aanpak, die we in dit college zullen behandelen, is van recenter datum.

STOCHASTIEK. We kunnen een onderscheid maken tussen deterministische modellen en stochastische modellen. Deterministisch zijn b.v. de wetten van Newton (b.v. $F = m \cdot a$). Stochastische modellen hebben een bepaalde mate van onzekerheid ingebouwd. De reden kan gebrek aan gegevens zijn, maar vaak ziet men onzekerheid als inherent aan de natuur.

Vb. We kunnen nooit voorspellen hoe lang morgen de file op de A4 zal zijn, ook al ondervragen we iedereen, of ze al of niet met de auto gaan, waar naar toe en hoe laat, en gebruiken we het beste weerbericht.

Zelfs in het hypothetische geval dat we alles over de huidige toestand en het verleden weten, kunnen we niet zeggen wat er over één seconde gebeurt. Het is een filosofisch probleem om uit te maken of dit nu door onze onmacht komt of niet. In de natuurkunde heeft

men zolangzamerhand overeenstemming bereikt: de natuur is stochastisch van nature. Waarschijnlijkheidsrekening en statistiek houden zich bezig met stochastische modellen, zonder daar verder filosofisch over te doen. We merken nog op dat voor sommige problemen, die in principe deterministisch van aard zijn, het handig kan zijn er een kansmechanisme aan op te leggen.

Vb. Bij het zoeken naar het maximum van een grillige functie van twee variabelen, kan men systematisch de functiewaarde in een rooster van punten berekenen, maar het is soms makkelijker om volkomen onsystematisch te werk te gaan, d.w.z. de punten worden als bij “staartje prik” gekozen.

1. Waarschijnlijkheidsrekening

1.1. Empirische wet van de grote aantallen

DOBBELSPEL. In de 17-de eeuw is empirisch het volgende vastgesteld: als men $100\times$ met een dobbelsteen gooit vindt men in ongeveer $1/6$ van de gevallen een 6 (evenzo voor een ander aantal ogen). Dit heeft geleid tot de

HYPOTHESE: als men ∞ vaak met een zuivere slijtvaste dobbelsteen gooit vindt men in precies $1/6$ van de gevallen een 6 (evenzo voor een ander aantal ogen).

De hypothese is nooit te controleren, want (i) we kunnen niet oneindig vaak gooien (en zelfs al zou dat kunnen, een slijtvaste dobbelsteen bestaat niet) en (ii) het begrip *zuiver* is niet goed gedefinieerd. Men zou kunnen zeggen dat een zuivere dobbelsteen per definitie een slijtvaste dobbelsteen is die bij ∞ vaak gooien in precies $1/6$ van de gevallen een 6, in $1/6$ van de gevallen een 5, enz. geeft. Daarmee begeben we ons op abstract terrein.

TERMINOLOGIE. We spreken over een *experiment*, en de verzameling van alle mogelijke uitkomsten noemen we de *uitkomstenruimte* Γ . *Herhaalde* experimenten zijn verscheidene uitvoeringen van hetzelfde experiment. De herhaalde experimenten vormen tezamen weer een experiment met gecompliceerdere uitkomstenruimte. Bij herhaalde experimenten kan men spreken van de frequentie van een gebeurtenis. Dit is het aantal keren dat de gebeurtenis optreedt gedeeld door het aantal experimenten. Bij n experimenten waarbij de gebeurtenis S $n(S)$ keer optreedt is dus

$$f_q(S) = n(S)/n$$

de frequentie van gebeurtenis S .

Vb.

Experiment: gooien met een dobbelsteen

Uitkomstenruimte: $\{1, 2, 3, 4, 5, 6\}$

Herhaalde experimenten: $n\times$ gooien met een dobbelsteen

$n(\{6\})$ = het aantal keren dat 6 is gegooit

$f_q(\{6\}) = n(\{6\})/n$ = de frequentie van 6

Empirisch vastgesteld: $f_q(\{6\}) \approx 1/6$ als n groot.

Vb. Binaire getallen.

Een binair getal ω tussen 0 en 1 kan men schrijven als $\omega = 0.\omega_1\omega_2\omega_3\dots$ met $\omega_i \in \{0, 1\}$. Bekijk nu $f_q(\{1\}) :=$ de fractie énen in de eerste n digits. Het blijkt dat als ω een *willekeurig* gekozen getal tussen 0 en 1 is, dan $\lim_{n \rightarrow \infty} f_q(\{1\}) = 1/2$. Willekeurig gekozen betekent hier dat iedere digit wordt bepaald door het opgooien van een muntje. Dit komt overeen met het blindelings kiezen van een getal tussen 0 en 1.

1.2. Gebeurtenissen

Een verzameling bestaat formeel uit onderscheidbare elementen. Wij bekijken een speciale verzameling, n.l. de verzameling van alle mogelijke uitkomsten van een experiment (notatie: Γ): de uitkomstenruimte. We kunnen deelverzamelingen van Γ beschouwen, en vervolgens operaties uitvoeren, zoals doorsneden nemen e.d.. Zo komen we terecht bij de verzamelingstheorie.

DEFINITIE. Een *gebeurtenis* is een deelverzameling van Γ . We noemen een gebeurtenis ook wel een *eventualiteit* (Engels: *event*).

Vb. Het nemen van een sok uit een kast met rode en groene al of niet kapotte sokken.

$\Gamma = \{(r, k), (r, h), (g, k), (g, h)\}$ met r =rood, g =groen, k =kapot en h =heel. Dus b.v. (r, k) is de uitkomst van een rode kapotte sok. Een gebeurtenis is dan b.v. $A = \{(r, k), (r, h)\}$, d.w.z. een rode sok.

Vb. Het werpen met een dobbelsteen.

$\Gamma = \{1, 2, 3, 4, 5, 6\}$ en een gebeurtenis is b.v. $A = \{1, 3, 5\}$, een oneven getal.

OPMERKING. Een uitkomst is een deelverzameling van Γ bestaande uit maar één element.

VERZAMELINGENLEER. Op verzamelingen A en B kan men de volgende operaties uitvoeren: $A \cap B$: A door(snedes met) B . Dit is de verzameling van alle elementen die zowel in A als in B zitten. We zeggen ook wel dat gebeurtenissen A en B allebei optreden.

$A \cup B$: A verenigd met B . Dit is de verzameling van elementen die in A of in B zitten, of in beide. We zeggen ook wel dat gebeurtenis A of B optreedt.

\bar{A} : Het complement van A . Dit zijn alle elementen die niet in A zitten. We zeggen ook wel dat de gebeurtenis A niet optreedt.

Als $B \subset A$, d.w.z. B is een deelverzameling van A , dan zitten alle elementen van B ook in A .

Als $A \cap B = \emptyset$, de lege verzameling, dan hebben A en B geen elementen gemeen. We zeggen ook wel dat de gebeurtenissen A en B niet tegelijk kunnen optreden.

1.3. Axiomatische opzet

We hebben een algemeen kansbegrip nodig, dat ook toepasbaar is als de uitkomstenruimte oneindig groot is. Voorbeelden van zulke uitkomstenruimten zijn:

Vb. Het aantal keren dat men moet gooien met een dobbelsteen totdat voor de eerste keer een 6 wordt gevonden. Dan $\Gamma = \{1, 2, 3, \dots\}$.

Vb. Men springt in een zwembad en wil weten op welke plaats men dan terecht komt. Dan b.v. $\Gamma = \{ \text{alle punten in het zwembad} \}$.

AXIOMA'S. We noemen P een kans op de gebeurtenissen in Γ als

(1) $0 \leq P(A) \leq 1$ voor alle gebeurtenissen $A \subset \Gamma$,

(2) $P(\emptyset) = 0$,

(3) $P(\Gamma) = 1$,

(4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ voor alle gebeurtenissen $A, B \subset \Gamma$,

(5) Als A_1, A_2, \dots disjuncte gebeurtenissen zijn (d.w.z. de doorsnede van ieder tweetal is leeg), dan $P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$.

Een kans kan beschouwd worden als een maat voor de gebeurtenissen. We spreken dan ook vaak van een kansmaat.

OPMERKINGEN.

(1) Net als bij maten veronderstellen we dat kansen positieve getallen zijn. We stellen $P(A) \leq 1$, net als bij frequenties. Dit is niet meer dan een afspraak, we hadden b.v. ook $P(A) \leq 100$ kunnen nemen, zoals bij procenten (in feite doen we dat soms ook, n.l. als we het hebben over een kans van $x\%$).

(2) Het hoeft niet zo te zijn dat $P(A) = 0$ impliceert $A = \emptyset$; er kunnen ook andere onmogelijke gebeurtenissen zijn.

Vb. Het gooien met een dobbelsteen. Men mag $\Gamma = \{1, 2, 3, 4, 5, 6, 7\}$ nemen. Dan $P(\{7\}) = 0$.

(3) Ook hier hoeft de omkering niet te gelden. In het bovenstaande voorbeeld geldt b.v. $P(\{1, 2, 3, 4, 5, 6\}) = 1$ maar $\{1, 2, 3, 4, 5, 6\} \neq \Gamma = \{1, 2, 3, 4, 5, 6, 7\}$. Een ander voorbeeld is

Vb. Men trekt 3 sokken uit een kast met rode en groene sokken. Dan is $P(\text{een paar van één kleur}) = 1$.

(4) Dit is ook weer net als bij aantallen, frequenties en oppervlakten. Zo is, als A en B deelverzamelingen van het vlak zijn, de oppervlakten van $A \cup B$ gelijk aan de oppervlakte van A plus de oppervlakte van B min de oppervlakte van $A \cap B$, want als men A en B in het vlak legt komt er op het gebied $A \cap B$ een dubbele laag.

(5) Met deze eis zullen we in dit college niet veel te maken krijgen. We geven één voorbeeld.

Vb. Stel $\Gamma = (0, 1]$ en $P(A)$ = de lengte van A . Noem nu $A_1 = (1/2, 1]$, $A_2 = (1/4, 1/2]$, $A_3 = (1/8, 1/4]$, enz..

Dan $P(A_1) = 1/2$, $P(A_2) = 1/4$, $P(A_3) = 1/8$, enz.. Verder zijn de intervallen A_1, A_2, \dots disjunct en $A_1 \cup A_2 \cup \dots = (0, 1] = \Gamma$, dus $P(A_1 \cup A_2 \cup \dots) = P(\Gamma) = 1$. Er moet dus volgens (5) gelden dat $1 = 1/2 + 1/4 + 1/8 + \dots$. Dit is inderdaad het geval.

VOORBEELD VAN EEN KANS. Beschouw een verzameling van N elementen $\{a_1, \dots, a_N\}$, waarvan er R kenmerk S bezitten. Definieer

$$P(a_i) = 1/N, \quad i = 1, \dots, N.$$

Dan geldt $P(S) = R/N$. In woorden: als men *aselect* een element kiest, dan is de kans op kenmerk S gelijk aan R/N .

Vb. Beschouw een kast met 10 rode en 20 groene sokken. Trek er *aselect* één uit. Dan is $\Gamma = \{r_1, \dots, r_{10}, g_1, \dots, g_{20}\}$, zeg. Bekijk nu het kenmerk $S =$ een rode sok. Dan $P(S) = 10/30$.

We hebben nu een uitkomstenruimte Γ en een kansmaat P op deelverzamelingen van Γ , met de rekenregels (1) t/m (5). M.b.v. deze axioma's kan men verdere rekenregels afleiden. Belangrijke gevolgen zijn:

- Als $A \cap B = \emptyset$, dan $P(A \cup B) = P(A) + P(B)$,
- $A \cap \bar{A} = \emptyset$ en $A \cup \bar{A} = \Gamma$, dus $P(A) + P(\bar{A}) = 1$,
- Als $B \subset A$, dan $P(A) = P(B) + P(A \cap \bar{B}) \geq P(B)$,
- $A \cap B \subset A \subset A \cup B$ dus $P(A \cap B) \leq P(A) \leq P(A \cup B)$.

VOORWAARDELIJKE KANS.

DEFINITIE. Als $P(B) \neq 0$, dan is de *voorwaardelijke kans op A gegeven B*:

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

Voorwaardelijke kansen voldoen ook aan axioma's (1) t/m (5). Door voorwaardelijke kansen gegeven B te nemen maakt men in feite een overgang naar een nieuwe uitkomstenruimte met nieuwe kansen erop (b.v. alleen uitkomsten in B kunnen nog positieve kans hebben).

Vb. Men gooit drie keer met een zuivere munt. Dan

$$\Gamma = \left\{ \begin{array}{l} (M, M, M), (M, M, K), (M, K, M), (M, K, K), \\ (K, M, M), (K, M, K), (K, K, M), (K, K, K) \end{array} \right\}$$

en de kans op iedere uitkomst is $1/8$. Wat is nu de kans op minstens $1 \times$ kruis gegeven minstens $2 \times$ munt? Neem $B = \{(M, M, K), (M, K, M), (K, M, M), (M, M, M)\}$. Dus $P(B) = 4/8$. Als A minstens $1 \times$ kruis is, dan $A \cap B = \{(M, M, K), (M, K, M), (K, M, M)\}$. Dus $P(A \cap B) = 3/8$ en $P(A|B) = 3/4$.

Soms zijn alleen voorwaardelijke kansen gegeven. De onvoorwaardelijke kansen kan men dan terugvinden m.b.v. de eigenschap hieronder. Eerst hebben we een definitie nodig.

DEFINITIE.

B_1, \dots, B_k heet een *partitie* van Γ als B_1, \dots, B_k disjunct zijn en $B_1 \cup \dots \cup B_k = \Gamma$.

EIGENSCHAP. Als B_1, \dots, B_k een partitie van Γ vormt, dan voor iedere gebeurtenis A ,

$$P(A) = \sum_{i=1}^k P(A \cap B_i) = \sum_{P(B_i) \neq 0} \frac{P(A \cap B_i)}{P(B_i)} P(B_i) = \sum_{P(B_i) \neq 0} P(A|B_i)P(B_i).$$

Vb. Twee vrienden J en K worden gedwongen te kiezen uit 3 chocolaatjes, waarvan er één vergiftigd is. Het gekozen chocolaatje dient meteen genuttigd te worden. Stel dat J eerst kiest. Dan $P(J\text{overleeft}) = 2/3$. Verder $P(K\text{overleeft}|J\text{overleeft}) = 1/2$ en $P(K\text{overleeft}|J\text{overleeft niet}) = 1$. Hieruit volgt dat

$$\begin{aligned} P(K\text{overleeft}) &= P(K\text{overleeft}|J\text{overleeft})P(J\text{overleeft}) \\ &\quad + P(K\text{overleeft}|J\text{overleeft niet})P(J\text{overleeft niet}) \\ &= \frac{1}{2} \frac{2}{3} + 1 \frac{1}{3} = \frac{2}{3}. \end{aligned}$$

M.a.w. K heeft dezelfde kans om te overleven als J .

1.4. Combinatoriek

Stel dat men de kans op een gebeurtenis A wil weten, bijvoorbeeld bij het aselekt kiezen uit een verzameling van N elementen. Dan is het van belang te weten hoeveel uitkomsten er in A zitten. Daarbij is enige kennis van de combinatoriek goed bruikbaar. Om na te gaan hoeveel mogelijkheden er zijn is het soms te doen om alle mogelijkheden gewoon uit te schrijven.

Vb. Op een menukaart staan 3 soorten soep, 2 hoofdgerechten, en 3 toetjes. Het aantal manieren waarop men dan een menu kan samenstellen is $3 \times 2 \times 3 = 18$.

Vb. Men gooit $6 \times$ met een dobbelsteen. Het aantal mogelijke uitkomsten is $6 \times 6 \times 6 \times 6 \times 6 \times 6 = 46656$. Het aantal mogelijke uitkomsten met voor iedere worp een ander aantal ogen is $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$. De kans dat alle aantallen ogen verschillend zijn bij $6 \times$ gooien met een dobbelsteen is dus $720/46656 = 0.015432$. We gebruiken hier: de kans op kenmerk S is gelijk aan het aantal mogelijke uitkomsten met kenmerk S gedeeld door het totaal aantal mogelijke uitkomsten.

REGELS.

(A)

Vb. Het aantal telefoonnummers van 4 cijfers is $10 \times 10 \times 10 \times 10 = 10\,000 = 10^4$.

Algemeen: Het aantal rijtjes van lengte k van n symbolen is n^k .

(B)

Vb. Het aantal manieren om een voorkeursrangschikking van 5 soorten koffie te geven is $5 \times 4 \times 3 \times 2(\times 1) = 5! = 120$.

Algemeen: Het aantal manieren om n symbolen te rangschikken is $n \times (n-1) \times (n-2) \times \dots \times 3 \times 2(\times 1) = n!$ (spreek uit: n faculteit).

(C)

Vb. Het aantal telefoonnummers van 4 verschillende cijfers is $10 \times 9 \times 8 \times 7 = 5040$.

Algemeen: Het aantal rijtjes van lengte k van n symbolen zodanig dat niet twee keer dezelfde optreedt is $n!/(n-k)!$. We definiëren $0! = 1$ (dus voor het geval $n = k$ zijn we terug in situatie (B)).

(D)

Vb. Het aantal manieren om van 5 koffiemarken er 3 te kiezen is als volgt. We hebben $5 \times 4 \times 3 = 60$ geordende rijtjes (d.w.z. rijtjes waarbij we voor de drie gekozen koffiemarken ook een voorkeursrangschikking geven). Er zijn $3!$ mogelijke ordeningen van een rijtje van 3 koffiemarken. Dus het aantal manieren om er 3 te kiezen, zonder voorkeursrangschikking is $60/3! = 10$.

Algemeen: Het aantal manieren om uit n symbolen er k te kiezen is $\binom{n}{k} = n!/(k!(n-k)!)$ (spreek uit: n boven k). Men noemt $\binom{n}{k}$ een *binomiaal coëfficiënt*. Het verschil met (C) is dat we niet op de ordening letten. Merk op dat het aantal manieren om er k te kiezen gelijk is aan het aantal manieren om er $(n-k)$ (niet) te kiezen, d.w.z. $\binom{n}{k} = \binom{n}{n-k}$.

Vb. Van n symbolen kan men $n(n-1)$ geordende paren vormen, en $n(n-1)/2 = \binom{n}{2}$ ongeordende paren.

EIGENSCHAPPEN EN TOEPASSINGEN VAN BINOMIAAL COËFFICIËNTEN.

(1) De *driehoek van Pascal* is

$$\begin{array}{ccccccc} & & & & & & 1 \\ & & & & & & 1 & 1 \\ & & & & & 1 & 2 & 1 \\ & & & 1 & 3 & 3 & 1 \\ & 1 & 4 & 6 & 4 & 1 \\ & & & & & & & \dots \end{array}$$

Op de $(n+1)$ -ste rij van de driehoek vindt men de binomiaal coëfficiënten

$$\binom{n}{0}, \binom{n}{1}, \dots, \binom{n}{k}, \dots, \binom{n}{n-1}, \binom{n}{n}.$$

(2) Er geldt:

$$\binom{n}{0} = \binom{n}{n} = 1,$$

en

$$\binom{n}{1} = \binom{n}{n-1} = n,$$

en de symmetrie $\binom{n}{k} = \binom{n}{n-k}$. Verder ziet men aan de driehoek van Pascal dat

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}.$$

(3) Het *binomium van Newton* is de formule

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}.$$

Vb. $(a+b)^2 = a^2 + 2ab + b^2$ en $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$.

Vb. $2^n = (1+1)^n = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n}$, d.w.z. het aantal rijtjes van 0'en en 1'en is gelijk aan het aantal rijtjes met alleen 0'en + het aantal rijtjes met één 1 en verder alleen 0'en + ... + het aantal rijtjes met alleen 1'en.

(4) Bij het n keer gooien met een munt bezit het aantal keren dat men kruis werpt een z.g. *binomiale verdeling*. Noem p de kans op kruis bij één keer gooien. Als $p = 1/2$, dan is de munt zuiver. Bij n keer gooien is de kans op een geordend rijtje met (precies) k keer kruis gelijk aan $p^k(1-p)^{n-k}$ (b.v. als $n = 3$ dan is de kans op (K, K, M) gelijk aan $p^2(1-p)$). Het aantal rijtjes met k keer kruis is $\binom{n}{k}$ (b.v. als $n = 3$ dan komt kruis 2 keer voor in de rijtjes $(K, K, M), (K, M, K), (M, K, K)$, en het aantal is dus $3 = \binom{3}{2}$). We vinden zo:

$$P(\text{(precies) } k \times \text{ kruis}) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n.$$

De uitkomstenruimte is hier $\Gamma = \{0, 1, \dots, n\}$. Omdat $P(\Gamma) = 1$ moet dus gelden:

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Dat dit inderdaad klopt, volgt b.v. uit het binomium van Newton.

(5) Steekproefcontrole.

Als voorbeeld bekijken we een partij van N chips, waarvan een onbekend aantal, zeg R , kapot is. Definieer $p = R/N$. Dus p is de fractie kapotte chips in de partij. We willen nu iets te weten komen over p , maar het is teveel werk om alle chips in de partij te controleren. We nemen daarom slechts een steekproef van n chips. Dit kan op twee manieren:

(a) Steekproef *met* teruglegging.

Trek n keer aselekt een chip, noteer of deze chip functioneert, en leg de getrokken chip vervolgens weer terug in de partij. De kans op een kapotte chip bij één keer aselekt trekken is dan p . Dus bij n keer trekken is het aantal kapotte chips in de steekproef binomiaal verdeeld:

$$P(k \text{ kapotte chips in de steekproef}) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \dots, n, \quad p = R/N.$$

(b) Steekproef *zonder* teruglegging.

Trek n keer aselect een chip, en leg deze apart (we veronderstellen hier dat $n \leq N$). Het aantal manieren waarop men n elementen uit N kan kiezen is $\binom{N}{n}$. Het aantal manieren om k elementen te kiezen uit R , en $n - k$ uit de overige $N - R$ is $\binom{R}{k} \binom{N-R}{n-k}$. Dus

$$P(k \text{ kapotte chips in de steekproef}) = \frac{\binom{R}{k} \binom{N-R}{n-k}}{\binom{N}{n}}.$$

Dit geldt voor $0 \leq k \leq \min(n, R)$, en $0 \leq n - k \leq \min(n, N - R)$. We noemen dit de *hypergeometrische verdeling*.

Volgens de empirische wet van de grote aantallen geldt zowel in geval (a) als in geval (b) (met N groot), dat als n groot is de fractie kapotte chips in de steekproef wel ongeveer gelijk zal zijn aan de fractie kapotte chips in de partij. In die zin geeft de steekproef dus informatie over de onbekende fractie p .

1.5. Stochastische onafhankelijkheid

DEFINITIE. Twee gebeurtenissen A en B heten *onderling onafhankelijk* (afgekort: *o.o.*) als $P(A \cap B) = P(A)P(B)$.

EIGENSCHAPPEN.

-Als $P(B) = 0$, dan $P(A \cap B) = P(A)P(B) = 0$, dus dan zijn A en B o.o. voor iedere A .

-Als $P(B) \neq 0$, dan zijn A en B dan en slechts dan o.o. als $P(A|B) = P(A)$.

-Als A en B o.o. èn A en B zijn disjunct, dan

$$0 = P(A \cap B) = P(A)P(B),$$

dus dan $P(A) = 0$ of $P(B) = 0$.

DEFINITIE. Een rij van gebeurtenissen A_1, \dots, A_k heet *paarsgewijs onafhankelijk* als ieder tweetal o.o. is. De rij heet o.o. als voor iedere keuze van $J \subset \{1, \dots, k\}$ geldt

$$P(\cap_{j \in J} A_j) = \prod_{j \in J} P(A_j).$$

Een drietal van gebeurtenissen A_1, A_2, A_3 is dus paarsgewijs onafhankelijk als geldt:

$P(A_1 \cap A_2) = P(A_1)P(A_2)$, $P(A_1 \cap A_3) = P(A_1)P(A_3)$ en $P(A_2 \cap A_3) = P(A_2)P(A_3)$.

Ze zijn o.o. als bovendien geldt: $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$.

We merken op dat onderlinge onafhankelijkheid paarsgewijze onafhankelijkheid impliceert, maar dat het omgekeerde niet hoeft te gelden.

VOORBEELDEN.

(1) In Zeeland is een bouwwerk gemaakt bestaande uit 60 pijlers, die bij storm neergelaten kunnen worden zodat ze een dam vormen. De kans dat één zo'n pijler functioneert op het moment dat de dam in werking wordt gezet is vrij groot, ongeveer 95 %. De pijlers functioneren op onderling onafhankelijke wijze. Als één pijler het niet doet ontstaat er door de sterke stroming een enorm gat in de dam, zodat er toch overstromingen zullen

zijn. Het is dus van belang dat alle 60 pijlers goed functioneren. De kans hierop is echter ongeveer $(0.95)^{60} < 0.05!$

(2) Om de veiligheid van een kerncentrale te vergroten, bouwt men diverse veiligheidsmechanismen in. Slechts als àl deze mechanismen haperen kan er een kernramp gebeuren. Men zegt nu dat de kans op een kernramp erg klein is omdat het wel toevallig zou zijn als alle veiligheidsvoorzorgen tegelijkertijd het laten afweten. Vaak is impliciet in deze redenering, de veronderstelling dat de veiligheidsmechanismen o.o. zijn. Immers, dan is de kans dat àlle veiligheidsmechanismen niet werken gelijk aan het product van de kansen dat één veiligheidsmechanisme niet werkt. Deze kans is dan kleiner naarmate er meer veiligheidsmechanismen zijn. Bij een risico-analyse is het daarom van groot belang om na te gaan of de veronderstelling van onderlinge onafhankelijkheid wel klopt.

(3) Dit is een voorbeeld van gebeurtenissen wel paarsgewijs onafhankelijk zijn, maar niet o.o.. Neem een vierkantige zuivere dobbelsteen.

De uitkomstenruimte is $\Gamma = \{1, 2, 3, 4\}$ en $P(\{k\}) = 1/4$, $k = 1, \dots, 4$. Beschouw de gebeurtenissen $A_1 = \{1, 2\}$, $A_2 = \{1, 3\}$ en $A_3 = \{1, 4\}$. Dan $P(A_1) = P(A_2) = P(A_3) = 1/2$ en $P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = 1/4$, dus A_1 , A_2 en A_3 zijn paarsgewijs onafhankelijk. Maar $P(A_1 \cap A_2 \cap A_3) = 1/4 \neq 1/8 = P(A_1)P(A_2)P(A_3)$, dus A_1 , A_2 en A_3 zijn niet o.o..

1.6. Urnmodellen

In voorbeeld (5) van Paragraaf 1.4 merkten we al op dat er twee methoden zijn voor het nemen van een steekproef uit een eindige populatie, n.l. *met* terugleggen en *zonder* terugleggen. Het zal duidelijk zijn dat de individuele trekkingen in het laatste geval niet o.o. zijn. Immers, als b.v. op een bepaald moment alle elementen met kenmerk S al getrokken zijn, dan is het niet langer mogelijk S nog terug te vinden in nieuwe trekkingen. In deze paragraaf gaan we hier wat dieper op in. De situatie in voorbeeld (5) is net als bij het trekken van knikkers uit een vaas met twee kleuren knikkers, zeg blauwe en groene. Vandaar dat men spreekt over *urnmodellen*.

We citeren het dictaat STATISIEK VOOR STUDENTEN CIVIELE TECHNIEK (A106, Deel 1), (P. GROENEBOOM (Delft, 1991)): ‘*In traditionele teksten over kansrekening en statistiek (van vóór de “computerrevolutie”), moet men vaak 100 of meer bladzijden vol dobbelstenen en gooien van ballen in dozen doorworstelen voordat het begrip “continue dichtheid” aan de orde komt.*’ Ons dictaat heeft dan ook wat traditionele trekjes, maar de 100 bladzijden worden niet gehaald. We bevestigen hier wel dat de belangrijkste theorie pas behandeld wordt in de volgende paragraaf, zodat het niet aan te raden is om veel met dobbelstenen, ballen en knikkers te blijven spelen.

We bekijken een vaas met N knikkers, waarvan er R blauw zijn en $N - R$ groen. Bij één trekking is de uitkomstenruimte $\{b_1, \dots, b_R, g_1, \dots, g_{N-R}\}$. De kans op een element uit deze uitkomstenruimte is $1/N$ voor alle elementen. Bij n trekkingen noemen we B_i de gebeurtenis dat er bij de i -de trekking een blauwe knikker wordt gevonden, en G_i de gebeurtenis van een groene knikker $i = 1, \dots, n$. Beschouw nu het resultaat van n aselechte trekkingen. Aselect wil zeggen dat de kans op een gebeurtenis in de uitkomstenruimte van alle trekkingen tezamen, gelijk is aan het aantal manieren om die gebeurtenis voor elkaar te krijgen gedeeld door het totaal aantal mogelijkheden van het samengestelde experiment.

(a) *Met* terugleggen.

Wat is b.v. de kans dat de i -de trekking een blauwe knikker oplevert? Het totaal aantal mogelijke uitkomsten van n trekkingen met terugleggen is N^n (zie 1.4(A)). Het aantal manieren om een rijtje van n knikkers te vormen, zodanig dat de i -de knikker blauw is, is $N^{n-1}R$ want voor iedere plaats hebben we N mogelijkheden, behalve voor de i -de plaats, waar er R mogelijkheden zijn. Dus we vinden:

$$P(B_i) = \frac{N^{n-1}R}{N^n} = \frac{R}{N}.$$

Er volgt dan ook dat $P(G_i) = 1 - P(B_i) = (N - R)/N$. We kunnen zo de kansen op alle mogelijke gebeurtenissen bepalen. B.v. het aantal rijtjes van knikkers zo dat de eerste blauw is, de tweede groen en de derde weer blauw, is $R(N - R)RN^{n-3}$, dus

$$P(B_1 \cap G_2 \cap B_3) = \frac{R(N - R)RN^{n-3}}{N^n} = \left(\frac{R}{N}\right)^2 \frac{N - R}{N}.$$

Maar ook

$$P(B_1)P(G_2)P(B_3) = \frac{R}{N} \frac{N - R}{N} \frac{R}{N} = \left(\frac{R}{N}\right)^2 \frac{N - R}{N},$$

enz.. We zien zo dat de gebeurtenis van een blauwe of groene knikker in de trekkingen o.o. zijn. We zeggen ook wel dat de n trekkingen (n experimenten) o.o. zijn. Verder geldt

$$P(k \text{ blauwe knikkers}) = \binom{n}{k} \left(\frac{R}{N}\right)^k \left(\frac{N-R}{N}\right)^{n-k}, \quad k = 1, \dots, n.$$

(b) *Zonder terugleggen.*

Wat is $P(B_i)$ in dit geval? Het aantal rijtjes van lengte n van N verschillende elementen is $N!/(N-n)!$ (zie 1.4(C)). Dit is het totaal aantal mogelijke uitkomsten van n trekkingen zonder terugleggen. Het aantal rijtjes hiervan met op de i -de plaats een blauwe knikker is het aantal rijtjes van lengte $n-1$ van $N-1$ verschillende elementen maal het aantal mogelijkheden om op de i -de plaats een blauwe knikker te verkrijgen. Dus dit is $(N-1)!/((N-1)-(n-1))! \times R = (N-1)!/(N-n)! \times R$. Zo vinden we dat

$$P(B_i) = \frac{(N-1)!/(N-n)! \times R}{N!/(N-n)!} = \frac{R}{N},$$

net als in (a). Een andere manier om dit in te zien is als volgt. We hebben

$$P(B_1) = \frac{R}{N},$$

net als in (a), omdat bij de eerste trekking het al of niet terugleggen nog geen rol speelt. Verder geldt

$$\begin{aligned} P(B_2) &= P(B_1 \cap B_2) + P(G_1 \cap B_2) \\ &= \frac{R}{N} \frac{R-1}{N-1} + \frac{N-R}{N} \frac{R}{N-1} = \frac{R}{N}, \end{aligned}$$

enz..

Het is eenvoudig in te zien dat de gebeurtenissen niet o.o. zijn. Immers, neem b.v. B_1 en B_2 . Dan $P(B_2|B_1) = (R-1)/(N-1)$ want als de eerste knikker blauw is, dan zijn er nog maar $(R-1)$ blauwe knikkers over in de vaas met $(N-1)$ knikkers. Maar zoals we hebben gezien is $P(B_2) = R/N$, dus $P(B_2|B_1) \neq P(B_2)$. Dus B_1 en B_2 zijn niet o.o.. Daarom zijn B_1, \dots, B_n niet paarsgewijs onafhankelijk, dus zeker niet o.o.. De verdeling van het aantal blauwe knikkers in de steekproef is de hypergeometrische verdeling:

$$P(k \text{ blauwe knikkers}) = \frac{\binom{R}{k} \binom{N-R}{n-k}}{\binom{N}{n}}, \quad k \in \{\max(0, n - (N - R)), \dots, \min(n, R)\}.$$

Samenvattend: het aantal elementen met een zeker kenmerk S bezit een binomiale verdeling als het gaat om een steekproef met terugleggen, en een hypergeometrische verdeling bij een steekproef zonder terugleggen. Als N en R groot zijn t.o.v. n , zal het echter niet zoveel meer uitmaken of men al dan niet teruglegt. Dit kan wiskundig ook bewezen worden. Noem $P_B(k)$ de kans op k bij de binomiale verdeling en $P_H(k)$ de kans op k bij de hypergeometrische verdeling.

Lemma.

$$\lim_{N-R \rightarrow \infty, R \rightarrow \infty} \frac{P_H(k)}{P_B(k)} = 1.$$

BEWIJS.

$$\begin{aligned} \frac{P_H(k)}{P_B(k)} &= \left(\frac{R!(N-R)!n!(N-n)!}{k!(R-k)!((N-R)-(n-k))!(n-k)!N!} \right) / \left(\frac{n!}{k!(n-k)!} \frac{R^k (N-R)^{n-k}}{N^{n-k}} \right) \\ &= \frac{R!}{(R-k)!R^k} \frac{(N-R)!}{((N-R)-(n-k))!(N-R)^{n-k}} \frac{(N-n)!N^n}{N!}. \end{aligned}$$

Er geldt

$$\frac{N!}{(N-n)!} = N \times (N-1) \times \dots \times (N-n+1) \approx N \times N \times \dots \times N = N^n.$$

Analoog voor de andere termen:

$$\frac{R!}{(R-k)!} \approx R^k, \quad \frac{(N-R)!}{((N-R)-(n-k))!} \approx (N-R)^{n-k}.$$

Vul dit in en het lemma is bewezen. \square

Als getallenvoorbeeld nemen we het geval $p = 1/2$ en $n = 6$. Bij de hypergeometrische verdeling nemen we eerst $N = 10$, $R = 5$ ($R/N = p = 1/2$) en dan $N = 100$, $R = 50$ ($R/N = p = 1/2$). Het is duidelijk te zien dat voor de grotere waarden van N en R de hypergeometrische verdeling goed benaderd wordt door de binomiale verdeling.

k	0	1	2	3	4	5	6			
$P_B(k)$, $p = 1/2$, $n = 6$				0.015	0.094	0.234	0.312	0.234	0.094	0.015
$P_H(k)$, $R/N = 1/2$, $n = 6$, $N = 10$	0.000	0.024	0.238	0.476	0.238	0.024	0.000			
$P_H(k)$, $R/N = 1/2$, $n = 6$, $N = 100$	0.013	0.089	0.237	0.322	0.237	0.089	0.013			

1.7. Stochastische grootheden

Een *stochastische grootheid* beschrijft de uitkomst van een experiment. We gebruiken de afkorting *s.g.*. Stochastische grootheden worden meestal met hoofdletters (X, Y , etc.) aangegeven. (Vanaf Paragraaf 2.3 zullen we een notatie met hoofdletters en vetgedrukte letters door elkaar gaan gebruiken).

Men kan een *s.g.* beschouwen als een codering van de uitkomsten van een experiment.

Vb. Experiment: het één keer gooien met een munt.

$\Gamma = \{\text{kruis, munt}\}$

Beschouw nu de *s.g.* $X \in \{0, 1\}$. De gebeurtenis $\{X = 0\}$ laten we b.v. corresponderen met $\{\text{kruis}\}$ en $\{X = 1\}$ met $\{\text{munt}\}$.

Vb. Experiment: het uit laten voeren van een programma.

X is dan b.v. de executietijd.

Vb. Experiment: het n keer aselekt trekken uit een vaas met blauwe en groene knikkers. Laat X_i de uitslag van de i -de trekking zijn, waarbij $\{X_i = 1\}$ correspondeert met een blauwe knikker, en $\{X_i = 0\}$ met een groene knikker. Dan is $X = \sum_{i=1}^n X_i$ het aantal blauwe knikkers in de steekproef.

Als X een stochastische grootheid is, dan geldt altijd $X \in \mathbf{R}$, de reële getallen. Soms is dat natuurlijk, b.v. als X de executietijd van een programma is, soms is het echter een codering. Antwoorden $\{ja, nee\}$ op een vraag kan men met $\{1, 0\}$ coderen. Gebeurtenissen zijn nu van de vorm $\{X \in A\}$ met $A \subset \mathbf{R}$.

Er is een onderscheid te maken tussen discrete s.g.ⁿ en continue s.g.ⁿ, en iets daartussen in. Het laatste geval zullen we in dit college niet tegenkomen.

VOORBEELDEN.

-Het aantal functionerende verbindingen in een electriciteitscircuit is een discrete s.g..

-Het aantal ogen bij het gooien van een dobbelsteen is een discrete s.g..

-De executietijd van een programma is een continue s.g..

-Het dioxinegehalte in melk is een continue s.g..

1.7.1. Discreet verdeelde stochastische grootheden

DEFINITIE. X is een *discrete* s.g. als X maar eindig of aftelbaar veel waarden kan aannemen.

Vb. Laat X het aantal keren zijn, dat men een dobbelsteen moet gooien om een 6 te vinden. Dan $X \in \{1, 2, 3, \dots\}$. Dit is dus een discrete s.g. die aftelbaar veel waarden kan aannemen.

DISCRETE VERDELING. Stel X is een discrete s.g. met waarden in $\{x_1, x_2, \dots\}$. (Een speciaal geval is de situatie waar X maar eindig veel waarden kan aannemen.) Definieer

$$p_k = P(X = x_k), \quad k = 1, 2, \dots$$

Dus p_k is de kans op uitkomst x_k . Als p_1, p_2, \dots gegeven zijn, kan men de kans op iedere gebeurtenis uitrekenen. We zeggen dat p_1, p_2, \dots de *verdeling* van X beschrijven. De s.g. x legt kansmassa p_k op x_k , $k = 1, 2, \dots$

De verdelingsfunctie van X is gedefinieerd als

$$F(x) = \sum_{x_k \leq x} p_k = P(X \leq x).$$

F is dus een functie van \mathbf{R} naar $[0, 1] \subset \mathbf{R}$. Als de verdelingsfunctie $F(x)$ gegeven is voor alle x , dan zijn ook de kansen p_k terug te vinden. De verdelingsfunctie is daarom een volledige beschrijving van de verdeling van X .

EIGENSCHAPPEN.

(i) $0 \leq p_k \leq 1$ en $\sum_{k=1}^{\infty} p_k = 1$,

- (ii) $0 \leq F(x) \leq 1$ en $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow \infty} F(x) = 1$,
- (iii) $F(x)$ is een stijgende (d.w.z. niet-dalende) functie,
- (iv) $F(x)$ springt p_k omhoog bij x_k , $k = 1, 2, \dots$ en is constant tussen de sprongpunten.

VOORBEELDEN VAN DISCRETE VERDELINGEN.

(1) Ontaarde verdeling (gedegenereerde verdeling).

X bezit een *ontaarde* verdeling als X maar één waarde kan aannemen. d.w.z. als voor zeker getal x_0 geldt $P(X = x_0) = 1$. De verdelingsfunctie $F(x)$ is dan constant gelijk aan nul voor $x < x_0$ en constant gelijk aan één voor $x \geq x_0$.

Vb. Trek 3 sokken uit een kast met rode en groene sokken. Noem X het aantal paren van één kleur dat gevonden wordt. Dan $P(X = 1) = 1$ en dus ook $P(X \neq 1) = 0$.

(2) Alternatieve verdeling met parameter p .

X bezit een *alternatieve* verdeling als X slechts 2 waarden kan aannemen, zeg $X \in \{x_1, x_2\}$. Noem $P(X = x_1) = p$. Dan $P(X = x_2) = 1 - p$. Als (zonder verlies van algemeenheid) $x_1 < x_2$, dan springt de verdelingsfunctie een afstand p bij x_1 en springt verder naar 1 bij x_2 (de tweede spronggrootte is dus $1 - p$).

(3) Binomiale verdeling met parameters n en p .

De binomiale verdeling zijn we al tegengekomen in Paragraaf 1.4 en 1.6. X bezit een binomiale verdeling als

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n,$$

waarbij n en p parameters zijn.

Vb. Bekijk n onafhankelijk werkende computerprogramma's. Noem $\{X_i = 1\}$ de gebeurtenis dat het i -de programma succesvol gedraaid heeft, en $\{X_i = 0\}$ de gebeurtenis dat het i -de programma op een mislukking uitdraait. Noem $p = P(X_i = 1)$ de succeskans. We veronderstellen dat p voor alle programma's hetzelfde is. Nu is $X = \sum_{i=1}^n X_i$ het aantal succesvolle programma's, en X bezit een binomiale verdeling met parameters n en p .

(4) Hypergeometrische verdeling.

Ook de hypergeometrische verdeling zijn we al eerder tegengekomen. X bezit een hypergeometrische verdeling als

$$P(X = k) = \frac{\binom{R}{k} \binom{N-R}{n-k}}{\binom{N}{n}}, \quad k = \max(0, n - (N - R)), \dots, \min(n, R).$$

(5) Poissonverdeling met parameter μ .

X bezit een *Poisson*verdeling met parameter $\mu > 0$ als

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu} := p_k, \quad k = 0, 1, \dots$$

Er moet gelden dat $\sum_{k=0}^{\infty} p_k = 1$. Dat dit inderdaad het geval is kan men inzien door e^μ in een Taylorreeks te ontwikkelen rond $\mu = 0$:

$$e^\mu = 1 + \mu + \frac{\mu^2}{2} + \frac{\mu^3}{3!} + \dots$$

Hoe komt men nu op een dergelijke verdeling? De interpretatie zullen we aan de hand van een voorbeeld trachten duidelijk te maken.

Vb. Stel dat X het aantal logins op een mainframe gedurende tijdsperiode $[0, T]$ is. We willen de verdeling van X weten. Daartoe verdelen we het interval $[0, T]$ in n kleine deelintervalletjes van lengte T/n .

Definieer X_i = het aantal logins in intervalletje i . Stel

(a) De kans op één login in intervalletje i is ongeveer evenredig met de lengte van dat intervalletje: $P(X_i = 1) \approx \lambda T/n$. Hier is λ de evenredigheidsconstante.

(b) De kans op meer dan één login in een klein intervalletje is ongeveer nul: $P(X_i > 1) \approx 0$.

(c) Het aantal logins in een klein intervalletje is onafhankelijk van het aantal logins in een ander intervalletje.

Nu is $X = \sum_{i=1}^n X_i$. De bovenstaande veronderstellingen zeggen dat X_i ongeveer alternatief verdeeld is met parameter $p = \lambda T/n$. De onafhankelijkheidsveronderstelling (c) geeft dan

$$P(X = k) \approx \binom{n}{k} (\lambda T/n)^k (1 - \lambda T/n)^{n-k}, \quad k = 1, \dots, n$$

(zie ook voorbeeld (3) in deze paragraaf). Herschrijven geeft

$$P(X = k) \approx \frac{n!}{(n-k)!n^k} \frac{(\lambda T)^k}{k!} \left(1 - \frac{\lambda T}{n}\right)^{n-k}.$$

Er geldt

$$\lim_{n \rightarrow \infty} \frac{n!}{(n-k)!n^k} = 1$$

(zie ook het bewijs van het lemma in de vorige paragraaf). Verder,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda T}{n}\right)^{n-k} \\ &= \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda T}{n}\right)^n = e^{-\lambda T}. \end{aligned}$$

Dus

$$P(X = k) = \lim_{n \rightarrow \infty} \frac{n!}{(n-k)!n^k} \frac{(\lambda T)^k}{k!} \left(1 - \frac{\lambda T}{n}\right)^{n-k} = \frac{(\lambda T)^k}{k!} e^{-\lambda T}, k = 0, 1, \dots$$

M.a.w. X is Poisson verdeeld met parameter $\mu = \lambda T$. We noemen μ de intensiteit. Als μ groot is betekent dat dat het druk is bij het mainframe.

(6) Negatief binomiale verdeling met parameters k en p .

Vb. Stel X is het aantal keren dat een computerprogramma gedraaid heeft totdat het voor de eerste keer fout liep. Noem

$$Y_i = \begin{cases} 1, & \text{als het bij de } i\text{-de keer draaien fout loopt;} \\ 0, & \text{als het bij de } i\text{-de keer draaien goed gaat,} \end{cases}$$

en zij $p = P(Y_i = 1)$. Dan, onder de aanname dat het al of niet fout lopen voor de individuele executies o.o. zijn,

$$P(X = n) = P(Y_1 = Y_2 = \dots = Y_{n-1} = 0, Y_n = 1) = (1-p)^{n-1}p, n = 1, 2, \dots$$

Dit noemt men de *geometrische verdeling*. De geometrische verdeling is een speciaal geval van de *negatief binomiale verdeling*. De laatste krijgt men, als men naar de verdeling kijkt van de s.g. \tilde{X} , de wachttijd tot het voor de k -de keer fout loopt. Dan

$$\begin{aligned} P(\tilde{X} = n) &= P\left(\sum_{i=1}^{n-1} Y_i = k-1, Y_n = 1\right) \\ &= \binom{n-1}{k-1} p^k (1-p)^{n-k}, n = k, k+1, \dots \end{aligned}$$

Voor $k = 1$ is dit de geometrische verdeling.

1.7.2. Continu verdeelde stochastische grootheden

DEFINITIE. X is een *continue* s.g. als X alle waarden in een zeker interval kan aannemen.

Vb. X = de executietijd van een programma, kan alle waarden > 0 aannemen.

Met een continue s.g. associëren we een *dichtheid* $f(x)$, die de *aannemelijkheid* van de waarde x aangeeft.

Vb. X = de lichaamstemperatuur van een mens. Dit is een continue s.g. die alle waarden tussen 35° en 42° kan aannemen. (Lagere of hogere waarden zijn misschien ook wel mogelijk, maar dan hebben we in ieder geval niet met een gezond persoon te maken). Voor de meeste mensen schommelt X rond de 37° . De dichtheid $f(x)$ zal een maximum hebben in de buurt van $x = 37$.

Vb. Stel X is een aselekt gekozen digitaal getal tussen 0 en 1. Aselekt wil zeggen dat iedere waarde in $(0, 1]$ even aannemelijk wordt geacht. Een binair getal in $(0, 1]$ is te schrijven als

$\omega = 0.\omega_1\omega_2\dots$ met $\omega_i \in \{0, 1\}$, $i = 1, 2, \dots$. Bekijk nu Ω_n = de verzameling van alle binaire getallen waarvoor de eerste n digits overeenkomen met die van ω . Voor een aselekt gekozen getal is de kans op een nul op de i -de plaats gelijk aan de kans op een één (dus gelijk aan $1/2$) en onafhankelijk van de uitkomst op een andere plaats. D.w.z. iedere uitkomst van de eerste n digits heeft dezelfde kans. Dus

$$P(X \in \Omega_n) = 2^{-n}.$$

We vinden dus dat voor alle ω

$$P(X = \omega) = \lim_{n \rightarrow \infty} 2^{-n} = 0.$$

M.a.w. iedere uitkomst heeft kans nul! We stellen nu de kans op een interval, zeg $(x_1, x_2]$, gelijk aan de lengte van dat interval. Dan

$$P(x_1 < X \leq x_2) = x_2 - x_1.$$

Verder stellen we $f(x) = 1$ voor alle x tussen 0 en 1. Dit geeft weer dat we alle uitkomsten x even aannemelijk vinden, want f is constant op $(0, 1]$. Voor een willekeurige gebeurtenis $A \subset (0, 1]$ stellen we

$$P(X \in A) = \int_A f(x) dx.$$

Zo hebben we een beschrijving van de verdeling van X . In praktijk kunnen we slechts tot een paar cijfers achter de komma meten, d.w.z. we nemen alleen Ω_n waar. Dit correspondeert met een discrete s.g. waarvan alle 2^n uitkomsten gelijke kans hebben.

EIGENSCHAP. Als X een continue s.g. is, dan bezit X (onder zekere voorwaarden) een dichtheid $f(x)$ zó dat

$$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x) dx.$$

In dit college bekijken we alleen continue s.g.ⁿ die een dichtheid bezitten met bovenstaande eigenschap, maar er zijn dus in principe ook andere continue s.g.ⁿ. Verder merken we nog op, dat ook de kans op een willekeurige gebeurtenis (dus niet alleen op intervallen) berekend kan worden m.b.v. de dichtheid, n.l. door te integreren over die gebeurtenis. Omdat dit wel eens lastige integraalrekening kan opleveren laten we dit verder buiten beschouwing.

De kans op een interval is gelijk aan de oppervlakte onder de grafiek van f , bij dat interval. Het is duidelijk dat dit alleen zinnig is als f niet-negatief is, want kansen moeten niet-negatief zijn. Bovendien moet de oppervlakte onder de gehele grafiek gelijk aan 1 zijn, want $P(X \in \mathbf{R}) = 1$. Het blijkt nu dat iedere functie die aan deze twee voorwaarden voldoet, gezien kan worden als een dichtheid, d.w.z. er kan een continue s.g. mee geassocieerd worden. We kunnen daarom zeggen:

$$f(x) \text{ is dichtheid} \Leftrightarrow \begin{cases} (i) f(x) \geq 0 \text{ voor alle } x, \\ (ii) \int_{-\infty}^{\infty} f(x) dx = 1. \end{cases}$$

De verdelingsfunctie $F(x)$ van een continue s.g. is net zo gedefinieerd als bij discrete stochastische grootheden, n.l.

$$F(x) = P(X \leq x).$$

Bij een continue s.g. betekent dit dat

$$F(x) = \int_{-\infty}^x f(t)dt,$$

zodat F een primitieve van f is, ofwel $f(x) = dF(x)/dx$. Bij een continue s.g. is de verdelingsfunctie ook continu (terwijl de verdelingsfunctie van een discrete s.g. een trapfunctie is).

We vatten wat eigenschappen samen (vergelijk (i) t/m (iv) met die voor discrete stochastische grootheden):

- (i) $f(x) \geq 0$, $\int_{-\infty}^{\infty} f(x)dx = 1$,
- (ii) $0 \leq F(x) \leq 1$, $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow \infty} F(x) = 1$,
- (iii) $F(x)$ is een stijgende (d.w.z. niet-dalende) functie,
- (iv) $F(x)$ is continu,
- (v) $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$,
- (vi) $P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2) = P(x_1 \leq X < x_2) = P(x_1 < X \leq x_2)$ en $P(X = x) = 0$.

Het maakt volgens (vi) niet uit of we de eindpunten van het interval al of niet meenemen. De kans op zo'n eindpunt is toch nul.

VOORBEELDEN VAN CONTINUE VERDELINGEN

(1) Homogene verdeling (uniforme verdeling) op $[a, b]$.

Aan het begin van deze subparagraaf bekeken we al als voorbeeld de *homogene* verdeling op $(0, 1]$. Dit is hetzelfde als de homogene verdeling op $[0, 1]$ of $(0, 1)$ of $[0, 1)$. We nemen nu een algemener interval met eindpunten $a < b$, zeg $[a, b]$ en dichtheid

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{anders} \end{cases}.$$

De s.g. X bezit een homogene verdeling op $[a, b]$ als X bovenstaande dichtheid heeft. De dichtheid is constant, zeg gelijk aan c , op $[a, b]$ en we hebben c zó gekozen dat f tot 1 integreert. De verdelingsfunctie wordt

$$F(x) = \begin{cases} 0, & x \leq a, \\ \int_{-\infty}^x \frac{1}{b-a} dx = \frac{x-a}{b-a}, & x \in [a, b], \\ 1, & x \geq b, \end{cases}$$

en

$$P(x_1 \leq X \leq x_2) = \frac{x_2 - x_1}{b - a} = \frac{\text{langte subinterval}}{\text{langte hele interval}},$$

voor alle $a \leq x_1 \leq x_2 \leq b$.

(2) Normale verdeling met parameters μ en σ^2 .

X bezit een *normale* verdeling als de dichtheid is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

Hier zijn $\mu \in \mathbf{R}$ en $\sigma^2 > 0$ parameters, en σ is de positieve wortel uit σ^2 . Omdat we hier μ en σ^2 niet verder specificeren, beschrijven we in feite een *klasse* van verdelingen. De parameter μ geeft het maximum van $f(x)$ aan, en $\mu \pm \sigma$ zijn de buigpunten. De breedte van de grafiek wordt bepaald door σ .

De normale verdeling is ingevoerd omdat het blijkt dat er makkelijk mee te rekenen valt, maar ook omdat deze verdeling een goede benadering is voor bepaalde andere verdelingen. We komen hier nog op terug. De notatie voor de normale verdeling is: $N(\mu, \sigma^2)$ -verdeling. We schrijven soms $X \sim N(\mu, \sigma^2)$, waarmee dan bedoeld wordt dat X normaal verdeeld is met parameters μ en σ^2 .

De *standaard* normale verdeling ($N(0, 1)$ -verdeling) betreft het geval $\mu = 0$, $\sigma^2 = 1$. De dichtheid is dan

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

en de standaard normale verdelingsfunctie is

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt.$$

Deze kan verder niet expliciet worden uitgerekend, maar er bestaan wel tabellen van (zie blz.).

Stel nu dat $X \sim N(\mu, \sigma^2)$. Dan $Y := (X - \mu)/\sigma \sim N(0, 1)$. Andersom geldt ook: als $Y \sim N(0, 1)$, dan $X := \sigma Y + \mu \sim N(\mu, \sigma^2)$. Dus als $F(x)$ de verdelingsfunctie van X is, dan

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

Zo kan men m.b.v. de tabel voor de standaard normale verdeling, de verdelingsfunctie voor iedere andere normale verdeling berekenen. Nu is $\Phi(x)$ meestal alleen getabelleerd voor $x \geq 0$. Omdat $\phi(x)$ symmetrisch rond $x = 0$ is, geldt echter

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

zodat $\Phi(x)$ voor negatieve waarden van x ook uit de tabel af te lezen is.

Vb. Laat X de omgevingstemperatuur in de zomer zijn. Stel dat $X \sim N(21^\circ, (3^\circ)^2)$. Wat is dan de kans dat de temperatuur hoogstens 15° is? We willen dan $P(X \leq 15)$ weten. Dat is:

$$\begin{aligned} P(X \leq 15) &= P\left(\frac{X - 21}{3} \leq \frac{15 - 21}{3}\right) = P\left(\frac{X - 21}{3} \leq -2\right) \\ &= \Phi(-2) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228. \end{aligned}$$

(3) Exponentiële verdeling met parameter λ .

X bezit een *exponentiële* verdeling als de dichtheid is:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Hier is $\lambda > 0$ weer een parameter. De verdelingsfunctie is nu

$$F(x) = 1 - e^{-\lambda x}.$$

We zullen een interpretatie geven aan de hand van een voorbeeld.

Vb. Laat X het tijdsinterval zijn, dat verloopt tussen twee opeenvolgende auto's dat langs een vast punt langs de snelweg raast. Noem Y_T het aantal auto's dat langs dit punt komt gedurende een tijdsinterval van lengte T . Stel dat Y_T Poissonverdeeld is met parameter λT (zie voorbeeld (5) van de discrete verdelingen). Dan

$$P(Y_T = k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T},$$

voor $k \in \{0, 1, \dots\}$. Zo vinden we

$$\begin{aligned} P(X \leq x) &= P(\text{minstens 1 auto in tijdsinterval met lengte } x) \\ &= 1 - P(\text{geen auto's in tijdsinterval met lengte } x) = 1 - P(Y_x = 0) = 1 - e^{-\lambda x}. \end{aligned}$$

We zien dat X exponentieel verdeeld is met parameter λ . Als de intensiteit λ groot is, komen er veel auto's langs, en zal men over het algemeen niet lang op de volgende auto hoeven te wachten.

1.8. Simultane verdelingen

In de vorige paragraaf behandelden we stochastische grootheden $X \in \mathbf{R}$. Men kan ook de verdeling van stochastische *vectoren* beschouwen. Deze zijn van de vorm (X_1, \dots, X_n) , met X_1, \dots, X_n stochastische grootheden. We zullen vooral ingaan op de verdeling van paren van stochastische grootheden ($n = 2$). Het geval $n > 2$ is een rechtstreekse uitbreiding.

1.8.1. Simultane discrete verdelingen

Vb. We bekijken een kast met poststukken, die geordend zijn naar twee aspecten, n.l. gefrankeerd v.s. ongefrankeerd en binnenland v.s. buitenland. $X = 1$ is gefrankeerd, $X = 0$ is ongefrankeerd, en $Y = 1$ is binnenland, $Y = 0$ is buitenland. De aantallen zijn gegeven in de volgende tabel:

	$X = 1$	$X = 0$	
$Y = 1$	6	5	11
$Y = 0$	3	0	3
	9	5	14

Trekken we nu aselekt een poststuk uit de kast, dan vinden we

$$P(X = 1, Y = 1) = 6/14, \quad P(X = 0, Y = 1) = 5/14,$$

$$P(X = 1, Y = 0) = 3/14, \quad P(X = 0, Y = 0) = 0,$$

en

$$P(X = 1) = 9/14, \quad P(X = 0) = 5/14,$$

$$P(Y = 1) = 11/14, \quad P(Y = 0) = 3/14.$$

De kansen op een uitkomst voor het paar (X, Y) noemt men de *simultane* verdeling van (X, Y) . De kansen op een uitkomst voor X heet dan *marginale* verdeling, en evenzo voor Y . Meestal laat men echter de woorden *simultaan* of *marginiaal* weg.

DEFINITIE. Stel X is een s.g. met waarden in $\{x_1, x_2, \dots\}$ en Y is een s.g. met waarden in $\{y_1, y_2, \dots\}$. Dan heet

$$p_{ij} = P(X = x_i, Y = y_j), \quad i, j = 1, 2, \dots$$

de *simultane* verdeling van X en Y ,

$$p_{i+} = P(X = x_i) = \sum_j p_{ij}, \quad i = 1, 2, \dots$$

de *marginale* verdeling van X en

$$p_{+j} = P(Y = y_j) = \sum_i p_{ij}, \quad j = 1, 2, \dots$$

de *marginale* verdeling van Y .

De som van de kansen is natuurlijk weer één:

$$\sum_i \sum_j p_{ij} = 1.$$

Vb. Men gooit $5 \times$ met een dobbelsteen. Zij X het aantal 1'en dat gevonden wordt en Y het aantal 2'en. Dan

$$P(X = 0, Y = 0) = \left(\frac{4}{6}\right)^5,$$

$$P(X = 0, Y = 1) = \frac{1}{6} \left(\frac{4}{6}\right)^4 \times 5,$$

en zo vinden we algemeen

$$P(X = x, Y = y) = \left(\frac{1}{6}\right)^x \left(\frac{1}{6}\right)^y \left(\frac{4}{6}\right)^{5-(x+y)} \times c,$$

waarbij c het aantal rijtjes is met x 1'en y 2'en de overige elementen in $\{3, 4, 5, 6\}$. Dus

$$c = \binom{5}{x} \binom{5-x}{y} = \frac{5!}{x!y!(5-(x+y))!}.$$

Als we n keer met de dobbelsteen gooien vinden we

$$P(X = x, Y = y) = \frac{n!}{x!y!(n-(x+y))!} p^{x+y} (1-2p)^{n-(x+y)},$$

met $p = 1/6$. Dit geldt voor alle $x \in \{0, \dots, n\}$ en $y \in \{0, \dots, n-x\}$. De marginale verdeling van X en van Y weten we al, n.l. dat is de binomiale verdeling met parameters n en $p = 1/6$. We kunnen dit nog eens controleren. B.v. de marginale verdeling van X is

$$\begin{aligned} P(X = x) &= \sum_{y=0}^{n-x} P(X = x, Y = y) \\ &= \sum_{y=0}^{n-x} \frac{n!}{x!y!(n-(x+y))!} p^{x+y} (1-2p)^{n-(x+y)} \\ &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \sum_{y=0}^{n-x} \frac{(n-x)!}{y!(n-x-y)!} \left(\frac{p}{1-p}\right)^y \left(1 - \frac{p}{1-p}\right)^{n-x-y} \\ &= \binom{n}{x} p^x (1-p)^{n-x} C, \end{aligned}$$

waarbij

$$C = \sum_{y=0}^{n-x} \binom{n-x}{y} \left(\frac{p}{1-p}\right)^y \left(1 - \frac{p}{1-p}\right)^{n-x-y} = 1,$$

want dit is de som van alle kansen op de uitkomsten $0, 1, \dots, n-x$ van een binomiale verdeling met parameters $n-x$ en $p/(1-p)$.

We kunnen ook de verdeling van functies van X en Y uitrekenen.

Vb. Laat X de wachttijd zijn totdat we voor de eerste keer 6 gooien met een dobbelsteen, en Y de wachttijd tot voor de eerste keer 6 gooien, vanaf X gerekend. Dan is $X + Y$ de wachttijd totdat we voor de tweede keer 6 gooien. De verdeling van $X + Y$ weten we in principe al, het is n.l. de negatief binomiale verdeling. Maar we kunnen deze ook als volgt vinden. We hebben

$$P(X = k) = P(Y = k) = (1 - p)^{k-1}p$$

met $p = 1/6$. Dus

$$P(X = k, X + Y = n) = P(X = k, Y = n - k) = (1 - p)^{k-1}p(1 - p)^{n-k-1}p = (1 - p)^{n-2}p^2.$$

Hieruit volgt dat

$$P(X + Y = n) = \sum_{k=1}^{n-1} (1 - p)^{n-2}p^2 = (n - 1)(1 - p)^{n-2}p^2,$$

inderdaad de negatief binomiale verdeling met parameters 2 en p . In dit voorbeeld is het ook interessant om de voorwaardelijke verdeling van X gegeven $X + Y$ te bekijken. Deze ziet er n.l. nogal simpel uit:

$$P(X = k | X + Y = n) = \frac{P(X = k, X + Y = n)}{P(X + Y = n)} = \frac{1}{n - 1}, \quad k = 1, \dots, n - 1.$$

Gegeven $X + Y$ zijn dus alle mogelijke waarden van X even waarschijnlijk.

1.8.2. Continue simultane verdelingen

Vb. Laat (X, Y) de plaats van een algje bekeken onder de microscoop zijn, d.w.z. X is de x -coördinaat en Y de y -coördinaat. We willen de kans weten dat het algje zich in een bepaald gebied, zeg A , bevindt. Stel het medium bevat meer voedsel op de linkerhelft van de kweek, dan is het aannemelijker dat het algje zich links bevindt. We geven de aannemelijkheid van de plaats (x, y) weer aan m.b.v. een dichtheid $f(x, y)$, die de kansen beschrijft:

$$P((X, Y) \in A) = \int_A \int f(x, y) dx dy.$$

Een dichtheid kunnen we interpreteren aan de hand van het fysisch analogon. Zo is in de natuurkunde de dichtheid van een bepaald soort deeltjes gelijk aan de concentratie (d.w.z. het aantal deeltjes per oppervlakte-eenheid). De concentratie kan afhangen van de plaats waar men aan het kijken is, m.a.w. concentratie is een functie van plaats (een functie van 2 variabelen als men het over concentraties in oppervlakten heeft). Als de concentratie constant is kan men de totale hoeveelheid deeltjes in een gebiedje uitrekenen door de concentratie te vermenigvuldigen met de oppervlakte van dat gebiedje. Als de

concentratie echter afhangt van plaats vindt men de totale hoeveelheid door de concentratie over het gebiedje te integreren.

EIGENSCHAP.

Een 2-dimensionale stochastische vector (X, Y) , met X en Y continu verdeelde s.g.ⁿ bezit onder zekere voorwaarden een *dichtheid* $f(x, y)$ met

- (i) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$,
- (ii) $f(x, y) \geq 0$ voor alle x en y ,
- (iii) $P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x, y) dx dy$.

Eigenschappen (i) en (ii) zijn weer nodig en voldoende voorwaarden opdat $f(x, y)$ een dichtheid is. Omdat de kans op de hele uitkomstenruimte (in dit geval \mathbf{R}^2) één is moet (i) gelden. Dit zegt dat de inhoud onder de grafiek één is. Kansen zijn niet-negatief, dus we hebben ook (ii) nodig. Eigenschap (iii) zegt dat de kans op een rechthoekje gelijk is aan de inhoud onder de grafiek van de functie op dat rechthoekje. Men kan ook weer kansen op algemenere gebieden definiëren, maar dan wordt de integraalrekening lastiger. We laten dit achterwege.

Het maakt weer niet uit of men de zijden van het rechthoekje gegeven door de punten (x_1, y_1) en (x_2, y_2) al of niet meeneemt, omdat de kans op zo'n zijde nul is. Het berekenen van een dubbele integraal komt neer op het twee maal uitrekenen van een enkele integraal, en het maakt niet uit of men eerst over x integreert en dan over y of andersom:

$$\int_{y_1}^{y_2} \left\{ \int_{x_1}^{x_2} f(x, y) dx \right\} dy = \int_{x_1}^{x_2} \left\{ \int_{y_1}^{y_2} f(x, y) dy \right\} dx.$$

We zeggen dat $f(x, y)$ de dichtheid van de *simultane* verdeling van X en Y is. De *marginale* dichtheid van X is

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy,$$

en de marginale verdeling van Y is

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

Dus als de simultane dichtheid bekend is kan men de marginale dichtheden ook vinden. Andersom geldt niet: als men de marginale verdelingen kent hoeft men niet de simultane verdeling te kennen. Dit komt doordat X en Y met elkaar in verband kunnen staan. Zo'n eventueel verband is niet terug te vinden in de marginale verdelingen.

Vb. Twee-dimensionale homogene verdeling.

(X, Y) bezit een homogene verdeling op de rechthoek $[a, b] \times [c, d]$ als de dichtheid is:

$$f(x, y) = \begin{cases} \frac{1}{(b-a)(d-c)}, & \text{als } a \leq x \leq b \text{ en } c \leq y \leq d, \\ 0, & \text{anders.} \end{cases}$$

Dan

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \frac{1}{(b-a)(d-c)} \int_{x_1}^{x_2} \int_{y_1}^{y_2} dy dx$$

$$= \frac{(x_2 - x_1)(y_2 - y_1)}{(b - a)(d - c)} = \frac{\text{oppervl.}([x_1, x_2] \times [y_1, y_2])}{\text{oppervl.}([a, b] \times [c, d])}$$

voor $a \leq x_1 \leq x_2 \leq b$ en $c \leq y_1 \leq y_2 \leq d$.

Dit is ook de inhoud onder de grafiek. Merk op dat de marginale verdeling van X de homogene verdeling op $[a, b]$ is, en de marginale verdeling van Y is de homogene verdeling op $[c, d]$.

Vb. Stel (X, Y) heeft dichtheid

$$f(x, y) = \begin{cases} 2y, & \text{als } 0 \leq x \leq 1 \text{ en } 0 \leq y \leq 1, \\ 0, & \text{anders.} \end{cases}$$

Dan is X homogeen verdeeld op $[0, 1]$ en de dichtheid van Y is

$$h(y) = 2y, \quad 0 \leq y \leq 1.$$

1.8.3. Onafhankelijke stochastische grootheden

Voor een stochastische vector (X_1, \dots, X_n) is de definitie van de simultane verdeling de voor de hand liggende uitbreiding van het twee-dimensionale geval. Als X_1, \dots, X_n discrete stochastische grootheden zijn wordt de simultane verdeling gegeven door alle kansen $P(X_1 = x_1, \dots, X_n = x_n)$, waarbij x_i de mogelijke waarden van X_i doorloopt. X_1, \dots, X_n heten dan onderling onafhankelijk als $P(X_1 = x_1, \dots, x_n = x_n) = P(X_1 = x_1) \dots P(X_2 = x_2)$, voor alle mogelijke waarden van x_1, \dots, x_n .

Als X_1, \dots, X_n continu zijn wordt de simultane verdeling gegeven door de simultane dichtheid $f(x_1, \dots, x_n)$. De rij X_1, \dots, X_n bestaat uit o.o. stochastische grootheden als geldt:

$$f(x_1, \dots, x_n) = f_1(x_1) \dots f_n(x_n),$$

voor alle $(x_1, \dots, x_n) \in \mathbf{R}^n$. Hier is $f_i(x_i)$ de marginale dichtheid van X_i in het punt x_i , $i = 1, \dots, n$.

Vb. Stel $f(x, y) = 2y$ voor $x \in [0, 1]$ en $y \in [0, 1]$ en $f(x, y) = 0$ anders. De marginale dichtheden zijn dan $g(x) = 1$ en $h(y) = 2y$ op $[0, 1]$. Dus $f(x, y) = g(x)h(y)$ voor alle x en y , d.w.z. X en Y zijn o.o..

Vb. Stel $f(x, y) = x + y$ op $[0, 1] \times [0, 1]$. De marginale dichtheden zijn dan $g(x) = x + 1/2$ op $[0, 1]$ en $h(y) = y + 1/2$ op $[0, 1]$, zodat $f(x, y) \neq g(x)h(y)$. Dus in dit geval zijn X en Y niet o.o..

1.9. De verdeling van functies van stochastische grootheden

DE VERDELING VAN EEN FUNCTIE VAN EEN DISCRETE S.G.

Laat X een discrete s.g. zijn met waarden in $\{x_1, x_2, \dots\}$ en met kansen $p_i = P(X = x_i)$, $i = 1, 2, \dots$. Noem $Z = g(X)$, met g één of andere functie. Dan is Z ook weer een discrete s.g. met waarden in $\{g(x_1), g(x_2), \dots\}$ en met kansen

$$P(Z = z) = \sum_{\{i: g(x_i)=z\}} p_i.$$

Vb. Stel $P(X = 1) = P(X = -1) = P(X = 0) = 1/3$ en laat $Z = X^2$ zijn. Dan $P(Z = 1) = 2/3$ en $P(Z = 0) = 1/3$.

DE VERDELING VAN EEN FUNCTIE VAN EEN CONTINUE S.G.

Laat X een continue s.g. zijn, met dichtheid $f_X(x)$. In het algemeen kan het best lastig zijn om nu de verdeling van een functie $Z = g(X)$ te bepalen. We bekijken hier alleen wat voorbeelden. De algemene methode is om eerst de verdelingsfunctie van Z te bepalen uit de verdelingsfunctie $F_X(x)$ van X , en vervolgens (indien mogelijk) te differentiëren.

Vb. Stel $Z = \mu + \sigma X$ met $\sigma > 0$. Dan

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(\mu + \sigma X \leq z) \\ &= P(X \leq \frac{z - \mu}{\sigma}) = F_X(\frac{z - \mu}{\sigma}). \end{aligned}$$

Dus de dichtheid van Z is

$$f_Z(z) = \frac{dF_Z(z)}{dz} = f_X(\frac{z - \mu}{\sigma}) \frac{1}{\sigma}$$

(gebruik de kettingregel). Een speciaal geval is als $X \sim N(0, 1)$. Dan

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

en

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2(\frac{z-\mu}{\sigma})^2}.$$

Dit is de dichtheid van een $N(\mu, \sigma^2)$ -verdeling.

Vb. Stel $Z = X^2$, dan is de verdelingsfunctie van Z :

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X^2 \leq z) = P(-\sqrt{z} \leq X \leq \sqrt{z}) \\ &= F_X(\sqrt{z}) - F_X(-\sqrt{z}). \end{aligned}$$

De dichtheid van Z is

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{dF_X(\sqrt{z})}{dz} - \frac{dF_X(-\sqrt{z})}{dz}$$

$$= \frac{f_X(\sqrt{z})}{2\sqrt{z}} + \frac{f_X(-\sqrt{z})}{2\sqrt{z}}, \quad z > 0,$$

want $d\sqrt{z}/dz = 1/(2\sqrt{z})$, $z > 0$. Het geval $z = 0$ hoeven we niet te bekijken, want $P(X = 0) = 0$ voor een continue s.g. X , dus ook $P(Z = 0) = 0$. Een speciaal geval is als $X \sim N(0, 1)$. Dan

$$F_Z(z) = \frac{1}{\sqrt{2\pi z}} e^{-(1/2)z}, \quad z > 0.$$

(Dit heet de *chi-kwadraat*-verdeling met één vrijheidsgraad.)

DE VERDELING VAN DE SOM VAN TWEE (O.O.) DISCRETE STOCHASTISCHE GROOTHEDEN.

Laat X , resp. Y , een discrete s.g. zijn met waarden in $\{x_1, x_2, \dots\}$, resp. $\{y_1, y_2, \dots\}$. Bekijk $Z = X + Y$. Dan is de verdeling van Z :

$$P(Z = z) = P(X + Y = z) = \sum_{i=1}^{\infty} P(X = x_i, Y = z - x_i).$$

Dus de verdeling van Z volgt uit de simultane verdeling van X en Y . I.h.b. als X en Y o.o. zijn:

$$P(Z = z) = \sum_{i=1}^{\infty} P(X = x_i)P(Y = z - x_i).$$

Men noemt dit wel de *convolutie* van de verdeling van X en Y . In het laatste voorbeeld van Paragraaf 1.8.1 vonden we zo de verdeling van $X + Y$ voor het geval dat X en Y o.o. geometrisch verdeeld zijn.

DE VERDELING VAN DE SOM VAN TWEE O.O. CONTINUE STOCHASTISCHE GROOTHEDEN.

Stel X , resp. Y , is continu verdeeld met dichtheid $f_X(x)$, resp. $f_Y(y)$. We nemen bovendien aan dat X en Y o.o. zijn (omdat we dit geval het meest zullen tegenkomen). Dan is de dichtheid $f_Z(z)$ van $Z := X + Y$:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx.$$

Dit heet weer de *convolutie* van $f_X(x)$ en $f_Y(y)$.

Vb. Stel X en Y zijn o.o. en homogeen verdeeld op $[0, 1]$. Dan is de dichtheid van $Z = X + Y$:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx = \int_0^1 f_Y(z-x)dx,$$

want $f_X(x) = 1$ voor $0 \leq x \leq 1$ en $f_X(x) = 0$ voor $x < 0$ of $x > 1$. Verder is $f_Y(z-x) = 1$ voor $0 \leq z-x \leq 1$ en anders is $f_Y(z-x) = 0$. Hieruit volgt dat

$$f_Z(z) = \begin{cases} z, & 0 \leq z \leq 1, \\ 2-z, & 1 \leq z \leq 2, \\ 0, & \text{anders.} \end{cases}$$

Vb. Stel X en Y zijn o.o. normaal verdeeld. Dan kan men m.b.v. bovenstaande formule afleiden dat $X + Y$ ook weer normaal verdeeld is. Ook als X en Y niet o.o. zijn is dit het geval. We zullen dit echter niet bewijzen, want het is nogal technische integraalrekening.

1.10. Verwachting en variantie

DEFINITIE. Stel X is een discrete stochastische grootte met waarden in $\{x_1, x_2, \dots\}$. Dan heet

$$E(X) = \sum_i x_i P(X = x_i)$$

de *verwachting* van X , en

$$Eg(X) = \sum_i g(x_i) P(X = x_i)$$

de verwachting van de functie $g(X)$ van X .

Vb. Laat X het aantal ogen zijn bij één keer gooien met een dobbelsteen. Dan $E(X) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$ en b.v. $E(X^2) = (1 + 4 + 9 + 16 + 25 + 36)/6 = 91/6$.

Het fysisch analogon van verwachting is *zwaartepunt*. Stel we leggen massa's π_1, π_2, \dots op de punten x_1, x_2, \dots . Dan is het zwaartepunt $(\sum_i x_i \pi_i) / (\sum_i \pi_i) = \sum_i x_i p_i$ met $p_i = \pi_i / (\sum_i \pi_i)$, $i = 1, 2, \dots$.

Vb. Beschouw een populatie van N elementen, waarvan er R kenmerk S bezitten. Neem een aselechte steekproef van grootte n en laat X het aantal elementen in de steekproef met kenmerk S zijn.

(a) Met terugleggen.

$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, met $p = R/N$ en $k \in \{0, 1, \dots, n\}$. Dus

$$\begin{aligned} E(X) &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} = np, \end{aligned}$$

want de som van alle kansen op de uitkomsten van een binomiale verdeling met parameters $(n-1)$ en p is 1.

(b) Zonder terugleggen.

$P(X = k) = \binom{R}{k} \binom{N-R}{n-k} / \binom{N}{n}$, voor $k \in \{0, \dots, n\}$, waarbij we voor het gemak $n \leq R$ en $n \leq N - R$ veronderstellen. Dus

$$E(X) = \sum_{k=0}^n k \binom{R}{k} \binom{N-R}{n-k} / \binom{N}{n}$$

$$= nR/N \sum_{k=1}^n \binom{R-1}{k-1} \binom{(N-1)-(R-1)}{(n-1)-(k-1)} / \binom{N-1}{n-1} = nR/N = np.$$

DEFINITIE. Stel X is een continue s.g. met dichtheid $f(x)$, dan heet

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

de verwachting van X , en

$$Eg(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

de verwachting van de functie $g(X)$ van X .

Vb. Stel X is homogeen verdeeld op het interval $[0, 1]$. Dan

$$E(X) = \int_0^1 xdx = 1/2$$

en b.v.

$$E \cos(X) = \int_0^1 \cos(x)dx = -\sin(1).$$

We noemen de verwachting van een s.g. (discreet of continu) ook wel eens het *gemiddelde*. Daarmee wordt dus niet eenvoudig het gemiddelde van een aantal getallen bedoeld! Voor discrete X is het over het algemeen een *gewogen gemiddelde* (met de kansen als gewichten) en voor continue X is het een integraal.

EIGENSCHAP. Zij X en Y twee stochastische grootheden (discreet of continu) en a, b, c getallen. Dan

$$E(aX + bY + c) = aE(X) + bE(Y) + c.$$

Als X_1, \dots, X_n een rij van stochastische grootheden is, dan vinden we door het bovenstaande herhaald toe te passen:

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

In woorden: de verwachting van de som is de som van de verwachtingen.

Vb. Beschouw nog eens n aselechte trekkingen uit een populatie van N elementen, waarvan er R kenmerk S bezitten. Noem

$$X_i = \begin{cases} 1, & \text{als bij de } i\text{-de trekking kenmerk } S \text{ wordt gevonden,} \\ 0, & \text{anders.} \end{cases}$$

Dan is $X = \sum_{i=1}^n X_i$ het aantal elementen in de steekproef met kenmerk S . Nu hebben we al gezien dat met of zonder teruglegging

$$P(X_i = 1) = R/N = p.$$

Voor een s.g. X_i die alleen de waarden 0 of 1 kan aannemen is

$$E(X_i) = P(X_i = 1),$$

dus $E(X_i) = p$, $i = 1, \dots, n$. Daarom is

$$E(X) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n p = np.$$

DEFINITIE. De *variantie* van een s.g. X is de verwachte kwadratische afwijking van het gemiddelde:

$$\text{var}(X) = E(X - EX)^2.$$

De *standaardafwijking* van X is

$$\sigma(X) = \sqrt{\text{var}(X)}.$$

We schrijven ook wel $\sigma^2(X)$ voor de variantie van X .

De standaardafwijking is een maat voor de *spreiding*.

EIGENSCHAP. $\text{var}(X) = EX^2 - (EX)^2$, want, als we $EX = \mu$ noemen, dan $\text{var}(X) = E(X - \mu)^2 = EX^2 - 2\mu EX + \mu^2 = EX^2 - 2\mu^2 + \mu^2 = EX^2 - \mu^2$.

Vb. Bij het roulettespel zetten we één gulden in op oneven. De kans om een gulden te winnen is $18/37$ en de kans om een gulden te verliezen is $19/37$ (nul is hier een even getal). Dus als X onze winst is, dan $EX = -1/37$ en $EX^2 = 1$, dus $\text{var}(X) = 1 - (1/37)^2 = 0.9993$.

Vb. Bij het roulettespel zetten we één gulden in op 23. De winst is dan $X = 35$ met kans $1/37$ en $X = -1$ met kans $36/37$. Dus $EX = -1/37$, net als in het vorige voorbeeld. Maar $EX^2 = (35)^2/37 + 36/37 = 1261/37$ zodat $\text{var}(X) = 1261/37 - (1/37)^2 = 34.0803$.

Vb. Laat X homogeen verdeeld zijn op $[0, 1]$. Dan $EX = 1/2$ en

$$EX^2 = \int_0^1 x^2 dx = 1/3.$$

Dus $\text{var}(X) = 1/3 - (1/2)^2 = 1/12$.

EIGENSCHAPPEN.

(i) $\text{var}(X) = E(X - EX)^2 \geq 0$. Hier volgt ook uit dat $EX^2 \geq (EX)^2$, want we hebben dat $\text{var}(X) = EX^2 - (EX)^2$.

(ii) $\text{var}(X) = 0$ dan en slechts dan als X een ontaarde verdeling bezit, d.w.z. voor zekere constante c is $P(X = c) = 1$. Deze constante is dan $c = EX$ (een s.g. die alleen de waarde c kan aannemen heeft natuurlijk ook verwachting c). We zeggen ook wel dat X geconcentreerd is in c . (In het algemeen geeft ook de variantie de mate van concentratie van X rond de verwachting aan.) Als X een discrete s.g. is met waarden in $\{x_1, x_2, \dots\}$ en met verwachting μ , dan is per definitie $\text{var}(X) = \sum_i (x_i - \mu)^2 P(X = x_i)$. Dit kan alleen gelijk aan nul zijn als alle termen nul zijn. Dus dan moet wel gelden dat $X = \mu$ met kans 1. Voor het geval van continue stochastische grootheden kan men de eigenschap op analoge wijze inzien.

(iii) Als a en b getallen zijn, dan $\text{var}(aX + b) = a^2 \text{var}(X)$. Immers, noem weer $EX = \mu$. Dan $E(aX + b) = a\mu + b$ en $\text{var}(aX + b) = E((aX + b) - (a\mu + b))^2 = E(aX - a\mu)^2 = E(a^2(X - \mu)^2) = a^2 E(X - \mu)^2 = a^2 \text{var}(X)$.

VOORBEELDEN.

(1) Stel X bezit de Poissonverdeling:

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu}, \quad k = 0, 1, \dots$$

Dan

$$\begin{aligned} EX &= \sum_{k=0}^{\infty} k \frac{\mu^k}{k!} e^{-\mu} \\ &= \mu \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} e^{-\mu} = \mu, \end{aligned}$$

en

$$EX^2 = \sum_{k=0}^{\infty} k^2 \frac{\mu^k}{k!} e^{-\mu}.$$

Nu is $1/(k-1)! + 1/(k-2)! = k/(k-1)!$, $k \geq 2$. Hieruit vinden we dat

$$EX^2 = \sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!} e^{-\mu} + \sum_{k=2}^{\infty} \frac{\mu^k}{(k-2)!} e^{-\mu} = \mu + \mu^2.$$

Dus $\text{var}(X) = \mu + \mu^2 - \mu^2 = \mu$. Bij de Poissonverdeling zijn verwachting en variantie gelijk.

(2) Stel Y is $N(0, 1)$ -verdeeld. Dan

$$EY = \int_{-\infty}^{\infty} y \phi(y) dy.$$

Omdat $\phi(y)$ symmetrisch is rond $y = 0$ is $EY = 0$. Verder kan men m.b.v. partiële integratie inzien dat

$$EY^2 = \int_{-\infty}^{\infty} y^2 \phi(y) dy$$

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (-y) d(e^{-(1/2)y^2}) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1/2)y^2} dy = \int_{-\infty}^{\infty} \phi(y) dy = 1.
\end{aligned}$$

Dus ook $\text{var}(Y) = 1$.

Als nu X $N(\mu, \sigma^2)$ -verdeeld is dan is $Z := (X - \mu)/\sigma$ $N(0, 1)$ -verdeeld. We zien dat $EX = E(\sigma Z + \mu) = \mu$ en $\text{var}(X) = \text{var}(\sigma Z + \mu) = \sigma^2$.

(3) Stel X is exponentieel verdeeld met parameter λ . Dan is de dichtheid

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Dus

$$EX = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda},$$

en

$$EX^2 = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2},$$

zodat $\text{var}(X) = 1/\lambda^2$.

DE VERWACHTING VAN FUNCTIES VAN STOCHASTISCHE GROOTHEDEN.

Laat $g : \mathbf{R}^2 \rightarrow \mathbf{R}$ één of andere functie van twee variabelen zijn, en X en Y twee s.g.ⁿ. Dan is $g(X, Y)$ een stochastische grootheid, met verwachting

$$Eg(X, Y) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} g(x_i, y_j) P(X = x_i, Y = y_j),$$

als X en Y discreet zijn met waarden in $\{x_1, x_2, \dots\}$ respectievelijk $\{y_1, y_2, \dots\}$ en

$$Eg(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy,$$

als X en Y continu zijn met dichtheid $f(x, y)$.

Vb. Laat (X, Y) dichtheid $f(x, y) = x + y$ op $[0, 1] \times [0, 1]$ hebben. Dan

$$E(XY) = \int_0^1 \int_0^1 xy(x + y) dx dy = \frac{1}{3}.$$

Stelling. Stel X en Y zijn o.o., dan

$$E(XY) = (EX)(EY).$$

BEWIJS. Als X en Y discreet zijn geldt

$$\begin{aligned} E(XY) &= \sum_i \sum_j x_i y_j P(X = x_i, Y = y_j) = \sum_i \sum_j x_i y_j P(X = x_i) P(Y = y_j) \\ &= \sum_i x_i P(X = x_i) \sum_j y_j P(Y = y_j) = (EX)(EY). \end{aligned}$$

Het bewijs is analoog als X en Y continu zijn. \square

DEFINITIE. X en Y heten *ongecorreleerd* als $E(XY) = (EX)(EY)$.

We hebben gezien dat twee o.o. stochastische grootheden ongecorreleerd zijn. Het omgekeerde hoeft echter niet te gelden, zoals blijkt uit het volgende voorbeeld.

Vb. Stel X is homogeen verdeeld op $[-1/2, 1/2]$ en $Y = X^2$. Dan zijn X en Y duidelijk niet o.o., want als men X weet, weet men Y ook. Maar $EX = 0$, $E(XY) = EX^3 = 0$, dus $E(XY) = (EX)(EY) = 0$, d.w.z. X en Y zijn wel ongecorreleerd.

DEFINITIE. De *covariantie* tussen twee stochastische grootheden X en Y is gedefinieerd als

$$\text{cov}(X, Y) = E((X - EX)(Y - EY)).$$

EIGENSCHAPPEN.

- (i) $\text{cov}(X, X) = \text{var}(X)$,
- (ii) $\text{cov}(X, Y) = E(XY) - (EX)(EY)$, want als we schrijven $EX = \mu$ en $EY = \nu$, dan $\text{cov}(X, Y) = E((X - \mu)(Y - \nu)) = E(XY) - \mu EY - \nu EX + \mu\nu = E(XY) - \mu\nu$.

Merk nu op dat als X en Y o.o. zijn, dan $\text{cov}(X, Y) = 0$, maar dat het omgekeerde niet waar hoeft te zijn.

Vb. Stel (X, Y) heeft dichtheid $f(x, y) = x + y$ op $[0, 1] \times [0, 1]$. Dan $E(XY) = 1/3$. De dichtheid van X is $g(x) = x + 1/2$ op $[0, 1]$. Dus $EX = 7/12$. Analoog $EY = 7/12$. De covariantie is dus

$$\text{cov}(X, Y) = 1/3 - (7/12)^2 = -1/144.$$

De covariantie is een maat voor een *lineaire* verband tussen stochastische grootheden. We zeggen dat er een *exact* lineair verband is tussen X en Y als voor zekere α en β geldt $Y = \alpha + \beta X$. In het algemeen is er natuurlijk geen exact lineair verband, maar we verwachten wel vaak een relatie in de trant van: "hoe groter X , des te groter Y " (bv. bij lichaamslengte en lichaamsgewicht) of juist: "hoe groter X des te kleiner Y ".

Vb. Stel $Y = \alpha + \beta X + V$, waarbij V en X onafhankelijk zijn. Men kan V interpreteren als een verstoring van het lineaire verband. Nu is $E(XY) = E(X(\alpha + \beta X + V)) = \alpha EX + \beta EX^2 + E(XV)$, en $(EX)(EY) = (EX)(\alpha + \beta EX + EV) = \alpha EX + \beta (EX)^2 + (EX)(EV)$. Dus $\text{cov}(X, Y) = \beta \text{var}(X)$. We zien dat de covariantie positief is als $\beta > 0$, en anders is de covariantie negatief (of nul).

In het algemeen noemen we het geval $\text{cov}(X, Y) > 0$ een positief verband en $\text{cov}(X, Y) < 0$ een negatief verband. Als $\text{cov}(X, Y) = 0$ zijn X en Y per definitie ongecorreleerd, maar er kan nog best een zeker verband zijn (er is alleen geen *lineair* verband).

Lemma. $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$.

BEWIJS. Noem $EX = \mu$ en $EY = \nu$. Er geldt

$$\begin{aligned}\text{var}(X + Y) &= E(X + Y - (\mu + \nu))^2 = E((X - \mu)^2 + (Y - \nu)^2 + 2(X - \mu)(Y - \nu)) \\ &= E(X - \mu)^2 + E(Y - \nu)^2 + 2E((X - \mu)(Y - \nu)) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y).\end{aligned}$$

□

GEVOLG. Als X en Y ongecorreleerd zijn (i.h.b. als X en Y o.o. zijn), dan

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

UITBREIDING. Door bovenstaand lemma herhaald toe te passen vindt men

$$\text{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i=2}^n \sum_{j=1}^{i-1} \text{cov}(X_i, X_j).$$

Als X_1, \dots, X_n ongecorreleerd zijn (i.h.b. als ze o.o. zijn), dan

$$\text{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{var}(X_i),$$

d.w.z. dan is de variantie van de som de som van de varianties.

Vb. Bekijk n aselechte trekkingen, uit een populatie van N elementen waarvan er R kenmerk S bezitten. Noem

$$X_i = \begin{cases} 1, & \text{als } S \text{ wordt gevonden in de } i\text{-de trekking,} \\ 0, & \text{anders.} \end{cases}$$

Bij een steekproef met of zonder terugleggen geldt

$$P(X_i = 1) = \frac{R}{N}, \quad i = 1, \dots, n.$$

Dit impliceert dat $E(X_i) = R/N$, $E(X_i^2) = R/N$ en $\text{var}(X_i) = R/N - (R/N)^2 = R/N(1 - R/N)$. Zij nu weer $X = \sum_{i=1}^n X_i$ het aantal elementen in de steekproef met kenmerk S .

(a) Met terugleggen.

X_1, \dots, X_n zijn o.o., waaruit volgt dat

$$\text{var}(X) = \sum_{i=1}^n \text{var}(X_i) = \sum_{i=1}^n \frac{R}{N} \left(1 - \frac{R}{N}\right) = n \frac{R}{N} \left(1 - \frac{R}{N}\right).$$

(b) Zonder terugleggen.

Er geldt voor $j \neq i$,

$$E(X_i X_j) = P(X_i = 1, X_j = 1) = \frac{R}{N} \frac{R-1}{N-1}.$$

Dus

$$\text{cov}(X_i, X_j) = \frac{R}{N} \frac{R-1}{N-1} - \left(\frac{R}{N}\right)^2 = \frac{R}{N} \frac{N-R}{N} \frac{1}{N-1}.$$

We vinden zo

$$\begin{aligned} \text{var}(X) &= \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i=2}^n \sum_{j=1}^{i-1} \text{cov}(X_i, X_j) \\ &= \sum_{i=1}^n \frac{R}{N} \frac{N-R}{N} + 2 \sum_{i=1}^n \sum_{j=1}^{i-1} -\frac{R}{N} \frac{N-R}{N} \frac{1}{N-1} \\ &= n \frac{R}{N} \frac{N-R}{N} + 2 \left(\frac{n(n-1)}{2}\right) \left(-\frac{R}{N} \frac{N-R}{N} \frac{1}{N-1}\right) \\ &= n \frac{R}{N} \frac{N-R}{N} \frac{N-n}{N-1}. \end{aligned}$$

De variantie bij een steekproef zonder terugleggen is kleiner dan bij een steekproef met terugleggen.

Vb. Stel X is binomiaal verdeeld met parameters n en p . Dan is $EX = np$ en $\text{var}(X) = np(1-p)$. Dit volgt onmiddellijk, doordat X weer te schrijven is als $X = \sum_{i=1}^n X_i$, met X_1, \dots, X_n o.o. met $P(X_i = 1) = 1 - P(X_i = 0) = p$, en door dezelfde redenering toe te passen als in bovenstaand voorbeeld, geval (a).

DEFINITIE. Stel X is een stochastische grootte met verwachting $EX = \mu$ en variantie $\text{var}(X) = \sigma^2$. Dan heet

$$\tilde{X} = X - \mu$$

de *gereduceerde* van X , en

$$X^* = (X - \mu)/\sigma$$

de *gestandaardiseerde* van X .

EIGENSCHAP. $E(\tilde{X}) = E(X^*) = 0$ en $\text{var}(X^*) = 1$.

DEFINITIE. De *correlatie* tussen X en Y is

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.$$

EIGENSCHAPPEN.

(i) De correlatie is, in tegenstelling tot de covariantie, een dimensieloos begrip, d.w.z. het is onafhankelijk van de meeteenheid. Of X en/of Y b.v. in centimeters of in meters wordt gemeten heeft geen invloed op de correlatiecoëfficiënt. Als \tilde{X} resp. X^* en \tilde{Y} resp. Y^* de gereduceerde resp. gestandaardiseerde van respectievelijk X en Y zijn, dan

$$\rho(X, Y) = \rho(\tilde{X}, \tilde{Y}) = \rho(X^*, Y^*).$$

(ii) Het blijkt dat

$$-1 \leq \rho(X, Y) \leq 1.$$

1.11. De wet van de grote aantallen

Vb. M.b.v. aselechte getallen en een computer simuleren we 1000000 keer gooien met een dobbelsteen. We vinden de uitkomsten $x_1 = 2, x_2 = 5, \dots, x_{1000000} = 3$. (Men kan deze getallen interpreteren als *realisaties* van stochastische grootheden $X_1, \dots, X_{1000000}$.) Het gemiddelde $\bar{x} = (1/(1000000))(x_1 + \dots + x_{1000000})$ blijkt $\bar{x} = 3.500867$ te zijn.

Het wiskundige kansbegrip is gefundeerd op het idee dat vaak herhaalde experimenten een gemiddelde uitkomst oplevert, die altijd ongeveer gelijk zal zijn. We kunnen nu een stelling formuleren, die zegt dat het gemiddelde van de uitkomsten ongeveer gelijk zal zijn aan de verwachting van de uitkomst van één experiment. We moeten hier wel goed bedenken wat we precies bedoelen met gemiddelde en verwachting. Om het onderscheid te benadrukken noemen we de eerste soms wel het steekproefgemiddelde of het empirische gemiddelde, en de tweede het populatiegemiddelde of het theoretische gemiddelde.

Onderstaand lemma hebben we nodig om de (theoretische) wet van de grote aantallen te bewijzen.

Lemma. (Een Chebyshev ongelijkheid.) Laat Z een stochastische grootheid zijn, dan geldt voor alle $c > 0$,

$$P(|Z| > c) \leq \frac{EZ^2}{c^2}.$$

BEWIJS. We tonen het alleen aan voor een discrete s.g. Z . Als Z continu verdeeld is verloopt het bewijs analoog. Per definitie

$$EZ^2 = \sum_i z_i^2 P(Z = z_i),$$

waarbij $\{z_i\}$ de mogelijke uitkomsten van Z zijn. We kunnen dit opsplitsen in twee delen:

$$EZ^2 = \sum_{|z_i| \leq c} z_i^2 P(Z = z_i) + \sum_{|z_i| > c} z_i^2 P(Z = z_i).$$

Als we hier de eerste term weglaten wordt het resultaat hoogstens kleiner (want de termen zijn ≥ 0). Wat betreft de tweede term merken we op dat als $|z_i| > c$, dan $z_i^2 > c^2$, dus

$$\sum_{|z_i| > c} z_i^2 P(Z = z_i) \geq c^2 \sum_{|z_i| > c} P(Z = z_i).$$

Nu is $\sum_{|z_i| > c} P(Z = z_i)$ precies de kans dat $|Z| > c$. Zo vinden we

$$EZ^2 \geq c^2 P(|Z| > c),$$

ofwel

$$P(|Z| > c) \leq \frac{EZ^2}{c^2}.$$

□

DEFINITIE. Het gemiddelde van een rij van stochastische grootheden X_1, \dots, X_n is

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

Als X_1, \dots, X_n o.o. zijn, en alle dezelfde verdeling hebben, noemen we de rij wel een *steekproef* uit die verdeling. Meestal zullen we een dergelijke situatie bekijken. Voor de wet van de grote aantallen hoeven we niet te veronderstellen dat ze alle dezelfde verdeling hebben, maar is het voldoende aan te nemen dat ze o.o. zijn met alle dezelfde verwachting en variantie. De verwachting en variantie van \bar{X} zijn dan als volgt:

EIGENSCHAP. Stel X_1, \dots, X_n zijn o.o. met $EX_i = \mu$ en $\text{var}(X_i) = \sigma^2$ voor alle $i \in \{1, \dots, n\}$, dan

(i) $E\bar{X} = (1/n) \sum_{i=1}^n \mu = \mu,$
(ii) $\text{var}(\bar{X}) = (1/n)^2 \sum_{i=1}^n \sigma^2 = \sigma^2/n.$

Stelling. (Wet van de grote aantallen.) Stel voor $n = 1, 2, \dots$, dat X_1, \dots, X_n o.o. zijn met $EX_i = \mu$, $\text{var}(X_i) = \sigma^2$ voor alle $i \in \{1, \dots, n\}$, dan voor alle $c > 0$,

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| > c) = 0.$$

BEWIJS. Neem in bovenstaand lemma $Z = \bar{X} - \mu$. Dan $EZ^2 = E(\bar{X} - \mu)^2 = E(\bar{X} - E\bar{X})^2 = \text{var}(\bar{X}) = \sigma^2/n$. Pas nu het lemma toe:

$$P(|\bar{X} - \mu| > c) \leq \frac{\sigma^2/n}{c^2} = \frac{\sigma^2}{nc^2}.$$

Hieruit volgt dat

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| > c) \leq \lim_{n \rightarrow \infty} \frac{\sigma^2}{nc^2} = 0.$$

□

Er geldt dus voor willekeurige c , dat de kans dat \bar{X} meer dan c van μ afwijkt, willekeurig klein wordt als het aantal waarnemingen maar groot genoeg is. Het gemiddelde \bar{X} is een stochastische grootheid, zodat we alleen kansuitspraken over \bar{X} kunnen doen. Maar doordat de variantie van \bar{X} klein is als n groot is, concentreert \bar{X} zich rond μ . In de limiet wordt de variantie nul, en we hebben gezien dat een s.g. met variantie gelijk aan nul maar één waarde kan aannemen. n.l. zijn verwachting. Bij continu verdeelde stochastische grootheden zou de dichtheid van \bar{X} er ongeveer zó uit kunnen zien:

We kunnen de wet van de grote aantallen ook als volgt interpreteren. Stel X_1, \dots, X_n zijn o.o. met verwachting μ en variantie σ^2 . Noem $V_i = X_i - \mu$, $i = 1, \dots, n$. Dan zijn ook

V_1, \dots, V_n o.o. met $EV_i = 0$ en $\text{var}(V_i) = \sigma^2$ en verder geldt dat $X_i = \mu + V_i$. We kunnen zeggen dat X_i een meting is van μ , met meetfout V_i . Er is geen systematische fout in de meting, in die zin dat $EV_i = 0$. De nauwkeurigheid van de meting wordt weergegeven door de variantie van de meetfout. Als σ^2 groot is hebben we tamelijk onnauwkeurige metingen. Merk nu op dat $\bar{X} = \mu + \bar{V}$, waarbij $\bar{V} = (1/n) \sum_{i=1}^n V_i$. Dus \bar{X} meet μ met meetfout \bar{V} . De onnauwkeurigheid is kleiner geworden dan die van de individuele metingen X_i , want $\text{var}(\bar{V}) = \sigma^2/n$. De onnauwkeurigheid gaat naar nul als we steeds meer metingen verrichten.

SPECIAAL GEVAL: ALTERNATIEVE VERDELING.

Bekijk n o.o. experimenten, waarbij bij de individuele uitkomsten het kenmerk S al dan niet optreedt. Noem $\{X_i = 1\}$ de gebeurtenis dat kenmerk S gevonden wordt in het i -de experiment en $\{X_i = 0\}$ als dit niet het geval is, $i = 1, \dots, n$. Laat $p = P(X_i = 1)$ de kans op kenmerk S zijn. M.a.w. X_1, \dots, X_n zijn o.o. alternatief verdeeld met waarden in $\{0, 1\}$ en met succeskans p . De frequentie van kenmerk S is

$$\begin{aligned} f_q(S) &= \frac{\{\text{aantal keren dat } S \text{ optreedt}\}}{\{\text{aantal experimenten}\}} \\ &= \frac{\sum_{i=1}^n X_i}{n} = \bar{X}. \end{aligned}$$

Nu is

$$E(X_i) = 1 \times P(X_i = 1) + 0 \times P(X_i = 0) = P(X_i = 1) = p.$$

Volgens de wet van de grote aantallen geldt daarom:

$$\lim_{n \rightarrow \infty} P(|f_q(S) - p| > c) = 0,$$

voor alle $c > 0$. Dit zegt dat de frequentie van een gebeurtenis met grote kans in de buurt van de kans op die gebeurtenis ligt, als tenminste het aantal experimenten maar groot genoeg is.

1.12. De centrale limietstelling

Volgens de wet van de grote aantallen is met grote kans $\bar{X} \approx \mu$, en de afwijking $|\bar{X} - \mu|$ is i.h.a. kleiner als \bar{X} op meer experimenten gebaseerd is. Men kan nooit exact zeggen hoe groot $|\bar{X} - \mu|$ precies is, doordat \bar{X} een s.g. is waarvan de uitkomst onzeker is. Maar men kan wel een kansuitspraak over de afwijking doen. De wet van de grote aantallen is een vrij grove kansuitspraak. De centrale limietstelling geeft een benadering voor kansen van de vorm $P(|\bar{X} - \mu| > c)$. De stelling zegt dat \bar{X} ongeveer normaal verdeeld is. We zullen \bar{X} eerst standaardiseren (d.w.z. we trekken er de verwachting van af en delen door de standaarddeviatie) zodat het resultaat verwachting 0 en variantie 1 heeft. De centrale limietstelling beweert dat de gestandaardiseerde \bar{X} ongeveer standaardnormaal verdeeld is.

Centrale limietstelling. Laat X_1, \dots, X_n o.o. stochastische grootheden zijn die alle dezelfde verdeling hebben, met verwachting $\mu = E(X_i)$ en variantie $\sigma^2 = \text{var}(X_i)$, $i = 1, \dots, n$. Dan voor alle t

$$\lim_{n \rightarrow \infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq t\right) = \Phi(t).$$

BEWIJS laten we achterwege. \square

Onder de voorwaarden van bovenstaande stelling heeft \bar{X} verwachting μ en variantie σ^2/n (zie de vorige paragraaf). De standaarddeviatie van \bar{X} is dan σ/\sqrt{n} . Hieruit volgt dat $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ verwachting 0 en variantie 1 heeft. Het ligt voor de hand dat de benadering van verdeling van de gestandaardiseerde van \bar{X} ook verwachting 0 en variantie 1 moet hebben. In de centrale limietstelling benaderen we de verdelingsfunctie van de gestandaardiseerde van \bar{X} met de standaardnormale verdelingsfunctie Φ . De laatste is getabelleerd, zodat we nu specifieke kansen inderdaad kunnen uitrekenen.

Merk op dat er verschillende schrijfwijzen zijn:

$$\frac{\bar{X} - \mu}{\sigma\sqrt{n}} = \sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}.$$

Nu heeft $\sum_{i=1}^n X_i$ verwachting $n\mu$ en variantie $n\sigma^2$, zodat de laatste uitdrukking gezien kan worden als de gestandaardiseerde van $\sum_{i=1}^n X_i$. Het maakt natuurlijk niet uit of men eerst het gemiddelde neemt en dan deze standaardiseert of dat men $\sum_{i=1}^n X_i$ rechtstreeks standaardiseert.

In praktijk komt het er op neer dat \bar{X} ongeveer $N(\mu, \sigma^2/n)$ -verdeeld is, ofwel dat $\sum_{i=1}^n X_i$ ongeveer $N(n\mu, n\sigma^2)$ -verdeeld is. We schrijven

$$P(\bar{X} \leq t) \approx \Phi\left(\sqrt{n}\left(\frac{t - \mu}{\sigma}\right)\right),$$

en

$$P\left(\sum_{i=1}^n X_i \leq t\right) \approx \Phi\left(\frac{t - n\mu}{\sqrt{n}\sigma}\right).$$

De centrale limietstelling maakt deze uitspraak wiskundig precies. Bovendien moeten we altijd standaardiseren om de tabellen van de $N(0, 1)$ -verdeling te kunnen gebruiken. De stelling maakt duidelijk waarom de normale verdeling zo'n belangrijke verdeling is: wat de verdeling van X_i , $i = 1, \dots, n$ ook is, in de limiet 'vergeet' \bar{X} waar ie vandaan komt.

EIGENSCHAP VAN DE NORMALE VERDELING.

Laat Y_i een s.g. zijn die $N(\mu_i, \sigma_i^2)$ -verdeeld is, $i = 1, \dots, n$. Het blijkt dat dan iedere *lineaire combinatie* van Y_1, \dots, Y_n ook weer normaal verdeeld is. (Dit is een eigenschap die voor andere verdelingen niet hoeft te gelden, b.v. de som van twee exponentieel verdeelde stochastische grootheden is niet exponentieel verdeeld.) De normale verdeling wordt bovendien volledig beschreven door de verwachting en variantie: als men weet dat een s.g. normaal verdeeld is en men weet ook de verwachting en variantie, dan kent men de verdeling. (Dit is ook een eigenschap die voor andere verdelingen niet hoeft te gelden. We zijn er

echter geen voorbeelden van tegengekomen. Men kan zich misschien wel voorstellen dat er verdelingen zijn die door 3 of zelfs meer parameters worden beschreven.) Nu is het voor een lineaire combinatie niet moeilijk om verwachting en variantie uit te drukken in verwachtingen en (co)varianties van de oorspronkelijke variabelen. Er geldt voor de lineaire combinatie $a_0 + a_1Y_1 + \dots + a_nY_n$ (met a_0, \dots, a_n getallen)

$$E(a_0 + a_1Y_1 + \dots + a_nY_n) = a_0 + a_1\mu_1 + \dots + a_n\mu_n.$$

Stel nu voor het gemak dat Y_1, \dots, Y_n o.o. zijn, zodat we niet met covarianties te maken hebben. Dan geldt

$$\text{var}(a_0 + a_1Y_1 + \dots + a_nY_n) = a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2.$$

Zo vinden we dat als Y_1, \dots, Y_n o.o. normaal verdeeld zijn, dan is $a_0 + \sum_{i=1}^n a_iY_i$ normaal verdeeld met verwachting $(a_0 + \sum_{i=1}^n a_i\mu_i)$, en variantie $\sum_{i=1}^n a_i^2\sigma_i^2$. Een speciaal geval is: als X_1, \dots, X_n o.o. $N(\mu, \sigma^2)$ -verdeeld zijn, dan $\bar{X} \sim N(\mu, \sigma^2/n)$.

In het geval van normaal verdeelde stochastische grootheden is \bar{X} dus exact normaal verdeeld en in alle andere gevallen spreken we van een asymptotisch normale verdeling.

SPECIAAL GEVAL: ALTERNATIEVE VERDELING.

Laat X_1, \dots, X_n o.o. zijn met $p = P(X_i = 1) = 1 - P(X_i = 0)$. Dit is meestal een codering van het al of niet optreden van een bepaald kenmerk S , en $f_q(S) = \bar{X}$ is dan de frequentie van S . De verwachting van X_i is p . De variantie is

$$\text{var}(X_i) = EX_i^2 - p^2 = p - p^2 = p(1 - p).$$

Dus $E(\bar{X}) = p$ en $\text{var}(\bar{X}) = p(1 - p)/n$. Volgens de centrale limietstelling is nu $f_q(S)$ ongeveer $N(p, p(1 - p)/n)$ -verdeeld. Noem $X = \sum_{i=1}^n X_i$. Dit is het aantal keren dat kenmerk S optreedt. Dan is X ongeveer $N(np, np(1 - p))$ -verdeeld, ofwel

$$\frac{X - np}{\sqrt{np(1 - p)}}$$

is ongeveer standaard normaal verdeeld. De exacte verdeling van X weten we ook, dit is n.l. de binomiale verdeling met parameters n en p :

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n.$$

De centrale limietstelling zegt in dit geval dat

$$\sum_{k \leq t} \binom{n}{k} p^k (1 - p)^{n-k} \approx \Phi \left(\frac{t - np}{\sqrt{np(1 - p)}} \right).$$

Men zou zoiets ook analytisch kunnen bewijzen, maar het ziet er niet eenvoudig uit.

Vb. We nemen $n = 20$. Voor deze waarde van het aantal experimenten is de binomiale verdeling nog getabelleerd, omdat de benadering met de normale verdeling niet goed genoeg is. Laten we eens zien wat het verschil is voor $p = 0.40$ en $t = 5$. Uit een tabel halen we dat

$$P(X \leq 5) = 0.1256.$$

Men kan dit narekenen:

$$0.1256 = \sum_{k=0}^5 \binom{20}{k} (0.40)^k (0.60)^{20-k}.$$

Verder

$$\begin{aligned} \Phi\left(\frac{t - np}{\sqrt{np(1-p)}}\right) &= \Phi\left(\frac{5 - (20)(0.40)}{\sqrt{(20)(0.40)(0.60)}}\right) \\ &= \Phi(-1.37) = 1 - \Phi(1.37) = 1 - 0.9147 = 0.0853. \end{aligned}$$

Vergelijk deze uitkomst met het exacte resultaat 0.1256. De benadering is dus niet zo best.

CONTINUÏTEITSCORRECTIE. Als X een binomiale verdeling met parameters n en p bezit, dan kan X alleen de waarden $0, 1, \dots, n$ aannemen. Het is een beter om bij de benadering van zo'n discrete s.g. met de continue normale verdeling een continuïteitscorrectie toe te passen, m.n. als n klein is. Deze correctie is:

$$P(X = t) \approx \Phi\left(\frac{t - np - 1/2}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{t - np + 1/2}{\sqrt{np(1-p)}}\right),$$

voor $t \in \{0, 1, \dots, n\}$. In woorden: $P(X = t)$ benaderen we met de kans dat een $N(np, np(1-p))$ -verdeelde s.g. in het interval $[t - 1/2, t + 1/2]$ ligt.

$P(X \leq t)$ benaderen we dan met de kans dat een $N(np, np(1-p))$ -verdeelde s.g. in het interval $(-\infty, t + 1/2]$ ligt:

$$P(X \leq t) \approx \Phi\left(\frac{t - np + 1/2}{\sqrt{np(1-p)}}\right), \quad t \in \{0, 1, \dots, n\}.$$

Vb. Neem weer $n = 20$ en $p = 0.40$. Dan

$$P(X \leq 5) = 0.1256.$$

Gebruiken we de benadering met continuïteitscorrectie, dan vinden we

$$\Phi\left(\frac{5 - (20)(0.40) + 1/2}{\sqrt{(20)(0.40)(0.60)}}\right) = \Phi(-1.14) = 1 - \Phi(1.14) = 1 - 0.8709 = 0.1291.$$

Dit is inderdaad een verbetering. Bekijk ook $P(X = 8) = 0.1797$. Ga na dat de benadering is $\Phi(0.22) - \Phi(-0.22) = 0.1820$.

2. Mathematische statistiek

2.1. Inleiding

Een producent laat tien weken lang elke werkdag een batterij van de productie van die dag controleren. Zo worden gegevens x_1, \dots, x_n verkregen, met x_i de levensduur van batterij i , waarbij $n = 50$ het aantal controle-dagen is. De producent streeft ernaar dat de batterijen minstens acht uur meegaan. Door variatie in het productieproces e.d. hebben de batterijen niet alle dezelfde levensduur. Hoe kan men nu de kwaliteit van de batterijen beschrijven, en in welke zin geven de waarnemingen x_1, \dots, x_{50} informatie over de kwaliteit? Voor we deze vragen trachten te beantwoorden, zullen we eerst de situatie vertalen in statistische termen. De beschrijving van de situatie laat zich min of meer samenvatten in de punten

(i) t/m (iv):

(i) De levensduur varieert (op niet voorspelbare manier),

(ii) Levensduren van verschillende batterijen beïnvloeden elkaar niet,

(iii) Het productieproces is op iedere dag ongeveer hetzelfde, en er zijn daarom geen systematische verschillen tussen batterijen te verwachten,

(iv) Er is eventueel specialistische voorkennis over de levensduur van een batterij (b.v. iedere batterij die het doet gaat zeker 2 uur mee).

Een vertaling van bovenstaande assumpties in statistische termen is:

(i) De waargenomen levensduren x_1, \dots, x_n zijn *realisaties* van stochastische grootheden X_1, \dots, X_n ,

(ii) De s.g.ⁿ X_1, \dots, X_n zijn onderling onafhankelijk,

(iii) X_1, \dots, X_n hebben alle dezelfde verdeling (ofwel, ze hebben alle dezelfde verdelingsfunctie $F(x) = P(X_1 \leq x) = \dots = P(X_n \leq x)$),

(iv) Er zijn eventueel zekere aspecten van $F(x)$ bekend (b.v. $F(x) = 0$ voor $0 < x < 2$).

We noemen de statistische vertaling (i) t/m (iv) het *model*.

Hoe meten we nu de kwaliteit van een batterij? We stellen voor dit te doen door naar de verdeling van de X_i te kijken, en het begrip kwaliteit samen te vatten in een getal dat een eigenschap van die verdeling beschrijft. We zouden bijvoorbeeld kunnen zeggen dat p de kwaliteit is, waarbij $p = P(X_i > 8)$, $i = 1, \dots, n$. In woorden: de kwaliteit van een batterij is de kans dat ie langer dan acht uur meegaat. Omdat we veronderstellen dat alle batterijen dezelfde verdeling hebben is p niet alleen de kwaliteit van één specifieke batterij, maar het geeft ook de kwaliteit van het productieproces weer. Hoe groter p , des te beter is het productieproces. We willen nu een uitspraak doen over p op grond van de waarnemingen x_1, \dots, x_n . Iedere waarneming zegt iets over p , want het zijn alle realisaties van dezelfde verdeling. Iedere waarneming geeft een brokje informatie, en alle waarnemingen samen geeft ons een idee van p . Dit idee kan op de volgende manier verkregen worden. Noem

$$Y_i = \begin{cases} 1, & \text{als } X_i > 8 \\ 0, & \text{anders} . \end{cases}$$

Dan $P(Y_i = 1) = 1 - P(Y_i = 0) = p$, $i = 1, \dots, n$. M.a.w. Y_i , bezit een alternatieve verdeling met succeskans p . Hieruit volgt dat het aantal batterijen dat langer dan acht uur mee gaat (dit is $\sum_{i=1}^n Y_i$) een binomiale verdeling bezit met parameters n en p . Veronderstel dat we bij die 50 geteste batterijen er 40 vinden die langer dan acht uur meegaan. Dan

hebben we de realisatie $\sum_{i=1}^n y_i = 40$ van $\sum_{i=1}^n Y_i$, en verder $\bar{y} = (1/n) \sum_{i=1}^n y_i = 0.80$. D.w.z. 80 % gaat langer dan acht uur mee. Volgens de wet van de grote aantallen ligt $\bar{Y} = (1/n) \sum_{i=1}^n Y_i$ met grote kans in de buurt van p als n groot genoeg is. We noemen nu \bar{Y} een *schatter* van p en \bar{y} een *schatting*. De waarde van p komen we niet te weten, maar we hebben het idee dat p wel in de buurt van de 0.80 zal liggen.

Een andere manier om de kwaliteit te meten is door naar de verwachte waarde van een levensduur X_i te kijken. Noem $\mu = E(X_i)$, $i = 1, \dots, n$. Er vanuit gaand dat we inderdaad X_i kunnen observeren (sommige batterijen gaan misschien wel drie jaar mee, terwijl men de uitslag van de controle binnen drie maanden wil rapporteren), kunnen we μ schatten met $\bar{X} = (1/n) \sum_{i=1}^n X_i$. Volgens de wet van de grote aantallen is dit zinnige schatter. Om precies te weten met welke kans \bar{X} in de buurt van μ ligt, zouden we iets over de verdeling van \bar{X} moeten weten (d.w.z. iets specificeren bij assumptie (iv)). (Merk op dat in het vorige geval we de verdeling van \bar{Y} , afgezien van de onbekende p , wel kennen.) Het blijkt dat dankzij de centrale limietstelling het volgende recept toegepast kan worden: noem

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

dan ligt μ met ongeveer 95 % kans in het interval

$$\bar{X} \pm 2\sqrt{S^2/n}.$$

We hebben realisaties x_1, \dots, x_n van X_1, \dots, X_n en dus ook een realisatie s^2 van S^2 . In praktijk vul je deze getallen in in bovenstaand recept. In de komende paragrafen zullen we de theorie ontwikkelen die tot dit recept leidt.

De statistische theorie voor een bepaald probleem hangt sterk af van de vraagstelling. In bovenstaand voorbeeld is de vraag of men p of μ wil schatten (of beide) alleen te beantwoorden door de producent. Misschien wil hij/zij wel de hele verdelingsfunctie $F(x)$ schatten, en achteraf eens kijken of er aan de schatter interessante eigenschappen te ontdekken zijn. Door alleen naar p te kijken schat men in feite alleen $F(8)$, want $p = 1 - F(8)$. Andere waarden behalve $x = 8$ zijn misschien even belangrijk. Verder is er nog de keuze voor het soort uitspraak dat men uiteindelijk wil doen. Het kan zijn dat men zo goed mogelijk achter de waarde van p (of μ etc.) wil komen. Dan gebruikt men de z.g. *schattingstheorie*. Het is ook mogelijk dat men b.v. wil weten of $p > 1/2$ (of $\mu > 8$ etc.). In dat geval komt men terecht bij de *toetsingstheorie*. Ook hier kunnen we een vraag niet met zekerheid beantwoorden. De uitspraak is eerder van de vorm: er is reden om te beslissen vóór $p > 1/2$ en met hoogstens 5% kans is dit een foute beslissing.

In een kansmodel zijn de kansen min of meer volledig gespecificeerd. In een statistisch model is dat nooit het geval, en moet men op grond van waarnemingen proberen iets over de kansen te weten te komen. Formeel gesproken beschouwen we een vector (X_1, \dots, X_n) van observeerbare stochastische grootheden, en het statistische model is dat de verdeling van (X_1, \dots, X_n) behoort tot een zekere *klasse* van verdelingen \mathcal{P} . Zo'n klasse legt algemene karakteristieken vast. Een voorbeeld van een aanname (die we in dit college meestal zullen maken) is dat X_1, \dots, X_n o.o. zijn. Met de onafhankelijkheidsveronderstelling beperk je de klasse van mogelijke verdelingen \mathcal{P} . Een andere aanname kan zijn dat de waarnemingen

komen uit de normale verdeling met verwachting μ en variantie σ^2 , met μ en σ^2 niet gespecificeerd (d.w.z. onbekend). Zo'n veronderstelling in het algemeen geformuleerd luidt: X_1, \dots, X_n komen uit een verdeling die bekend is op een parameter θ na. Door θ te variëren beschrijf je zo een klasse van verdelingen. Het kan zijn dat er diverse onbekende parameters zijn. Dan is θ in feite een vector (b.v. θ is het paar (μ, σ^2) bij de normale verdeling). In latere paragrafen zal θ ook vaak de verdeling niet volledig vastleggen. Dan stoppen we in θ alleen de parameters die we interessant vinden, en alle overige onbekende parameters laten we op de achtergrond (b.v. bij de normale verdeling kan het zijn dat we alleen μ willen onderzoeken en σ^2 niet).

ASELECTE GETALLEN.

Een producent laat elke dag een batterij testen. De vraag is: welke? Laten we zeggen dat er iedere dag 20 batterijen worden gemaakt. Als je steeds de eerste van de band test kan er een scheef beeld ontstaan doordat b.v. de eerste batterij van de dag er last van heeft dat de machines nog niet warmgedraaid zijn. Om dergelijke systematische fouten te vermijden is het zaak blindelings één van de 20 batterijen te kiezen. In principe kan dit door ze aan het eind van de dag alle in een bak te gooien, eens goed te roeren, en dan er een uit te graaien. Maar het is makkelijker om zo'n randomisatie-procedure te simuleren. Dit gaat als volgt. Nummer de batterijen van 1 t/m 20. Het maakt niet uit hoe je dat doet, de nummering is slechts om de batterijen een naam te geven. Kies uit een tabel van aselecte getallen het eerste getal tussen 1 en 20, zeg dat dit k is. Test dan de k -de batterij. De volgende dag zou men b.v. het eerstvolgende getal tussen 1 en 20 uit de tabel kunnen nemen, etc..

De computer genereert z.g. aselecte binaire getallen. Dat zijn rijtjes bestaande uit nullen en énen, en theoretisch moet de kans op een nul of één voor iedere digit $1/2$ zijn. Maar een computer kan geen muntjes opgooien. Bovendien hebben we alleen een abstract kansbegrip, en niemand heeft ooit in zijn of haar leven een echte kans gezien! Een computer berekent binaire getallen volgens deterministische regels. Deze getallen worden vervolgens aan allerlei statistische toetsen onderworpen om te zien of ze met goed fatsoen voor aselecte getallen kunnen doorgaan.

Aselecte getallen worden bij veel proefopzetten gebruikt. In paragraaf 1.4 bespraken we al de steekproefcontrole, waarbij een partij goederen steekproefsgewijs, al of niet met terugleggen, wordt getest. Dit wordt in veel bedrijven toegepast, b.v. bij de controle van boekhoudkundige posten, of bij het bepalen van het verzendtarief van een partij van een postorderbedrijf. Bij de steekproefcontrole zijn de waarnemingen alleen stochastisch door het aselecte kiezen. We zullen ook vaak situaties tegenkomen waar de stochastiek meer rechtstreeks uit de aard der waarnemingen komt. Men kan b.v. de hoogwaterstanden van de afgelopen 10 jaar als realisaties van stochastische grootheden zien, hoewel we ze niet aselect gekozen hebben.

2.2. Schattingstheorie

TERMINOLOGIE. We beschouwen doorgaans een rij X_1, \dots, X_n van o.o. s.g.² met dezelfde verdeling, d.w.z.

$$P(X_1 \leq x) = P(X_2 \leq x) = \dots = P(X_n \leq x) = F(x), \text{ voor alle } x.$$

We zeggen dan dat X_1, \dots, X_n o.o. en *identiek* verdeeld zijn, en we noemen X_1, \dots, X_n een

steekproef uit de verdeling F . Een realisatie van (X_1, \dots, X_n) noteren we met (x_1, \dots, x_n) . Dit zijn de getallen die we hebben waargenomen nadat de steekproef daadwerkelijk is uitgevoerd.

De verdelingsfunctie $F(x)$ is geheel of gedeeltelijk onbekend. We nemen vaak iets aan over de vorm van $F(x)$. Dit is soms voor het wiskundig gemak, maar het kan ook b.v. zijn dat we het waardebereik van de X_i kennen, of iets anders over de verdeling van de X_i . Als b.v. $X_i \in \{0, 1\}$, dan weten we dat X_i een alternatieve verdeling bezit. De succeskans $p = P(X_i = 1)$ zullen we in het algemeen niet kennen, en we zeggen dan dat X_i een alternatieve verdeling met parameter p bezit. Een ander voorbeeld is dat we op grond van een redentatie als in Vb. 3 in Paragraaf 1.7.2 veronderstellen dat X_i een exponentiële verdeling bezit met parameter λ . Wij zullen vaak aannemen dat X_i normaal verdeeld is met parameters μ en σ^2 . Dit is dan meestal alleen voor het wiskundig gemak. In Paragraaf 2.3 zullen we de situatie bekijken waarbij we zo goed als niets over $F(x)$ veronderstellen. Het is vaak interessant om de onbekende verdelingsfunctie te schatten. In deze paragraaf gaan we in op het geval dat de onbekende verdeling afhangt van een onbekende parameter θ .

Vb. Stel X_1, \dots, X_n zijn o.o. en identiek verdeeld met verwachting $EX_i = \theta$. We nemen verder niets aan over de verdeling van de X_i . De parameter θ legt de verdeling van de X_i niet vast. Toch kunnen we wel iets zeggen. Bekijk n.l. $\bar{X} = (1/n) \sum_{i=1}^n X_i$. Er geldt dat $E\bar{X} = \theta$ (men noemt \bar{X} daarom een *zuivere* schatter van θ) en bovendien is de variantie van \bar{X} klein als n groot is. Dat is een plezierige eigenschap want het betekent dat \bar{X} zich voor grote n sterk rond θ concentreert. De wet van de grote aantallen en de centrale limietstelling geven nog wat extra benaderingen van het gedrag van \bar{X} als schatter van θ . Hoe klein de variantie van \bar{X} is weten we niet, want we kennen de variantie van de X_i niet. Daarom wordt ook vaak naar een schatter van deze variantie gezocht. Dat is m.a.w. om te schatten hoe goed de schatter is.

Stel X_1, \dots, X_n hebben een verdeling die (ondermeer) van een onbekende parameter θ afhangt. We veronderstellen hier dat θ een *getal* is (geen vector), en het representeert de grootheid die interessant geacht wordt.

DEFINITIE. Een *schatter* is een functie

$$T = t(X_1, \dots, X_n),$$

van X_1, \dots, X_n , die niet afhangt van θ of andere onbekende grootheden.

De reden waarom we eisen dat de functie T niet van onbekende grootheden mag afhangen, is dat we T in praktijk moeten kunnen uitrekenen. D.w.z. als we waarnemingen X_1, \dots, X_n hebben, dan is T ook waar te nemen. Bij realisaties x_1, \dots, x_n noemen we de realisatie $t = t(x_1, \dots, x_n)$ een schatting. Een schatting is dus een getal dat we berekenen uit de realisaties van x_1, \dots, x_n .

Vb. $T = \bar{X}$ of $T = X_1 + X_2^2$ zijn schatters, maar de functie $(X_1 + \theta)/2$ is géén schatter als θ onbekend is.

Een schatter hangt als functie van de waarnemingen dus niet van onbekende grootheden af, maar de *verdeling* van een schatter hangt meestal wel van onbekende grootheden af. Dit komt doordat de verdeling van de waarnemingen zelf van onbekende grootheden afhangt, m.n. van de onbekende parameter θ . In het bijzonder hangt de verwachting en variantie van een schatter T af van θ . Om dit te benadrukken schrijven we soms $E_\theta T$ voor de verwachting van T als de parameterwaarde θ is, en analoog: $\text{var}_\theta(T)$, $P_\theta(T > 2)$, etc.. Nu gaat het erom schatters te vinden die een idee geven van de parameterwaarde θ . M.a.w. we zoeken schatters die volgens een bepaald criterium *goed* zijn. Een voorbeeld van zo'n criterium is:

DEFINITIE. De verwachte kwadratische fout van een schatter T is

$$\text{MSE}_\theta(T) = E_\theta(T - \theta)^2$$

(MSE komt van het engelse begrip Mean Square Error).

DEFINITIE. De *onzuiverheid* van een schatter T is

$$\text{bias}_\theta(T) = E_\theta(T) - \theta$$

(bias is het engelse woord voor onzuiverheid). We noemen een schatter T *zuiver* (engels: *unbiased*) als

$$E_\theta(T) = \theta$$

voor *alle* mogelijke waarden van θ .

Een schatter moet bij voorkeur een kleine MSE hebben, en we willen ook graag dat een schatter zuiver is, of op z'n minst kleine onzuiverheid heeft. De relatie tussen de MSE en de bias is gegeven in het volgende

Lemma. $\text{MSE}_\theta(T) = \text{var}_\theta(T) + \text{bias}_\theta^2(T)$.

BEWIJS.

$$\begin{aligned} \text{MSE}_\theta(T) &= E_\theta(T - \theta)^2 \\ &= E_\theta((T - E_\theta T) + (E_\theta T - \theta))^2 \\ &= E_\theta(T - E_\theta T)^2 + (E_\theta T - \theta)^2 + 2E_\theta(T - E_\theta T)(E_\theta T - \theta) \\ &= \text{var}_\theta(T) + \text{bias}_\theta^2(T) + 0. \end{aligned}$$

□

Voor een zuivere schatter T is $\text{MSE}_\theta(T) = \text{var}_\theta(T)$. M.a.w. een zuivere schatter is *goed* als deze kleine variantie heeft. Een onzuivere schatter met kleine variantie kan onbruikbaar zijn, want zo'n schatter concentreert zich rond het verkeerde punt. Soms moet men een afweging maken: aan de éne kant wil men graag een zuivere schatter hebben en aan de andere kant wil men ook de variantie klein houden. Dit kunnen strijdige belangen zijn. Vaak houdt men vast aan de eis dat een schatter zuiver moet zijn, en zoekt men onder alle zuivere schatters diegene met de kleinste variantie. Dit kan twee problemen geven. Ten

eerste hoeft een zuivere schatter niet te bestaan. We eisen dan n.l. dat $E_\theta(T) = \theta$ voor alle mogelijke waarden van θ , en dit geldt soms voor geen enkele T . Ten tweede hangt de variantie van een schatter i.h.a. ook van de onbekende θ af, dus het kan gebeuren dat T voor zekere waarden van θ kleine variantie heeft, maar voor andere waarden juist grote variantie. Omdat we θ niet kennen weten we dus ook niet of de schatter zich nu wel of niet behoorlijk gedraagt. Toch komen er situaties voor waarbij men kan bewijzen dat een zuivere schatter de kleinste variantie onder alle zuivere schatters heeft, voor alle mogelijke waarden van θ .

DEFINITIE. T heet een *meest nauwkeurige* zuivere schatter als T zuiver is en als voor iedere andere zuivere schatter T^* geldt dat $\text{var}_\theta(T^*) \geq \text{var}_\theta(T)$ voor alle mogelijke waarden van θ . We noemen T dan ook wel een zuivere schatter met *minimale variantie*.

Het kan gebeuren dat er geen zuivere schatter bestaat, en als er al een bestaat, dan kan het nog gebeuren dat er geen zuivere schatter met minimale variantie bestaat. Dit is een gevolg van het feit dat de eisen betrekking hebben op alle mogelijke waarden van θ .

Er is een uitgebreide theorie over de existentie en constructie van zuivere schatters met minimale variantie. We gaan daar in dit college niet op in. We zullen wel af en toe roepen dat een bepaalde zuivere schatter het meest nauwkeurig is, zonder dit te bewijzen. Met het gereedschap wat we nu hebben kunnen we in ieder geval altijd nagaan of een schatter al of niet zuiver is, en als er diverse zuivere schatters zijn, kunnen we hun gedrag vergelijken door de varianties uit te rekenen. In het algemeen kan men schatters, al of niet zuiver, vergelijken door naar de MSE te kijken.

Hoog tijd dat we eens wat voorbeelden doorwerken. We beginnen met een heel eenvoudig geval, om het idee te verduidelijken.

Vb. Laat X_1, \dots, X_n o.o. zijn met alle verwachting μ en variantie σ^2 . Dan is \bar{X} een zuivere schatter van μ , want $E_\mu \bar{X} = \mu$, voor alle waarden van μ . De variantie van deze schatter is $\text{var}(\bar{X}) = \sigma^2/n$. Laten we dit eens vergelijken met de variantie van een andere zuivere schatter. De eerste waarneming X_1 heeft ook verwachting μ (en evenzo voor de andere $(n-1)$ waarnemingen). X_1 is m.a.w. een zuiver schatter van μ , met variantie $\text{var}(X_1) = \sigma^2$. We zien dat $\text{var}(\bar{X}) \leq \text{var}(X_1)$ (met gelijkheid dan en slechts dan als er maar één waarneming is). \bar{X} is daarom een betere schatter dan X_1 . Dit is ook intuïtief duidelijk, omdat \bar{X} gebruik maakt van alle waarnemingen, ofwel van alle informatie die je hebt, terwijl het gebruik van X_1 er op neer komt dat men alle overige waarnemingen weggooit. De waarnemingen schommelen rond de verwachting μ en over het algemeen middelt \bar{X} de schommeling rond μ er voor een deel uit.

Vb. Laat X_1, \dots, X_n o.o. homogeen verdeeld zijn op het interval $[0, \theta]$ met $\theta > 0$ onbekend. Het ligt voor de hand de schatter $T_1 := \max(X_1, \dots, X_n)$ te nemen. Het is duidelijk dat $T_1 \leq \theta$. Dit doet al vermoeden dat de schatter T_1 wel te klein zal uitvallen. Inderdaad, de bias van T_1 is negatief. Om dit te bewijzen gaan we eerst de dichtheid van T_1 afleiden. De verdelingsfunctie is:

$$\begin{aligned} F_{T_1}(t) &= P(T_1 \leq t) = P(\max(X_1, \dots, X_n) \leq t) \\ &= \left(\frac{t}{\theta}\right)^n = \left(\frac{1}{\theta}\right)^n t^n, \quad 0 \leq t \leq \theta. \end{aligned}$$

Door de afgeleide te nemen vind je de dichtheid:

$$f_{T_1}(t) = \left(\frac{1}{\theta}\right)^n n t^{n-1}, \quad 0 \leq t \leq \theta.$$

De verwachting van T_1 is:

$$E_{\theta}T_1 = \int_0^{\theta} t f_{T_1}(t) dt = \frac{n}{n+1}\theta.$$

M.a.w. T_1 ligt in het algemeen iets teveel naar links, maar als n groot is is de bias klein:

$$\text{bias}_{\theta}(T_1) = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta.$$

Het is nu niet moeilijk om een zuivere schatter voor θ te verzinnen. Neem $T_2 := ((n+1)/n)T_1$. Dit verschuift T_1 een beetje naar rechts. T_2 is een zuivere schatter:

$$E_{\theta}(T_2) = \frac{n+1}{n}E_{\theta}T_1 = \theta,$$

voor alle mogelijke waarden van θ . Laten we de variantie van T_2 uitrekenen. Er geldt:

$$E_{\theta}T_2^2 = \left(\frac{n+1}{n}\right)^2 E_{\theta}T_1^2 = \left(\frac{n+1}{n}\right)^2 \int_0^{\theta} t^2 f_{T_1}(t) dt = \frac{(n+1)^2}{n(n+2)}\theta^2.$$

Dus

$$\text{var}_{\theta}(T_2) = \frac{(n+1)^2}{n(n+2)}\theta^2 - \theta^2 = \frac{1}{n(n+2)}\theta^2.$$

Dit resultaat gaan we nu vergelijken met de variantie van een andere zuivere schatter, n.l. $T_3 = 2\bar{X}$. Omdat $E_{\theta}X_i = (1/2)\theta$, $i = 1, \dots, n$ is T_3 inderdaad zuiver. De variantie van X_i is

$$\text{var}_{\theta}(X_i) = \frac{\theta^2}{12}.$$

Hieruit volgt dat

$$\text{var}_{\theta}(\bar{X}) = \frac{\theta^2}{12n},$$

en

$$\text{var}_{\theta}(T_3) = 4\text{var}_{\theta}(\bar{X}) = \frac{\theta^2}{3n}.$$

We zien dat

$$\text{var}_{\theta}(T_2) \leq \text{var}_{\theta}(T_3),$$

voor alle θ , met gelijkheid alleen in het geval $n = 1$. Als $n = 1$ zijn T_2 en T_3 dezelfde schatter, wat verklaart waarom hun varianties dan gelijk zijn.

HET SCHATTEN VAN EEN VERWACHTING.

Als X_1, \dots, X_n o.o. zijn met verwachting μ en variantie σ^2 , dan is \bar{X} een zuivere schatter van μ met $\text{var}(\bar{X}) = \sigma^2/n$. Dit geldt dus ook in de volgende speciale gevallen:

(a) $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$ en o.o.. Dan is \bar{X} een zuivere schatter van μ met variantie $\text{var}_{(\mu, \sigma^2)}(\bar{X}) = \sigma^2/n$.

(b) X_1, \dots, X_n o.o., met $P(X_i = 1) = 1 - P(X_i = 0) = p$, $i = 1, \dots, n$. Dan is \bar{X} een zuivere schatter van p . Verder: $\sigma^2 = p(1 - p)$, dus $\text{var}(\bar{X}) = p(1 - p)/n$.

(c) X_1, \dots, X_n o.o. Poisson verdeeld met parameter μ . Dan is \bar{X} een zuiver schatter van μ met $\text{var}_{\mu}(\bar{X}) = \mu/n$.

(d) X_1, \dots, X_n o.o. exponentieel verdeeld met parameter λ . Dan is \bar{X} een zuivere schatter van $1/\lambda$ met $\text{var}_{\lambda}(\bar{X}) = 1/(n\lambda^2)$. Maar $1/\bar{X}$ is géén zuivere schatter van λ !

In bovenstaande gevallen (a) t/m (d) is \bar{X} ook de meest nauwkeurige zuivere schatter van de verwachting. We hebben echter gezien dat als X_1, \dots, X_n o.o. homogeen verdeeld zijn op $[0, \theta]$, dan is \bar{X} een zuivere schatter van $(1/2)\theta$, maar niet de meest nauwkeurige.

HET SCHATTEN VAN DE VARIANTIE.

Laat X_1, \dots, X_n o.o. zijn met verwachting μ en variantie σ^2 . Als μ bekend zou zijn is

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

een zuivere schatter van σ^2 . Immers

$$\begin{aligned} E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right) &= \frac{1}{n} \sum_{i=1}^n E(X_i - \mu)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \sigma^2 = \sigma^2. \end{aligned}$$

Maar in praktijk is μ meestal niet bekend. In dat geval vervang je in bovenstaande schatter μ door \bar{X} en $1/n$ door $1/(n-1)$. De resulterende schatter noemen we

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

Merk op dat we S^2 ook kunnen schrijven als

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

want

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

$$= \sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2.$$

Deze schrijfwijze is handig bij het bewijs dat S^2 een zuivere schatter van σ^2 is:

$$ES^2 = \frac{1}{n-1} \left(\sum_{i=1}^n EX_i^2 - nE\bar{X}^2 \right) = \frac{1}{n-1} \left(\sum_{i=1}^n (\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2) \right) = \sigma^2.$$

Bij dit soort berekeningen is de volgende truuk ook handig:

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

Nu staan de s.g.ⁿ in afwijking van de verwachting. Daarom mag je hier zonder verlies van algemeenheid veronderstellen dat ze verwachting nul hebben. Dit verklaart ook waarom in bovenstaand bewijs dat S^2 zuiver is, de μ 's tegen elkaar wegvallen.

Als schatter voor de standaarddeviatie σ gebruikt men meestal $S = \sqrt{S^2}$. Deze is echter niet zuiver, want aangezien $\text{var}(S) = ES^2 - (ES)^2 > 0$, is $(ES)^2 < ES^2 = \sigma^2$ ofwel $ES < \sigma$. We noemen S^2 de steekproefvariantie, en S de steekproefstandaarddeviatie.

2.3. De empirische verdelingsfunctie

Zoals we al aangaven in de inleiding van dit hoofdstuk, is bij een steekproef X_1, \dots, X_n uit F vaak niets bekend over F . We zullen nu een schatter van F introduceren, n.l. de *empirische verdelingsfunctie* \mathbf{F}_n . Voordat we deze definiëren merken we het volgende op. Stel men is geïnteresseerd in het schatten van één of ander kenmerk van F , d.w.z. in het schatten van een parameter θ . Men kan deze parameter altijd schrijven als functie van F , d.w.z.

$$\theta = T(F).$$

Daarom is een algemeen recept voor de constructie van een schatter van θ : vervang in bovenstaande formule de onbekende F door de schatter \mathbf{F}_n :

$$\hat{\theta} = T(\mathbf{F}_n).$$

Het steekproefgemiddelde is een voorbeeld van een dergelijke schatter, en evenzo voor de steekproefvariantie. We komen hier nog op terug.

Laat X_1, \dots, X_n o.o. s.g.ⁿ zijn met dezelfde verdelingsfunctie

$$F(x) = P(X_1 \leq x) = \dots = P(X_n \leq x), \quad x \in \mathbf{R}.$$

Neem een vaste x en noem

$$Y_i = \begin{cases} 1, & \text{als } X_i \leq x, \\ 0, & \text{anders.} \end{cases}$$

Dan $P(Y_i = 1) = 1 - P(Y_i = 0) = F(x)$, zodat Y_i alternatief verdeeld is met parameter $F(x)$. We hebben al gezien dat \bar{Y} een zuivere schatter is van $F(x)$ met $\text{var}(\bar{Y}) = F(x)(1 - F(x))/n$ (zie voorbeeld (b) in de vorige paragraaf). Merk nu op dat

$$\bar{Y} = \frac{1}{n} \{\text{aantal der } X_i \leq x\}.$$

We definiëren nu voor iedere x

$$\mathbf{F}_n(x) = \frac{1}{n} \{\text{aantal der } X_i \leq x\},$$

en we noemen \mathbf{F}_n de *empirische verdelingsfunctie*. Deze functie springt bij de punten X_i en is verder constant. Als alle waarnemingen verschillend zijn is de spronghoogte steeds $1/n$.

De empirische verdelingsfunctie hoort bij een discrete verdeling die kans $1/n$ toekent aan alle waarnemingen X_1, \dots, X_n . Omdat X_1, \dots, X_n stochastische grootheden zijn, is $\mathbf{F}_n(x)$ ook stochastisch. We geven dit aan door vetgedrukte letters te gebruiken. Een realisatie van \mathbf{F}_n noteren we met F_n .

Het steekproefgemiddelde \bar{X} is precies de verwachting van een stochastische grootheid met verdeling \mathbf{F}_n . Immers, stel Z is een discrete s.g. met waarden in $\{z_1, z_2, \dots\}$. Dan is

$$EZ = \sum_{i=1}^{\infty} z_i P(Z = z_i).$$

We schrijven dit vaak als

$$EZ = \int z F_Z(dz),$$

waarbij F_Z de verdelingsfunctie van Z is. Het voordeel van deze schrijfwijze is dat deze ook voor de verwachting van een continue s.g. is te gebruiken. Samenvattend: het steekproefgemiddelde

$$\bar{X} = \int x \mathbf{F}_n(dx)$$

is een schatter van het populatiegemiddelde

$$\mu = \int x F(dx).$$

Een s.g. met verdeling \mathbf{F}_n heeft variantie

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

\tilde{S}^2 is dan ook een vaak gebruikte schatter voor de variantie σ^2 van de X_i . Of, in de nieuwe schrijfwijze:

$$\tilde{S}^2 = \int (x - \bar{X})^2 \mathbf{F}_n(dx)$$

is een schatter van de populatievariantie

$$\sigma^2 = \int (x - \mu)^2 F(dx).$$

We gebruiken echter meestal een andere schatter, n.l.

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

omdat deze laatste zuiver is.

EEN SCHATTER VAN EEN DICHTHEID.

Stel dat X_1, \dots, X_n o.o. en continu verdeeld met dichtheid $f(x)$. Dan geldt

$$f(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h},$$

waarbij F weer de verdelingsfunctie is. Voor het schatten van $f(x)$ heeft het weinig zin om hier F door de empirische verdelingsfunctie \mathbf{F}_n te vervangen, want \mathbf{F}_n is niet differentieerbaar. Wat men wel kan doen is een vaste h kiezen en $f(x)$ schatten met

$$\hat{\mathbf{f}}(x) = \frac{\mathbf{F}_n(x+h) - \mathbf{F}_n(x)}{h}.$$

Dan geldt

$$E\hat{\mathbf{f}}(x) = \frac{F(x+h) - F(x)}{h} \approx f(x)$$

voor h klein. Dus $\hat{\mathbf{f}}(x)$ is niet zuiver, maar de bias is klein als h klein is. Merk op dat

$$\mathbf{F}_n(x+h) - \mathbf{F}_n(x) = \frac{1}{n} \{\text{aantal der } X_i \text{ met } x < X_i \leq x+h\}.$$

Dit is het steekproefgemiddelde van een steekproef uit een alternatieve verdeling met parameter $F(x+h) - F(x)$. Hieruit volgt dat

$$\text{var}(\hat{\mathbf{f}}(x)) = \frac{(F(x+h) - F(x))(1 - (F(x+h) - F(x)))}{h^2 n} \approx \frac{f(x)}{hn}.$$

De variantie is klein als h groot is. Om een redelijke MSE te krijgen moet men h daarom niet al te groot kiezen om de bias in de hand te houden, en niet al te klein om de variantie in de hand te houden. Men noemt h de *bandbreedte*.

Vb. Stel $f(x) = 2x$, $0 \leq x \leq 1$. Dan (ga na):

$$\text{bias}_f(\hat{\mathbf{f}}(x)) = h,$$

en

$$\text{var}_f(\hat{\mathbf{f}}(x)) = \frac{2x}{hn} + \frac{1 - 4x^2}{n}.$$

De MSE is:

$$\text{MSE}_f(\hat{\mathbf{f}}(x)) = h^2 + \frac{2x}{hn} + \frac{1 - 4x^2}{n}.$$

Deze is het kleinst voor $h = n^{-1/3}x^{1/3}$ (differentieer naar h , stel de afgeleide gelijk aan nul en los op voor h). Omdat we f niet kennen weten we de MSE_f echter niet, dus in praktijk weten we ook niet hoe we h optimaal kunnen kiezen.

Voor de keuze van de bandbreedte h is een uitgebreide theorie opgebouwd, waar we in dit college niet op ingaan. We merken alleen op dat i.h.a. wel iets over de optimale orde van grootte van h gezegd kan worden. In veel situaties moet h van dezelfde orde van grootte als $n^{-1/3}$ zijn.

Een *histogram* is een schatter van $f(x)$ voor alle waarden van x . Het waardebereik van de waarnemingen wordt verdeeld in intervalletjes van lengte h en met eindpunten a_0, a_1, \dots, a_T (dus $a_i = a_{i-1} + h$). Voor $x \in (a_{i-1}, a_i]$ schat men $f(x)$ met

$$\hat{\mathbf{f}}(x) = \frac{\mathbf{F}_n(x_i) - \mathbf{F}_n(x_{i-1})}{h}.$$

Vb. We berekenen het steekproefgemiddelde, de steekproefvariantie en een histogram, voor de volgende waargenomen waarden van X_1, \dots, X_n (met $n = 9$):

$$x_1 = 0.21, x_2 = 0.07, x_3 = 0.43, x_4 = 0.57, x_5 = 0.28, x_6 = 0.92,$$

$$x_7 = 0.73, x_8 = 0.72, x_9 = 0.61.$$

Zet deze eerst op volgorde:

$$x_{(1)} = 0.07, x_{(2)} = 0.21, x_{(3)} = 0.28, x_{(4)} = 0.43, x_{(5)} = 0.57, x_{(6)} = 0.61,$$

$$x_{(7)} = 0.72, x_{(8)} = 0.73, x_{(9)} = 0.92.$$

Er geldt $\bar{x} = 0.5045$ en de waarnemingen in afwijking van het gemiddelde zijn:

$$-0.4345, -0.2945, -0.2245, -0.0745, 0.0655, 0.1055, 0.2155, 0.2255, 0.4155.$$

Het kwadraat $(x_i - \bar{x})^2$ van deze getallen is

$$0.18879, 0.08673, 0.0504, 0.00555, 0.00429, 0.01113, 0.04644, 0.05085, 0.17264.$$

Dus $\sum (x_i - \bar{x})^2 = 0.61682$ en $\tilde{s}^2 = (0.61682/9) = 0.068536$, $s^2 = (0.61682/8) = 0.0771025$, $\tilde{s} = 0.2617$ en $s = 0.2776$.

Neem nu de bandbreedte $h = 0.25$ en $a_0 = 0$. Dan krijg je

$$\hat{f}(x) = \begin{cases} 8/9, & 0 < x \leq 0.25, \\ 8/9, & 0.25 < x \leq 0.50, \\ 16/9, & 0.50 < x \leq 0.75, \\ 4/9, & 0.75 < x \leq 1.00. \end{cases}$$

2.4. Meest aannemelijke schatters

In de vorige paragraaf noemden we al een methode om schatters te construeren, n.l. door de parameter θ als functie van F te schrijven en vervolgens F te vervangen door de empirische verdelingsfunctie. In deze paragraaf behandelen we een andere methode (of eigenlijk: een speciaal geval van de genoemde methode), die vooral werkt als de verdeling bekend is op enkele parameters na. Dit is de situatie waarbij over F veronderstelt wordt dat deze b.v. bij de Poisson verdeling, de normale verdeling, de exponentiële verdeling, enz. hoort. Het idee in deze paragraaf is die waarde als schatter van θ te kiezen, waarvoor de gevonden waarnemingen het meest aannemelijk zijn.

Vb. Laat

$$X_i = \begin{cases} 1, & \text{als het computerprogramma bij gegevensinvoer } i \text{ goed werkt,} \\ 0, & \text{anders.} \end{cases}$$

Stel dat $p = P(X_i = 1)$ de onbekende succeskans is, en dat p voor alle n soorten gegevensinvoer hetzelfde is. We hebben de realisatie

$$(x_1, \dots, x_{10}) = (1, 1, 1, 0, 1, 1, 1, 1, 1, 0)$$

waargenomen. Dus 8 van de 10 keer heeft het programma succesvol gedraaid. Op grond van deze waarneming is het aannemelijk dat p niet al te klein is, want anders zouden we i.h.a. wel meer mislukkingen hebben gevonden. De kans op $(1, 1, 1, 0, 1, 1, 1, 1, 1, 0)$ is

$$L(p) := p^8(1-p)^2$$

(waarbij we onderlinge onafhankelijkheid van de X_i veronderstellen). We noemen $L(p)$ de *aannemelijkheid* van de waarneming $(1, 1, 1, 0, 1, 1, 1, 1, 1, 0)$. Voor welke waarde van p is de aannemelijkheid nu het grootst? We zoeken dan het maximum van $L(p)$. Noem die waarde \hat{p} die $L(p)$ maximaliseert de *meest aannemelijke schatting*.

Maximaliseren van $L(p)$ kan d.m.v. de afgeleide nemen en die gelijk aan nul te stellen:

$$\frac{d}{dp}L(p) = \frac{d}{dp}(p^8 - 2p^9 + p^{10}) = 8p^7 - 18p^8 + 10p^9.$$

$$\frac{d}{dp}L(p) = 0 \Leftrightarrow 8p^7 - 18p^8 + 10p^9 = 0$$

$$\Leftrightarrow p = 0 \vee 8 - 18p + 10p^2 = 0$$

$$\Leftrightarrow p = 0 \vee p = 0.8 \vee p = 1,$$

waarbij we in de laatste stap gebruik maakten van de abc-formule. Nu zijn $p = 0 \vee p = 1$ minima van $L(p)$ en $p = 0.8$ is het maximum. De meest aannemelijke schatting in dit geval is daarom

$$\hat{p} = 0.8.$$

Vb. (Algemener) Stel X_1, \dots, X_n zij o.o. s.g.ⁿ met $P(X_i = 1) = 1 - P(X_i = 0) = p$ onbekend. Als (x_1, \dots, x_n) een realisatie is van (X_1, \dots, X_n) , dan is

$$L(p) = P(X_1 = x_1, \dots, X_n = x_n) = p^{(\sum_{i=1}^n x_i)}(1-p)^{(n - \sum_{i=1}^n x_i)}$$

de aannemelijkheidsfunctie.

Het maximum van $L(p)$ kan gevonden worden door de afgeleide gelijk aan nul te stellen en van de oplossingen na te gaan welke het maximum is. Maar het is handiger om $\log L(p)$ te maximaliseren. Dit is hetzelfde als $L(p)$ maximaliseren omdat het nemen van de logaritme een strict stijgende transformatie is. Met “log” bedoelen we de natuurlijke

logaritme (men mag ook de logaritme met een ander grondgetal kiezen, maar de natuurlijke logaritme blijkt vaak het gemakkelijkst te zijn). We hebben

$$\log L(p) = \sum_{i=1}^n x_i \log(p) + (n - \sum_{i=1}^n x_i) \log(1-p).$$

Zoek het maximum van $\log L(p)$:

$$\begin{aligned} \frac{d}{dp} \log L(p)|_{p=\hat{p}} &= \left(\frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} \right) |_{p=\hat{p}} = 0 \\ \Rightarrow \hat{p} &= \frac{\sum_{i=1}^n x_i}{n} = \bar{x}. \end{aligned}$$

De meest aannemelijke schatting is dus $\hat{p} = \bar{x}$.

Merk op dat \hat{p} van de uitkomsten x_1, \dots, x_n afhangt. Men kan dit aangeven met $\hat{p} = \hat{p}(x_1, \dots, x_n)$. We schrijven verder $\hat{\mathbf{p}} = \hat{p}(X_1, \dots, X_n)$ en noemen $\hat{\mathbf{p}}$ de meest aannemelijke *schatting*. De meest aannemelijke *schatting* is m.a.w. een realisatie van de meest aannemelijke *schatting*.

Er geldt nu dat $\hat{\mathbf{p}} = (\sum_{i=1}^n X_i)/n = \bar{X}$. We hebben al gezien dat dit een zuivere schatter van p is. Zonder bewijs merken we ook nog op dat \bar{X} onder alle zuivere schatters van p , minimale variantie heeft.

ALGEMENE DEFINITIE (discrete geval). Stel X_1, \dots, X_n hebben een simultane discrete verdeling, die van een onbekende parameter θ afhangt. Dan heet

$$L(\theta) = P_\theta(X_1 = x_1, \dots, X_n = x_n),$$

met x_1, \dots, x_n de waargenomen waarden, de *aannemelijkheidsfunctie* (Engels: *likelihood function*). De *meest aannemelijke schatting* $\hat{\theta}$ is gedefinieerd door:

$$L(\hat{\theta}) = \max_{\theta} L(\theta),$$

waarbij gemaximaliseerd wordt over alle mogelijke waarden van θ . Schrijven we $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$, dan heet $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ de meest aannemelijke *schatting*.

OPMERKINGEN.

(i) Het maximum van $L(\theta)$ kan vaak gevonden worden door de afgeleiden naar θ gelijk aan nul te stellen.

(ii) Vaak is het handig om $\log L(\theta)$ te maximaliseren, i.p.v. $L(\theta)$. M.n. als X_1, \dots, X_n o.o. zijn, omdat dan

$$L(\theta) = P_\theta(X_1 = x_1) \dots P_\theta(X_n = x_n),$$

en

$$\log L(\theta) = \sum_{i=1}^n \log P_\theta(X_i = x_i).$$

het is makkelijker om de *som* van een aantal termen te differentiëren, i.p.v. het *produkt* van een aantal termen.

Vb. Laat X_1, \dots, X_n o.o. identiek verdeeld zijn, met

$$P_\theta(X_i = k) = (1 - \theta)\theta^k, \quad k = 0, 1, 2, \dots, \quad 0 < \theta < 1, \quad i = 1, \dots, n.$$

Dan

$$\begin{aligned} L(\theta) &= P_\theta(X_1 = x_1, \dots, X_n = x_n) \\ &= (1 - \theta)\theta^{x_1}(1 - \theta) \dots (1 - \theta)\theta^{x_n} = (1 - \theta)^n \theta^{\sum_{i=1}^n x_i}, \end{aligned}$$

en

$$\begin{aligned} \log L(\theta) &= n \log(1 - \theta) + \sum_{i=1}^n x_i \log(\theta). \\ \frac{d}{d\theta} \log L(\theta)|_{\theta=\hat{\theta}} &= \left(-\frac{n}{1 - \theta} + \frac{\sum_{i=1}^n x_i}{\theta}\right)|_{\theta=\hat{\theta}} = 0 \\ \Rightarrow (1 - \hat{\theta}) \sum_{i=1}^n x_i - n\hat{\theta} &= 0 \Rightarrow \sum_{i=1}^n x_i - \left(\sum_{i=1}^n x_i + n\right)\hat{\theta} = 0 \\ \Rightarrow \hat{\theta} &= \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i + n} = \frac{\bar{x}}{\bar{x} + 1}. \end{aligned}$$

De meest aannemelijke schatter is daarom

$$\hat{\theta} = \frac{\bar{X}}{\bar{X} + 1}.$$

ALGEMENE DEFINITIE (continue geval). Stel X_1, \dots, X_n hebben een continue simultane verdeling die van een onbekende parameter θ afhangt. Zij $f_\theta(x_1, \dots, x_n)$ de dichtheid van X_1, \dots, X_n . Als x_1, \dots, x_n de waargenomen waarden van X_1, \dots, X_n zijn, dan heet

$$L(\theta) = f_\theta(x_1, \dots, x_n)$$

de *aannemelijkheidsfunctie*.

De meest aannemelijke schatting $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ is gedefinieerd door

$$L(\hat{\theta}) = \max_{\theta} L(\theta),$$

waarbij gemaximaliseerd wordt over alle mogelijke waarden van θ . De meest aannemelijke schatter is $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$.

OPMERKINGEN.

(i) Het maximum kan weer vaak gevonden worden door de afgeleiden gelijk aan nul te stellen.

(ii) Het is ook weer vaak handiger om de logaritme te nemen. Als X_1, \dots, X_n o.o. en identiek verdeeld zijn met (marginale) dichtheid $f_\theta(x)$, dan

$$\log L(\theta) = \sum_{i=1}^n \log f_\theta(x_i).$$

Vb. Stel X_1, \dots, X_n zijn homogeen verdeeld op $[0, \theta]$:

$$f_\theta(x) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta.$$

Dan

$$L(\theta) = \left(\frac{1}{\theta}\right)^n, \quad 0 \leq \min_{1 \leq i \leq n} x_i \leq \max_{1 \leq i \leq n} x_i \leq \theta.$$

Dus $\hat{\theta} = \max(x_1, \dots, x_n)$ is de meest aannemelijke schatting, en $\hat{\theta} = \max(X_1, \dots, X_n)$ is de meest aannemelijke schatter. We hebben gezien (zie het voorbeeld op blz. 49) dat deze schatter niet zuiver is.

Vb. Laat X_1, \dots, X_n o.o. $N(\mu, \sigma^2)$ -verdeeld zijn. Dan

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(1/2)\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

Geval 1: σ^2 bekend, μ onbekend.

$$\log L(\mu) = -n \log(\sqrt{2\pi}) - (n/2) \log(\sigma^2) - (1/2) \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^2}.$$

$$\frac{d}{d\mu} \log L(\theta)|_{\mu=\hat{\mu}} = \frac{\sum_{i=1}^n (x_i - \hat{\mu})}{\sigma^2} = 0$$

$$\Rightarrow \hat{\mu} = (1/n) \sum_{i=1}^n x_i = \bar{x}$$

De meest aannemelijke schatter van μ is dus $\hat{\mu} = \bar{X}$.

Geval 2: μ en σ^2 onbekend.

$$\log L(\mu, \sigma^2) = -n \log(\sqrt{2\pi}) - (n/2) \log(\sigma^2) - (1/2) \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^2}.$$

Door de afgeleide naar μ gelijk aan nul te stellen, vind je, net als in geval 1, dat $\hat{\mu} = \bar{x}$.
Verder:

$$\frac{d}{d\sigma^2} \log L(\mu, \sigma^2)|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + (1/2) \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{\hat{\sigma}^4} = 0$$
$$\Rightarrow \hat{\sigma}^2 = (1/n) \sum_{i=1}^n (x_i - \hat{\mu})^2$$

De meest aannemelijke schatters zijn dus $\hat{\mu} = \bar{X}$ en $\hat{\sigma}^2 = (1/n) \sum_{i=1}^n (X_i - \bar{X})^2$. De schatter $\hat{\mu}$ is zuiver, maar $\hat{\sigma}^2$ is niet zuiver. We zijn deze laatste al tegengekomen op blz. 54, waar we deze \tilde{S}^2 genoemd hebben.

2.5. Regressieanalyse

Vb. We beschikken over n datasets met omvang respectievelijk x_1, \dots, x_n . De datasets worden d.m.v. een computerprogramma gecontroleerd op coderingsfouten. Laat y_i de executietijd van het controleprogramma zijn, bij dataset i van omvang x_i , $i = 1, \dots, n$. We willen nu het verband tussen de omvang van een dataset en de executietijd onderzoeken. Het idee is dat we gegeven een dataset van omvang x_i , de executietijd niet precies kunnen voorspellen. D.w.z. y_i is een realisatie van een stochastische grootte Y_i . Gegeven x_i kunnen we iets over de verwachte waarde van Y_i zeggen, als we veronderstellen dat

$$E(Y_i) = g(x_i),$$

waarbij $g(x)$ één of andere functie is. Deze functie is in praktijk meestal onbekend, en moet geschat worden. We nemen wel altijd iets aan over de vorm van g . Als b.v. $g(x) = \alpha + \beta x$, dan spreken we van lineaire regressie. De vorm van g ligt dan vast (namelijk: g is lineair), maar de waarden van α en β zullen in het algemeen onbekend zijn.

Merk nu op dat we kunnen schrijven

$$Y_i = g(x_i) + V_i,$$

met $V_i = Y_i - E(Y_i)$. Dus $E(V_i) = 0$. Men kan V_i interpreteren als meetfout. Omdat V_i verwachting nul heeft, zeggen we dat er geen systematische fout is. Veronderstellen we dat Y_1, \dots, Y_n o.o. zijn, dan zijn ook V_1, \dots, V_n o.o., d.w.z. de meetfout in de éne meting heeft geen invloed op de meetfout in de andere meting.

DEFINITIE LINEAIRE REGRESSIEMODEL:

$$Y_i = \alpha + \beta x_i + V_i, \quad i = 1, \dots, n,$$

met V_1, \dots, V_n o.o. en $E(V_i) = 0$, $i = 1, \dots, n$, x_1, \dots, x_n gegeven getallen, en α en β onbekende parameters.

OPMERKING. Meestal neemt men ook aan dat $\text{var}(V_i)$ constant is voor alle i , zeg $\text{var}(V_i) = \sigma^2$ (met σ^2 i.h.a. onbekend). Dit zegt dat de nauwkeurigheid van de meting niet van i afhangt. Verder geldt dan ook $\text{var}(Y_i) = \sigma^2$ voor alle i .

KLEINSTE-KWADRATENSCHATTERS. Het idee is nu om die lijn $l(x)$ te zoeken die “het best past” bij de waarnemingen (puntenwolk) (x_i, y_i) , $i = 1, \dots, n$. We hanteren daarbij het criterium $\sum_{i=1}^n (y_i - l(x_i))^2$: dit moet voor zekere lijn $l(x)$ zo klein mogelijk zijn.

DEFINITIE. Noem

$$\mathbf{S}(\alpha, \beta) = \sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2.$$

De kleinste-kwadratenschatters (KK-schatters) $\hat{\alpha}$ en $\hat{\beta}$ zijn gedefinieerd door

$$\mathbf{S}(\hat{\alpha}, \hat{\beta}) = \min_{\alpha, \beta} \mathbf{S}(\alpha, \beta).$$

Een realisatie van $(\hat{\alpha}, \hat{\beta})$ noemt men een kleinste-kwadratenschatting.

We schrijven weer $\bar{x} = (1/n) \sum_{i=1}^n x_i$ en $\bar{Y} = (1/n) \sum_{i=1}^n Y_i$ ($\bar{y} = (1/n) \sum_{i=1}^n y_i$).

Lemma.

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x},$$

en

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

BEWIJS.

$$\frac{d}{d\alpha} \mathbf{S}(\alpha, \beta) = -2 \sum_{i=1}^n (Y_i - \alpha - \beta x_i)$$

$$\Rightarrow \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$\Rightarrow \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$

$$\frac{d}{d\beta} \mathbf{S}(\alpha, \beta) = -2 \sum_{i=1}^n (Y_i - \alpha - \beta x_i)x_i = -2 \left(\sum_{i=1}^n Y_i x_i - \alpha \sum_{i=1}^n x_i - \beta \sum_{i=1}^n x_i^2 \right)$$

$$\Rightarrow \sum_{i=1}^n Y_i x_i - \hat{\alpha} \sum_{i=1}^n x_i - \hat{\beta} \sum_{i=1}^n x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^n Y_i x_i (\bar{Y} - \hat{\beta}\bar{x}) \sum_{i=1}^n x_i - \hat{\beta} \sum_{i=1}^n x_i^2 = 0$$

$$\Rightarrow \hat{\beta} \sum_{i=1}^n (x_i^2 - x_i \bar{x}) = \sum_{i=1}^n (Y_i x_i - \bar{Y} x_i)$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^n x_i (Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x}) x_i} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

□

GETALLENVOORBEELD. Stel $n = 3$, $(x_1, x_2, x_3) = (1, 2, 3)$ en $(y_1, y_2, y_3) = (2, 1, 3)$. Dan $\bar{x} = 2$, $\bar{y} = 2$, $\sum_{i=1}^3 (x_i - \bar{x})^2 = 2$ en $\sum_{i=1}^3 (x_i - \bar{x})y_i = 1$. Dus $\hat{\beta} = 1/2$ en $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 1$.

EGENSCHAPPEN VAN KK-SCHATTERS. Deze zullen we in de vorm van twee lemma's geven.

Lemma. $\hat{\alpha}$ en $\hat{\beta}$ zijn zuivere schatters van α resp. β .

BEWIJS.

$$\begin{aligned} E(\hat{\beta}) &= E \left(\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = \frac{E(\sum_{i=1}^n (x_i - \bar{x})Y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})E(Y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

$$= \frac{\alpha \sum_{i=1}^n (x_i - \bar{x}) + \beta \sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta,$$

en

$$\begin{aligned} E(\hat{\alpha}) &= E(\bar{Y} - \hat{\beta}\bar{x}) = E(\bar{Y}) - E(\hat{\beta})\bar{x} \\ &= \alpha + \beta\bar{x} - \beta\bar{x} = \alpha. \end{aligned}$$

□

o

Lemma. Stel dat V_1, \dots, V_n o.o. $N(0, \sigma^2)$ -verdeeld zijn, dan zijn de KK-schatters $\hat{\alpha}$ en $\hat{\beta}$ ook de meest aannemelijke schatters.

BEWIJS. Omdat $E(Y_i) = \alpha + \beta x_i$ en $\text{var}(Y_i) = \sigma^2$ en bovendien Y_i een (nogal simpele) lineaire combinatie van de normaal verdeelde V_i is, is Y_i $N(\alpha + \beta x_i, \sigma^2)$ -verdeeld. De dichtheid van Y_i is dan

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right].$$

Omdat Y_1, \dots, Y_n o.o. zijn is de simultane dichtheid

$$\begin{aligned} f(y_1) \dots f(y_n) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}(y_1 - \alpha - \beta x_1)^2\right] \dots \exp\left[-\frac{1}{2\sigma^2}(y_n - \alpha - \beta x_n)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right] = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2} S(\alpha, \beta)\right]. \end{aligned}$$

De meest aannemelijke schattingen vind je door deze uitdrukking te maximaliseren. Dit is hetzelfde als $S(\alpha, \beta)$ minimaliseren. □

VOORBEELD VAN EEN LINEAIR MODEL. Laat x_i de druk zijn waaraan plastic buis i bloot staat, en Y_i de levensduur van deze buis. Stel men weet dat de verwachte levensduur van een plastic buis op een constante na omgekeerd evenredig is met de druk. In formule:

$$E(Y_i) = \alpha + \frac{\beta}{x_i}, \quad i = 1, \dots, n.$$

Door over te gaan op een nieuwe x -variabele kan je dit in de vorm van een lineair model gieten. Noem n.l. $\tilde{x}_i = (1/x_i)$. Dan $E(Y_i) = \alpha + \beta\tilde{x}_i$, $i = 1, \dots, n$. De parameters α en β kunnen weer met de kleinste-kwadratenmethode geschat worden, waarbij x_i nu steeds vervangen wordt door \tilde{x}_i .

STOCHASTISCHE VERKLARENDE VARIABELEN. De variabele x_i in het regressiemodel noemt men wel de verklarende variabele, en Y_i de te verklaren variabele. Het kan zijn dat x_i niet instelbaar is maar daarentegen een realisatie van een stochastische grootte X_i . Dit maakt in principe niets uit voor het regressiemodel, zolang men maar aanneemt dat X_i onafhankelijk van de meetfout V_i is. De modelaannamen moeten zodanig zijn, dat het regressiemodel geldt, *voorwaardelijk* op de waargenomen waarden x_1, \dots, x_n .

2.6. Toetsingstheorie

Veronderstel dat men wil nagaan of er verschil is in lichaamslengte tussen hoger, resp. lager opgeleide Nederlandse mannen. Een vereenvoudigde en geïdealiseerde procedure om hierover een uitspraak te doen, is als volgt:

-verdeel de opleidingsniveau's in twee categorieën, zeg X en Y .

-neem een steekproef van grootte m resp. n uit X resp. Y , en bepaal de gemiddelde lichaamslengte \bar{x} resp. \bar{y} in de steekproef uit X resp. Y .

-als $|\bar{x} - \bar{y}|$ groot is, zeg $|\bar{x} - \bar{y}| > c$, zet dan in de krant dat er een (oorzakelijk?) verband is tussen lichaamslengte en opleiding. Als $|\bar{x} - \bar{y}| \leq c$, zet dan geen bericht in de krant (terminologie: er is geen *significant* verschil).

De procedure is in principe hiermee welomschreven. Het enige probleem is de keuze van de *kritieke waarde* c . Men wil graag dat als men op grond van de steekproefuitkomsten besluit te publiceren dat er wel significant verschil is, er slechts een kleine kans bestaat, dat dit niet waar is. Laten we deze kans op 0.05 zetten. Dan moet c zo gekozen worden dat de kans dat $|\bar{x} - \bar{y}| > c$ hoogstens 0.05 is als er geen verband is tussen lichaamslengte en opleiding. Verder kiezen we c zo klein mogelijk, omdat we als er wel verband is, we dit graag willen publiceren.

Het bovenstaande probleem noemt men een toetsingsprobleem. In dit speciale geval gaat het om de toets voor twee steekproeven, waarbij men wil nagaan of de steekproeven uit dezelfde verdeling komen. In de nu volgende hoofdstukken zullen we eerst enkele toetsingsproblemen voor één steekproef behandelen.

2.7. Toets voor de alternatieve verdeling

Beschouw een computerpakket waarvan beweerd wordt dat deze kan voorspellen of de aandeelkoersen zullen stijgen, dan wel dalen. Het blijkt echter dat de voorspelling vaak niet uitkomt. Is het computerpakket dan eigenlijk wel wat waard? Als de voorspelling in 50% van de gevallen niet uitkomt, zou men net zo goed een muntje kunnen opgooien: als het munt is dalen de koersen, en anders stijgen de koersen. Noem nu p de kans dat het pakket goed voorspelt. Dan willen we toetsen

$$H_0 : p \leq 1/2,$$

tegen het alternatief

$$H_1 : p > 1/2.$$

Als $p \leq 1/2$ kan je het pakket wel weggooien. We noemen H_0 de *nulhypothese*. Dit is meestal de hypothese die je graag wil verwerpen, of de hypothese dat er niets aan de hand

is, o.i.d.. De *alternatieve hypothese* is meestal de hypothese waar je graag in wil geloven, of de hypothese dat er iets bijzonders gaande is. In ieder geval zijn H_0 en H_1 zo geformuleerd dat als H_0 verworpen wordt, dit vergaande consequenties kan hebben. Daarom wil men de kans op het ten onrechte verwerpen van H_0 klein houden. In dit geval hebben we het zo geformuleerd dat de klant die overweegt het computerpakket aan te schaffen, een kleine kans op problemen wil hebben, mocht hij/zij het pakket inderdaad aanschaffen. Met andere woorden: een conservatieve klant, die het voorspellen liever achterwege laat als het programma niet betrouwbaar genoeg blijkt te zijn.

Stel we laten het pakket 100 keer voorspellen (op verschillende dagen, met voldoende ruimte tussen de dagen). We vinden dat de voorspelling 60 keer uitkomt, en 40 keer niet. Is 60 keer raak voldoende om tot H_1 te concluderen? Om deze vraag te beantwoorden zullen we eerst de algemene strategie formuleren.

STRATEGIE. Noem X het aantal keren dat de voorspelling uitkomt. Kies een kritieke waarde x_r en verwerp H_0 als $X \geq x_r$. De keuze van x_r hangt af van de keuze van z.g. *onbetrouwbaarheidsdrempel* α_0 . Dit is de bovengrens die we nog accepteerbaar achten, voor de kans op ten onrechte verwerpen van H_0 . Er moet dus gelden

$$P(H_0 \text{ verwerpen} | H_0 \text{ waar}) \leq \alpha_0.$$

Ofwel, als $p \leq 1/2$ moet gelden

$$P_p(X \geq x_r) \leq \alpha_0.$$

Ofwel

$$\max_{p \leq 1/2} P_p(X \geq x_r) \leq \alpha_0.$$

Nu is het duidelijk dat hoe groter p , des te groter de kans is dat men veel goede voorspellingen vindt. Dus het maximum wordt aangenomen bij $p = 1/2$, zodat er moet gelden

$$P_{p=1/2}(X \geq x_r) \leq \alpha_0.$$

Voor α_0 kiest men meestal de waarde $\alpha_0 = 0.05$, maar als men erg zeker van de zaak wil zijn, neemt men ook wel $\alpha_0 = 0.01$ of nog kleiner.

ONBETROUWBAARHEID. De kans op H_0 verwerpen is $P_p(X \geq x_r)$. Deze hangt natuurlijk van de (onbekende) p af. Als $p \leq 1/2$ (d.w.z. als H_0 waar is) noemen we $P_p(X \geq x_r)$ de *onbetrouwbaarheid* α . We kiezen de kritieke waarde x_r zo dat $\alpha \leq \alpha_0$. Omdat we p niet kennen, weten we ook niet wat de onbetrouwbaarheid is als H_0 waar. We kunnen echter wel de onbetrouwbaarheid in het slechtste geval, n.l. het geval $p = 1/2$, bepalen. De onbetrouwbaarheid noemt men ook wel de *fout van de eerste soort*. De onbetrouwbaarheidsdrempel α_0 is dus de maximale fout van de eerste soort.

BEPALEN VAN x_r BIJ $n = 100$. X is het aantal voorspellingen dat uitkomt. Als we veronderstellen dat het al of niet uitkomen van een voorspelling voor de individuele voorspellingen o.o. zijn, dan bezit X een binomiale verdeling met parameters n en p . Nu is $n = 100$, wat groot genoeg is om de verdeling van X te benaderen met de normale verdeling met

verwachting np en variantie $np(1-p)$. Voor de zekerheid passen we de continuïteitscorrectie toe bij deze benadering. Het gaat er om x_r zo te kiezen dat

$$P_{p=1/2}(X \geq x_r) \leq \alpha_0,$$

en we gebruiken de benadering

$$\begin{aligned} P_{p=1/2}(X \geq x_r) &\approx 1 - \Phi\left(\frac{x_r - n(1/2) - 1/2}{\sqrt{n(1/2)(1 - (1/2))}}\right) \\ &= 1 - \Phi\left(\frac{x_r - 50.5}{5}\right). \end{aligned}$$

Neem $\alpha_0 = 0.05$ ($1 - \alpha_0 = 0.95$). Nu is $\Phi(1.65) = 0.9505 > 0.95$. Kies daarom

$$\frac{x_r - 50.5}{5} \geq 1.65,$$

ofwel

$$x_r \geq 58.75.$$

We nemen x_r verder zo klein mogelijk, en geheel, dus $x_r = 59$. De nulhypothese dat het pakket niet kan voorspellen, wordt verworpen als van de 100 voorspellingen er minstens 59 uitkomen. Ofwel: bij steekproefgrootte $n = 100$ verwerpen we H_0 als 59% van de voorspellingen uitkomen. Bij de waarneming $X = 60$ wordt H_0 dus verworpen. De kans dat dit ten onrechte is, is hoogstens 5%.

BEPALEN VAN x_r BIJ $n = 400$. Zij nu X het aantal goede voorspellingen in een steekproef van grootte $n = 400$. Dan bezit X een binomiale verdeling met parameters $n = 400$ en p , en we kiezen x_r zó dat

$$P_{p=1/2}(X \geq x_r) \leq \alpha_0.$$

Gebruik weer de normale benadering met continuïteitscorrectie:

$$P_{p=1/2}(X \geq x_r) \approx 1 - \Phi\left(\frac{x_r - 200 - 1/2}{10}\right).$$

Voor $\alpha_0 = 0.05$ nemen we

$$\frac{x_r - 200 - 1/2}{10} \geq 1.65,$$

en verder zo klein mogelijk en geheel. Dit levert $x_r = 217$. We verwerpen H_0 (met onbetrouwbaarheid hoogstens $\alpha_0 = 0.05$) als er minstens $(217/400)\% = 54.25\%$ goede voorspellingen zijn. Vergelijk dit met het geval $n = 100$. Bij grotere steekproeven en gelijke onbetrouwbaarheidsdrempel, verwerpen we dus voor kleinere percentages. D.w.z. bij grotere steekproeven zijn we eerder geneigd een uitspraak te doen, omdat grotere steekproeven meer informatie bevatten.

OVERSCHRIJDINGSKANS BIJ $n = 100$. Als we vinden $X = 60$ dan is de z.g. *overschrijdingskans* bij deze waarde

$$k_r = P_{p=1/2}(X \geq 60) \approx 1 - \Phi(1.9) = 0.0287.$$

Stel we verwerpen H_0 als $k_r \leq \alpha_0 = 0.05$, dan is dit equivalent met het verwerpen van H_0 voor $X \geq x_r$. Het voordeel van het werken met overschrijdingskansen is vaak dat deze gemakkelijker uit te rekenen zijn dan kritieke waarden.

EENZIJDIGE EN TWEEZIJDIGE TOETSEN. Voor de binomiale verdeling met parameters n en p kunnen we 3 soorten toetsen onderscheiden:

-rechtseenzijdig: $H_0 : p \leq p_0, H_1 : p > p_0$,

-linkseenzijdig: $H_0 : p \geq p_0, H_1 : p < p_0$,

-tweezijdig: $H_0 : p = p_0, H_1 : p \neq p_0$.

Hier is p_0 een gegeven getal (b.v. $p_0 = 1/2$). Bij de tweezijdige toets noemen we H_0 een *enkelvoudige* hypothese, omdat p daar maar één waarde kan aannemen. Als p meer waarden kan aannemen spreken we van een *samengestelde* hypothese.

LINKSEENZIJDIGE TOETS. We nemen $p_0 = 1/2, n = 100$ en $\alpha_0 = 0.05$. Waargenomen is de s.g. X met een binomiale verdeling met parameters $n = 100$ en p , en we willen toetsen $H_0 : p \geq 1/2$. H_0 wordt verworpen als $X \leq x_l$ met x_l zó gekozen dat

$$P_{p=1/2}(X \leq x_l) (\approx \Phi\left(\frac{x_l - 50 + 1/2}{5}\right)) \leq 0.05.$$

Nu is $\Phi(-1.65) = 0.0495$, dus we nemen

$$\frac{x_l - 50 + 1/2}{5} \leq -1.65,$$

ofwel

$$x_l \leq 41.25.$$

Verder nemen we x_l zo groot mogelijk en geheel, dus $x_l = 41$. Verwerp H_0 als $X \leq 41$ (vergelijk dit met de rechtseenzijdige toets, waar $H_0 : p \leq 1/2$ verworpen wordt als $X \geq 59$).

TWEEZIJDIGE TOETS. We nemen weer $p_0 = 1/2, n = 100$ en $\alpha_0 = 0.05$, en willen nu toetsen $H_0 : p = 1/2$. H_0 wordt verworpen als $X \geq x_r$ of $X \leq x_l$, voor zekere kritieke waarden x_r en x_l . Merk op dat

$$P_{p=1/2}(X \geq 59) + P_{p=1/2}(X \leq 41) \approx 0.05 + 0.05 = 0.10.$$

Daarom moet je bij een tweezijdige toets met onbetrouwbaarheid $\leq \alpha_0$, zorgen dat de onbetrouwbaarheid van de eenzijdige toetsen $\leq \alpha_0/2$ is. In dit geval $\alpha_0/2 = 0.025$. Er geldt $\Phi(1.96) = 0.975$. Kies daarom

$$\frac{x_r - 50.5}{5} \geq 1.96,$$

en zo klein mogelijk, en

$$\frac{x_l - 49.5}{5} \leq -1.96,$$

en zo groot mogelijk. Zo vind je $x_r = 61$ en $x_l = 39$. Verwerp H_0 als $X \geq 61$ of $X \leq 39$. Is de gevonden waarde $X = 60$ dan wordt H_0 niet verworpen. Als X het aantal goede voorspellingen van het computerpakket is, kan je bij deze uitkomst dus niet concluderen dat het pakket meer doet dan een muntje opgooien.

ONDERSCHIEDEND VERMOGEN. We hebben steeds de kritieke waarde x_r zo klein mogelijk genomen, en x_l zo groot mogelijk. De reden is dat als de nulhypothese niet waar is, we dit graag ook willen ontdekken. De kans op H_0 terecht verwerpen noemt men het *onderscheidend vermogen*.

-rechtseenzijdig. We verwerpen H_0 voor $X \geq x_r$. Definieer

$$\Pi(p) = P_p(X \geq x_r).$$

Dan stijgt $\Pi(p)$ in p (hoe groter p , des te groter de kans op veel goede voorspellingen). We kiezen x_r zó dat $\Pi(p) \leq \alpha_0$ voor alle $p \leq p_0$ (de kans op een fout van de eerste soort mag hoogstens α_0 zijn). Voor $p > p_0$ is $\Pi(p)$ het onderscheidend vermogen, en $1 - \Pi(p)$ de z.g. *fout van de tweede soort*.

-linkseenzijdig. We verwerpen H_0 voor $X \leq x_l$. Noem nu

$$\Pi(p) = P_p(X \leq x_l).$$

Dan daalt $\Pi(p)$ in p (hoe groter p , des te kleiner de kans op weinig goede voorspellingen). De kritieke waarde x_l is zó gekozen dat voor $p \geq p_0$, $\Pi(p) \leq \alpha_0$. Voor $p < p_0$ is $\Pi(p)$ het onderscheidend vermogen en $1 - \Pi(p)$ de kans op een fout van de tweede soort.

-tweezijdig. Verwerp H_0 als $X \geq x_r$ of $X \leq x_l$. Zij

$$\Pi(p) = P_p(X \geq x_r) + P_p(X \leq x_l).$$

We kiezen x_r en x_l zó dat $\Pi(p)$ minimaal is bij $p = p_0$, met $\Pi(p_0) \leq \alpha_0$. Voor $p \neq p_0$ is $\Pi(p)$ weer het onderscheidend vermogen en $1 - \Pi(p)$ de kans op een fout van de tweede soort. Als p veel van p_0 afwijkt is het onderscheidend vermogen groot, d.w.z. dan is de kans groot dat $H_0 : p = p_0$ wordt verworpen.

OPMERKING. We zorgen er altijd voor dat de maximale onbetrouwbaarheid hoogstens α_0 , maar verder zo groot mogelijk is, zodat het onderscheidend vermogen zo groot mogelijk

is. Doordat X discreet verdeeld is, kunnen we de maximale onbetrouwbaarheid niet altijd precies gelijk aan α_0 krijgen.

2.8. Toetsen voor de normale verdeling

MODEL: X_1, \dots, X_n o.o. $N(\mu, \sigma^2)$ -verdeeld.

We beschouwen m.a.w. een aselechte steekproef uit de normale verdeling met verwachting μ en variantie σ^2 . In deze paragraaf zullen we hypothesen over μ bespreken. In praktijk zal σ^2 meestal ook onbekend zijn. Het is wiskundig gezien echter gemakkelijker om eerst het geval σ^2 bekend te bekijken.

Merk op dat $\bar{X} (= (1/n) \sum_{i=1}^n X_i)$ een schatter is van μ . Hypothesen over μ zijn dan ook altijd op \bar{X} gebaseerd. Er geldt:

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

2.8.1. Toetsen van hypothesen over μ ; σ^2 bekend

-rechtsezijdige toets. Laat X_i bijvoorbeeld het dioxinegehalte in melk zijn (op dag i), en μ_0 de maximaal toegestane verwachte hoeveelheid. De toets is

$$H_0 : \mu \leq \mu_0,$$

tegen het alternatief

$$H_1 : \mu > \mu_0.$$

Neem als toetsingsgrootheid:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

en verwerp H_0 als Z groot is. De procedure is net als in de vorige paragraaf: kies onbetrouwbaarheid α_0 en baseer hierop de kritieke waarde voor Z . Dit gaat met de tabel voor de standaard normale verdeling. Het is handig om hiervoor een notatie in te voeren.

DEFINITIE. Het rechter α -punt u_α van de standaard normale verdeling is gedefinieerd door

$$1 - \Phi(u_\alpha) = \alpha.$$

Vb. $u_{0.05} \approx 1.65$, $u_{0.025} = 1.96$.

TOETS VOOR $H_0 : \mu \leq \mu_0$. Verwerp H_0 als $Z > u_{\alpha_0}$. Deze toets heeft onbetrouwbaarheid hoogstens α_0 . Dit kan als volgt worden ingezien. Laat $\Pi(\mu)$ de kans op H_0 verwerpen zijn, als de werkelijke waarde μ is (dus als $X_i \sim N(\mu, \sigma^2)$). In formule

$$\Pi(\mu) = P_\mu(Z > u_{\alpha_0}).$$

Nu is $\Pi(\mu)$ stijgend in μ , want als μ groot is zal de kans dat \bar{X} (en dus Z) groot is ook groot zijn. Dus voor $\mu \leq \mu_0$ is $\Pi(\mu) \leq \Pi(\mu_0)$. En

$$\Pi(\mu_0) = P_{\mu_0}(Z > u_{\alpha_0}) = \alpha_0,$$

want als $\mu = \mu_0$ is Z precies standaard normaal verdeeld. Samenvattend:

$$\max_{\mu \leq \mu_0} \Pi(\mu) = \Pi(\mu_0) = \alpha_0.$$

Ofwel, de kans op H_0 verwerpen als H_0 waar is, is hoogstens α_0 .

GETALLENVOORBEELD. Stel $n = 100$, $\sigma^2 = 0.25$, $\mu_0 = 0.1$ en $\alpha_0 = 0.05$. Dan $u_{\alpha_0} = 1.65$ (ongeveer). Laat $\bar{x} = 0.19$ de waargenomen waarde van \bar{X} zijn. De waargenomen waarde van Z is dan

$$z = \frac{0.19 - 0.1}{(0.5)/10} = 1.8.$$

Deze waarde is groter dan 1.65, zodat $H_0 : \mu \leq 0.1$ verworpen wordt.

ONDSCHIEDEND VERMOGEN. Voor $\mu > \mu_0$ heet $\Pi(\mu)$ weer het onderscheidend vermogen (de kans op H_0 verwerpen als H_0 niet waar is).

-linksezijdige toets.

$$H_0 : \mu \geq \mu_0,$$

$$H_1 : \mu < \mu_0.$$

Verwerp H_0 als

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -u_{\alpha_0},$$

ofwel als

$$\bar{X} < \mu_0 - \frac{\sigma}{\sqrt{n}} u_{\alpha_0}.$$

Noemen we nu

$$\Pi(\mu) = P_{\mu}(Z < -u_{\alpha_0}),$$

dan is $\Pi(\mu)$ dalend in μ , $\Pi(\mu) \leq \alpha_0$ voor $\mu \geq \mu_0$, en voor $\mu < \mu_0$ is $\Pi(\mu)$ het onderscheidend vermogen.

-tweezijdige toets.

$$H_0 : \mu = \mu_0,$$

$$H_1 : \mu \neq \mu_0.$$

Verwerp H_0 als

$$|Z| = \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > u_{\alpha_0/2},$$

ofwel als

$$\bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}} u_{\alpha_0/2} \text{ of } \bar{X} < \mu_0 - \frac{\sigma}{\sqrt{n}} u_{\alpha_0/2}.$$

In dit geval is de kans op verwerpen:

$$\Pi(\mu) = P_\mu(|Z| > u_{\alpha_0/2}).$$

We hebben hier in feite te maken met een combinatie van links- en rechtseenzijdige toets, ieder met onbetrouwbaarheid hoogstens $\alpha_0/2$. De tweezijdige toets heeft onbetrouwbaarheid α_0 , want als $\mu = \mu_0$ is Z precies standaard normaal verdeeld, zodat $\Pi(\mu_0) = \alpha_0$.

GETALLENVOORBEELD. Stel $n = 100$, $\sigma^2 = 0.25$, $\mu_0 = 0.1$ en $\alpha_0 = 0.05$. Dan $u_{\alpha_0/2} = 1.96$. We zagen al dat als $\bar{x} = 0.19$, dan $z = 1.8$. Dus dan $|z| = 1.8 < 1.96$, zodat $H_0 : \mu = 0.1$ niet verworpen kan worden.

OVERSCHRIJDINGSKANSEN.

-*rechtseenzijdig*. De rechtseenzijdige overschrijdingskans is

$$k_r(z) = 1 - \Phi(z),$$

met z de waargenomen waarde van Z . Verwerp H_0 als $k_r(z) \leq \alpha_0$.

Vb. ($n = 100$, $\sigma^2 = 0.25$, $\mu_0 = 0.10$, $\alpha_0 = 0.05$) Bij de waargenomen waarde $\bar{x} = 0.19$ is $z = 1.8$. We vinden $k_r(1.8) = 1 - 0.9641 = 0.0359 < \alpha_0$, dus $H_0 : \mu \leq 0.10$ wordt verworpen.

-*linkseenzijdig*. De linkseenzijdige overschrijdingskans is

$$k_l(z) = \Phi(z),$$

met z de waargenomen waarde van Z . Verwerp H_0 als $k_l(z) \leq \alpha_0$.

Vb. $z = 1.8$, dan $k_l(1.8) = 0.9641 > 0.05$, dus H_0 wordt niet verworpen op 95% niveau.

-*tweezijdig*. De tweezijdige overschrijdingskans is

$$k_{2z}(z) = 2 \min(k_l(z), k_r(z)),$$

met z de waargenomen waarde van Z . Verwerp H_0 als $k_{2z}(z) \leq \alpha_0$.

Vb. $z = 1.8$, dan $k_{2z}(1.8) = 2 \min(0.0359, 0.9641) = 0.0718 > 0.05$, dus H_0 wordt niet verworpen op 95% niveau.

BETROUWBAARHEIDSINTERVAL VOOR μ .

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

dus

$$P_\mu\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq u_{\alpha_0/2}\right) = 1 - \alpha_0.$$

We kunnen bovenstaande als volgt herschrijven:

$$\left| \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right| \leq u_{\alpha_0/2} \Leftrightarrow \bar{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_0/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_0/2},$$

dus

$$P_\mu \left(\bar{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_0/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_0/2} \right) = 1 - \alpha_0.$$

Men noemt

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_0/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_0/2} \right]$$

een $(1 - \alpha_0)$ -betrouwbaarheidsinterval voor μ . Voor $\alpha_0 = 0.05$ is dit dus een 95%-betrouwbaarheidsinterval. Dan $u_{\alpha_0/2} = u_{0.025} = 1.96 \approx 2$. Daarom gebruikt men vaak $\bar{X} \pm$ twee maal de standaarddeviatie van \bar{X} als betrouwbaarheidsinterval. Men zegt ook wel dat twee maal de standaarddeviatie van \bar{X} de *marge* is van de schatter van μ .

De lengte van het betrouwbaarheidsinterval is $(2\sigma/\sqrt{n})u_{\alpha_0/2}$. Men ziet dat als n groot is, dan is het interval klein. En als σ^2 groot is, of α_0 klein, dan is de lengte van het interval groot.

De tweezijdige toets voor $H_0 : \mu \leq \mu_0$ is equivalent met: verwerp H_0 als μ_0 niet in het $(1 - \alpha_0)$ -betrouwbaarheidsinterval ligt.

Vb. ($n = 100$, $\sigma^2 = 0.25$, $\alpha_0 = 0.05$.) Bij een waargenomen waarde $\bar{x} = 0.19$ is $[0.092, 0.288]$ een 95%-betrouwbaarheidsinterval voor μ . De waarde $\mu = 0.1$ ligt in dit betrouwbaarheidsinterval. Daarom wordt de hypothese $H_0 : \mu = 0.1$ niet verworpen op het 95% niveau.

-linkseenzijdig betrouwbaarheidsinterval:

$$\mu \geq \bar{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_0}$$

(hoort bij rechtseenzijdige toets).

-rechtseenzijdig betrouwbaarheidsinterval:

$$\mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_0}$$

(hoort bij linkseenzijdige toets).

2.8.2. Toetsen van hypothesen over μ ; σ^2 onbekend

In de vorige subparagraaf gebruikten we de toetsingsgrootte

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

en we verwierpen de nulhypothese voor extreme waarden van Z . Als nu σ^2 onbekend is, ligt het voor de hand in bovenstaande toetsingsgrootte σ^2 te vervangen door een schatter. De schatter die we hiervoor zullen gebruiken is

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

De resulterende toetsingsgrootheid noemen we

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

Nu is Z normaal verdeeld, en deze eigenschap gebruikten we om de kritieke waarde (b.v. u_{α_0}) te bepalen. De toetsingsgrootheid T is echter niet normaal verdeeld, doordat S een stochastische grootheid is. We moeten daarom de kritieke waarden voor T opnieuw bepalen, en daartoe dienen we de verdeling van T als $\mu = \mu_0$ te weten. De verdeling van T is bepaald door W.S. GOSSET (1876-1939) (schuilnaam: Student). Hoe de de analytische vorm er uit ziet doet er voor ons niet zo toe, het gaat er om dat de verdeling bekend is, en getabelleerd. Als $\mu = \mu_0$, noemt men de verdeling van T een *student*-verdeling (*t*-verdeling) met $(n - 1)$ vrijheidsgraden. Samenvattend:

RESULTAAT. Als X_1, \dots, X_n o.o. zijn, en $N(\mu_0, \sigma^2)$ -verdeeld, dan bezit

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

een *t*-verdeling met $(n - 1)$ vrijheidsgraden.

OPMERKINGEN.

-De *t*-verdeling is symmetrisch rond 0, d.w.z. als T een *t*-verdeling bezit, dan $P(T > a) = P(T < -a)$ voor alle a . Als $\mu \neq \mu_0$, dan bezit T een z.g. *niet-centrale t*-verdeling.

-De *t*-verdeling is getabelleerd. Noem $t_{n-1, \alpha}$ het rechter α -punt van de *t* verdeling met $(n - 1)$ vrijheidsgraden. In de tabel vinden we dan b.v. $t_{9, 0.05} = 1.83$ en $t_{99, 0.05} = 1.66$. Als T een *t*-verdeling bezit met $(n - 1)$ vrijheidsgraden, dan geldt dus $P(T > t_{n-1, \alpha}) = P(T < -t_{n-1, \alpha}) = \alpha$.

-Als n groot is, dan lijkt de *t*-verdeling met $(n - 1)$ vrijheidsgraden erg op de standaard normale verdeling. Dus voor n groot, $t_{n-1, \alpha} \approx u_\alpha$.

RECEPT. Het toetsen van hypothesen over μ is nu analoog aan de vorige subparagraaf. Het verschil is alleen dat overal σ wordt vervangen door S , en u_α door $t_{n-1, \alpha}$.

TOETSEN BIJ ONBETROUWBAARHEIDSDREMPEL α_0 :

-rechtseenzijdig: $H_0 : \mu \leq \mu_0$. Verwerp H_0 als $T > t_{n-1, \alpha_0}$, d.w.z. als

$$\bar{X} > \mu_0 + \frac{S}{\sqrt{n}} t_{n-1, \alpha_0}.$$

-linkseenzijdig: $H_0 : \mu \geq \mu_0$. Verwerp H_0 als

$$\bar{X} < \mu_0 - \frac{S}{\sqrt{n}} t_{n-1, \alpha_0}.$$

-tweezijdig: $H_0 : \mu = \mu_0$. Verwerp H_0 als

$$|\bar{X} - \mu_0| > \frac{S}{\sqrt{n}} t_{n-1, \alpha_0/2}.$$

Een tweezijdig betrouwbaarheidsinterval voor μ is nu

$$[\bar{X} - \frac{S}{\sqrt{n}}t_{n-1,\alpha_0/2}, \bar{X} + \frac{S}{\sqrt{n}}t_{n-1,\alpha_0/2}].$$

Vb. $n = 100$, $\alpha_0 = 0.05$, $t_{99,0.025} = 1.98$. Stel de waargenomen waarden zijn $\bar{x} = 0.19$ en $s^2 = 0.25$. Dan wordt het 95% betrouwbaarheidsinterval $[0.091, 0.289]$.

Vb. Stel $n = 6$ en de waargenomen waarden zijn

$$x_1 = 3, x_2 = 6, x_3 = 5, x_4 = 2, x_5 = 1, x_6 = 4.$$

Dan $\bar{x} = 3.5$ en $s^2 = 3.7$. Verder, $t_{5,0.025} = 2.57$, dus een 95%-betrouwbaarheidsinterval is

$$3.5 \pm \sqrt{\frac{3.7}{6}} 2.57 = 3.5 \pm 2.12.$$

2.9. Studenttoets voor paren

Door het verkeersbureau zijn op vast punt langs een tweebaans snelweg tellingen verricht. De tellingen gebeurden gedurende n werkdagen, 's morgens van 8.00 uur tot 8.01 uur. Laat x_i resp. y_i het aantal auto's dat geteld is op dag i zijn, voor de linkerrijbaan, resp. rechterrijbaan. Stel $n = 4$ en dat de volgende aantallen zijn waargenomen:

$$(x_1, y_1) = (9, 5), (x_2, y_2) = (11, 3), (x_3, y_3) = (6, 8), (x_4, y_4) = (10, 8).$$

We zien dat op de rechter rijbaan over het algemeen meer auto's geteld zijn dan op de linker rijbaan. Alleen de derde dag vormt hier een uitzondering op. We gaan nu toetsen of de verwachte aantallen verschillen.

MODELAANNAMEN. Laat $X_1, Y_1, \dots, X_n, Y_n$ o.o. en normaal verdeeld zijn. Veronderstel

$$X_i \sim N(\mu_i + \Delta, \sigma_1^2), Y_i \sim N(\mu_i, \sigma_2^2), i = 1, \dots, n.$$

OMERKINGEN.

-De aanname van normaliteit is dikwijls op z'n minst twijfelachtig. In Paragraaf 2.12 (over verdelingsvrije methoden) komen we hier op terug.

-We veronderstellen dus dat X_i en Y_i o.o. zijn. Deze eis kan men ook vervangen door de veronderstelling dat (X_i, Y_i) normaal verdeeld is met een covariantie die niet van i afhangt.

-De verwachting van X_i en Y_i mag wel van i afhangen, maar de variantie niet.

TOETS OVER Δ . Noem $Z_i = X_i - Y_i$. Dan is $Z_i \sim N(\Delta, \sigma^2)$, waarbij $\sigma^2 = \sigma_1^2 + \sigma_2^2$ (eventueel min twee maal de covariantie tussen X_i en Y_i). We kunnen nu de theorie van de vorige paragraaf toepassen. Voor het geval σ^2 onbekend nemen we de toetsingsgrootheid

$$T = \frac{\bar{Z}}{S/\sqrt{n}},$$

met $S^2 = (1/(n-1)) \sum_{i=1}^n (Z_i - \bar{Z})^2$. Als $\Delta = 0$ heeft T een t -verdeling met $(n-1)$ vrijheidsgraden.

-tweezijdig. Verwerp $H_0 : \Delta = 0$ als

$$|T| > t_{n-1, \alpha_0/2}.$$

-rechtseenzijdig. Verwerp $H_0 : \Delta \leq 0$ als

$$T > t_{n-1, \alpha_0}.$$

-linkseenzijdig. Verwerp $H_0 : \Delta \geq 0$ als

$$T < -t_{n-1, \alpha_0}.$$

VOORBEELD. In bovenstaand voorbeeld vinden we $(z_1, z_2, z_3, z_4) = (4, 8, -2, 2)$. Dus $\bar{z} = 3$. Verder, $s^2 = 52/3 = 17.33$ en $t = 3/(\sqrt{(17.33)/4}) = 1.44$. Nu geldt dat $t_{3, 0.025} = 3.18$ Dus

$H_0 : \Delta = 0$ kan niet verworpen worden op het 95% niveau. We kunnen dus niet concluderen dat er verschil is tussen de rijbanen. Dit was wel te verwachten, want 4 waarnemingen is erg weinig. Een 95%-betrouwbaarheidsinterval voor Δ is

$$\Delta \in 3 \pm \sqrt{\frac{17.3}{4}} 3.18 = 3 \pm 6.82.$$

2.10. Twee-steekproeven toets

Beschouw nog eens de situatie van Paragraaf 2.6. Hier wilden we nagaan of er verband is tussen lichaamslengte en opleidingsniveau. De toets van de vorige paragraaf kunnen we voor dit probleem niet gebruiken, want er is geen natuurlijke paring tussen mannen met een hoge en mannen met een lage opleiding. We beschikken over twee steekproeven X_1, \dots, X_m en Y_1, \dots, Y_n (en niet over een steekproef van paren $(X_1, Y_1), \dots, (X_n, Y_n)$).

MODELAANNAMEN. Laat $X_1, \dots, X_m, Y_1, \dots, Y_n$ o.o. zijn met

$$X_i \sim N(\mu_1, \sigma^2), Y_j \sim N(\mu_2, \sigma^2), i = 1, \dots, m, j = 1, \dots, n.$$

OPMERKINGEN.

-De aanname van normaliteit is vaak weer twijfelachtig.

-We veronderstellen dus dat X_i onafhankelijk is van Y_j .

-We nemen aan dat de verwachting van X_i niet van i afhangt, en de verwachting van Y_j niet van j . Verder stellen we ook nog dat de varianties van alle X_i en Y_j gelijk zijn.

TOETS OVER $\mu_1 - \mu_2$. Merk op dat \bar{X} een schatter is van μ_1 en \bar{Y} is een schatter van μ_2 . Als we een uitspraak willen doen over $\mu_1 - \mu_2$ ligt het daarom voor de hand naar $\bar{X} - \bar{Y}$ te kijken. Nu is

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma^2}{m}\right), \bar{Y} \sim N\left(\mu_2, \frac{\sigma^2}{n}\right).$$

Dus $\bar{X} - \bar{Y}$ is normaal verdeeld met verwachting $\mu_1 - \mu_2$ en variantie $\text{var}(\bar{X} - \bar{Y}) = \text{var}(\bar{X}) + \text{var}(\bar{Y}) = \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right) = \sigma^2(n+m)/(nm)$. Als $\mu_1 - \mu_2 = 0$ is

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma \sqrt{(n+m)/nm}} = \frac{\bar{X} - \bar{Y}}{\sigma} \sqrt{\frac{mn}{m+n}}$$

standaard normaal verdeeld. Als nu σ bekend is verwerpen we daarom b.v. $H_0 : \mu_1 = \mu_2$ als $|Z| > u_{\alpha_0/2}$. Als σ onbekend is moeten we er een schatter voor verzinnen. We nemen hiervoor

$$S_2 = \frac{1}{m+n-2} \left(\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2 \right).$$

Dit is een zuivere schatter van σ^2 (ga na). De toetsingsgrootte wordt

$$T = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{mn}{m+n}}.$$

RESULTAAT. Als $\mu_1 = \mu_2$, dan bezit T een t -verdeling met $(m + n - 2)$ vrijheidsgraden. (Dit zullen we niet bewijzen.)

-tweezijdige toets. Verwerp $H_0 : \mu_1 = \mu_2$ als

$$|T| > t_{m+n-2, \alpha_0/2}.$$

Een 95%-betrouwbaarheidsinterval voor $\mu_1 - \mu_2$ is dan ook

$$[\bar{X} - \bar{Y} - S\sqrt{\frac{m+n}{mn}}t_{m+n-2, \alpha_0/2}, \bar{X} - \bar{Y} + S\sqrt{\frac{m+n}{mn}}t_{m+n-2, \alpha_0/2}].$$

-rechtseenzijdige toets. Verwerp $H_0 : \mu_1 \leq \mu_2$ als

$$T > t_{m+n-2, \alpha_0}.$$

-linkseenzijdige toets. Verwerp $H_0 : \mu_1 \geq \mu_2$ als

$$T < -t_{m+n-2, \alpha_0}.$$

2.11. Toets voor het vergelijken van twee kansen

Laat $X_i = 1$ als het i -de programma gemaakt door programmeur A werkt, en $X_i = 0$ anders, met $i = 1, \dots, m$. Analoog, laat $Y_j = 1$ als het j -de programma gemaakt door programmeur B werkt, en $Y_j = 0$ anders, $j = 1, \dots, n$. Noem $A = \sum_{i=1}^m X_i$ het aantal werkende programma's van programmeur A, en $B = \sum_{j=1}^n Y_j$ het aantal werkende programma's van programmeur B. Als nu A $m = 12$ programma's heeft gemaakt, waarvan er $A = 8$ werken, en B heeft $n = 14$ programma's gemaakt, waarvan er $B = 8$ werken, kan men dan concluderen dat A een betere programmeur is dan B? Om deze vraag statistisch aan te pakken voeren we het volgende model in:

MODEL.

- X_1, \dots, X_m zijn o.o. met succeskans $p_1 = P(X_i = 1)$, $i = 1, \dots, m$.

- Y_1, \dots, Y_n zijn o.o. met succeskans $p_2 = P(Y_j = 1)$, $j = 1, \dots, n$.

- X_1, \dots, X_m en Y_1, \dots, Y_n zijn o.o..

We willen toetsen de hypothese

$$H_0 : p_1 = p_2.$$

TABEL VAN DE GEGEVENS (in dit voorbeeld).

	1	0		
A	$A = 8$	$C = 4$	$m = 12$	
B	$B = 8$	$D = 6$	$n = 14$	
	$R = 16$	$S = 10$	$m + n = 26$	

TOETSINGSGROOTHEID. Merk eerst op dat $A/m (= \bar{X})$ de fractie successen van programmeur A is, en $B/n (= \bar{Y})$ de fractie successen van programmeur B. Het ligt voor de hand een toets te baseren op het verschil $A/n - B/n (= \bar{X} - \bar{Y})$. Nu is A binomiaal verdeeld met parameters m en p_1 , en B is binomiaal verdeeld met parameters n en p_2 . De verdeling van $A/m - B/n$ hangt af van p_1 en p_2 , dus onder H_0 van de gezamenlijke waarde $p = p_1 = p_2$. Doordat we p niet kennen kunnen we er niet voor zorgen dat een toets gebaseerd op $A/m - B/n$ onbetrouwbaarheid $\leq \alpha_0$ heeft. Beschouw nu het totaal aantal successen $R = A + B$. Het totaal aantal successen zegt niets over het verschil tussen p_1 en p_2 , d.w.z. R bevat geen informatie over het al dan niet waar zijn van H_0 . Dit komt er op neer dat we de toets kunnen uitvoeren gegeven de waargenomen waarde r van R . Er geldt verder dat gegeven $R = r$,

$$A/m - B/n = \left(\frac{1}{m} + \frac{1}{n}\right)A - \frac{r}{n}.$$

Daarom kunnen we voor de voorwaardelijke toets net zo goed A als toetsingsgrootheid nemen.

VOORWAARDELIJKE VERDELING VAN A ONDER H_0 . Onder $H_0 : p_1 = p_2 = p$ geldt

$$\begin{aligned} P_{H_0}(A = a | R = r) &= \frac{P_{H_0}(A = a, R = r)}{P_{H_0}(R = r)} = \frac{P_{H_0}(A = a, B = r - a)}{P_{H_0}(R = r)} \\ &= \frac{\binom{m}{a} p^a (1-p)^{m-a} \binom{n}{r-a} p^{r-a} (1-p)^{n-(r-a)}}{\binom{m+n}{r} p^r (1-p)^{(m+n)-r}} \\ &= \frac{\binom{m}{a} \binom{n}{r-a}}{\binom{m+n}{r}}. \end{aligned}$$

Dus, gegeven dat het totaal aantal successen r is, is het aantal successen van programmeur A onder H_0 hypergeometisch verdeeld.

OVERSCHRIJDINGSKANSEN. Noem

$$\begin{aligned} k_r(j) &= P_{H_0}(A \geq j | R = r) = \sum_{k=j}^{\min(m,r)} \frac{\binom{m}{k} \binom{n}{r-k}}{\binom{m+n}{r}}, \\ k_l(j) &= P_{H_0}(A \leq j | R = r) = \sum_{k=\max(0,r-n)}^j \frac{\binom{m}{k} \binom{n}{r-k}}{\binom{m+n}{r}}, \end{aligned}$$

en

$$k_{2z}(j) = 2 \min(k_r(j), k_l(j)).$$

De toets wordt nu: verwerp $H_0 : p_1 = p_2$ als $A \geq a_r$ of $A \leq a_l$, waarbij de kritieke waarden a_r en a_l zó gekozen dienen te worden dat $k_r(a_r) \leq \alpha_0/2$ en $k_l(a_l) \leq \alpha_0/2$. Immers, dan is de kans op H_0 ten onrechte verwerpen, gegeven $R = r$:

$$P_{H_0}(A \geq a_r \text{ of } a \leq a_l | R = r) = k_r(a_r) + k_l(a_l) \leq \alpha_0/2 + \alpha_0/2 = \alpha_0.$$

Verder kiezen we a_r zo klein mogelijk en a_l zo groot mogelijk, om er voor te zorgen dat het onderscheidend vermogen zo groot mogelijk is. Het is eenvoudiger om met overschrijdingskansen te werken, want dan hoeft men a_r en a_l niet uit te rekenen. De toets wordt dan: verwerp H_0 als $k_{2z}(a) \leq \alpha_0$, met a de waargenomen waarde van A . Het is eenvoudig in te zien dat

$$k_{2z}(a) \leq \alpha_0 \Leftrightarrow a \geq a_r \text{ of } a \leq a_l,$$

dus het werken met overschrijdingskansen is hetzelfde als het werken met kritieke waarden.

EENZIJDIGE TOETSEN.

-rechtseenzijdig. Verwerp $H_0 : p_1 \leq p_2$ als $k_r(a) \leq \alpha_0$, met a de waargenomen waarde van A .

-linkseenzijdig. Verwerp $H_0 : p_1 \geq p_2$ als $k_l(a) \leq \alpha_0$, met a de waargenomen waarde van A .

BENADERING VAN DE VERDELING VAN A . Gegeven $R = r$ bezit A onder H_0 een hypergeometrische verdeling. Hieruit kunnen we kritieke waarden (of overschrijdingskansen) berekenen. Nu is het probleem dat de hypergeometrische verdeling afhangt van m , n , en r , en niet getabelleerd is voor alle waarden van deze parameters. Als m en n klein zijn, is het nog te doen om de overschrijdingskansen rechtstreeks uit te rekenen. Voor grotere waarden van m , n en r kan men gelukkig een benadering gebruiken. Laten we eerst A standaardiseren. We hebben al op blz. 30 aangetoond dat

$$E_{H_0}(A | R = r) = m \frac{r}{m+n}$$

(zie ook blz. 31). We gebruikten daar een andere notatie. Om ze uit elkaar te halen is het misschien handig te schrijven: $N^* = m+n$, $R^* = m$ en $n^* = r$, en de symbolen op blz. 30 of 31 door dezelfde symbolen met een sterretje te vervangen. Op blz. 36 lieten we zien dat

$$\text{var}_{H_0}(A | R = r) = r \left(\frac{m}{m+n} \right) \left(1 - \left(\frac{r}{m+n} \right) \right) \frac{m+n-r}{m+n-1} = \frac{mnrs}{(m+n)^2(m+n-1)},$$

met $s = m+n-r$. Noem nu

$$Z = \frac{A - (rm/(m+n))}{\sqrt{mnrs/((m+n)^2(m+n-1))}} = \frac{AD - BC}{\sqrt{mnrs}} \sqrt{m+n-1}.$$

Hier is $C = m - A$ en $D = n - B$ (zie de tabel). Dan heeft, gegeven $R = r$, de s.g. Z onder H_0 verwachting nul en variantie 1. Zo hebben we A gestandaardiseerd. Het blijkt nu dat,

als m , n en r groot zijn, Z (gegeven $R = r$) onder H_0 ongeveer standaard normaal verdeeld is. Daarom kan je dan de kritieke waarden uit de tabel voor de standaard normale verdeling halen. Een benadering van b.v. de tweezijdige toets wordt nu: verwerp $H_0 : p_1 = p_2$ als $|Z| > u_{\alpha_0/2}$.

VOORBEELD. Als $m = 12$, $n = 14$, en $r = 16$, dan is voor de waargenomen waarde $a = 8$, de waargenomen waarde van Z :

$$z = \frac{48 - 32}{\sqrt{(12)(14)(16)(10)}} \sqrt{25}.$$

Dit is kleiner dan $1.96 = u_{0.025}$, dus $H_0 : p_1 = p_2$ wordt niet verworpen als we de onbetrouwbaarheid hoogstens 0.05 willen hebben. (Merk op dat in dit voorbeeld de waarden van m , n en r niet al te groot zijn. Daarom is het eigenlijk beter de toets exact uit te voeren.)

2.12. Verdelingsvrije methoden

In Paragraaf 2.9 en Paragraaf 2.10 veronderstelden we dat de waarnemingen uit een normale verdeling komen. Zoals we daar al opmerkten is dit vaak niet realistisch. We zullen nu enkele toetsingsgrootheden introduceren, waarvan de verdeling (onder H_0) niet afhangt van de verdeling waar de waarnemingen uitkomen. Er hoeft dan ook bijna geen enkele veronderstelling over deze laatste te worden gemaakt.

2.12.1. Verdelingsvrije toetsen voor paren van waarnemingen.

Laat $(X_1, Y_1), \dots, (X_n, Y_n)$ o.o. paren van waarnemingen zijn. Veronderstel bovendien dat X_i en Y_i o.o. zijn, $i = 1, \dots, n$. Als voorbeeld kunt u denken aan de 1-minuut tellingen voor twee rijstroken (zie Paragraaf 2.9). We willen toetsen:

$H_0 : X_i$ en Y_i bezitten dezelfde verdeling, $i = 1, \dots, n$.

Het idee is nu: zoek een toetsingsgrootheid, en verwerp H_0 voor extreme waarden van deze toetsingsgrootheid. Zorg er bovendien voor dat de verdeling van de toetsingsgrootheid onder H_0 bekend is, zodat men kritieke waarden bij gegeven onbetrouwbaarheidsdrempel kan uitrekenen.

De tekentoets

TOETSINGSGROOTHEID VAN DE TEKENTOETS:

Noem K = het aantal paren (X_i, Y_i) met $X_i > Y_i$.

VERDELING VAN K ONDER H_0 .

(i) Bekijk eerst het geval dat X_i en Y_i continu verdeeld zijn. Dan geldt dat $P(X_i = Y_i) = 0$, d.w.z. dan kunnen er geen gelijke waarden voorkomen. Als X_i en Y_i nu bovendien dezelfde verdeling hebben, dan is de kans dat X_i groter dan Y_i is gelijk aan de kans dat Y_i groter is dan X_i . Deze kans moet dan wel gelijk zijn aan $1/2$. Dus

$$P_{H_0}(X_i > Y_i) = 1/2.$$

Nu kunnen we schrijven

$$K = \sum_{i=1}^n V_i,$$

met $V_i = 1$ als $X_i > Y_i$ en $V_i = 0$ als $Y_i > X_i$. Dus

$$P_{H_0}(V_i = 1) = 1 - P_{H_0}(V_i = 0) = 1/2.$$

Bovendien zijn V_1, \dots, V_n o.o.. Daarom bezit onder H_0 de toetsingsgrootheid K een binomiale verdeling met parameters n en $p = 1/2$. Dus

$$P_{H_0}(K = j) = \binom{n}{j} \left(\frac{1}{2}\right)^n, \quad j = 0, 1, \dots, n.$$

OVERSCHRIJDINGSKANSEN. Noem

$$k_r(j) = P_{H_0}(K \geq j) = \sum_{l=j}^n \binom{n}{l} \left(\frac{1}{2}\right)^n,$$

$$k_l(j) = P_{H_0}(K \leq j) = \sum_{l=0}^j \binom{n}{l} \left(\frac{1}{2}\right)^n,$$

en

$$k_{2z}(j) = 2 \min(k_r(j), k_l(j)).$$

De (tweezijdige) toets wordt: verwerp H_0 als $k_{2z}(k) \leq \alpha_0$, met k de waargenomen waarde van K . Als n groot is kan men weer de binomiale verdeling benaderen met de normale verdeling (eventueel met continuïteitscorrectie). De exacte toets of de benadering gaat net zo als in Paragraaf 2.7.

(ii) In bovenstaande veronderstelden we dat X_i en Y_i continu verdeeld zijn, zodat gelijke waarden niet kunnen voorkomen. In praktijk kan je gelijke waarden nooit uitsluiten. We kunnen te maken hebben met afrondingsfouten, of de verdeling is discreet van nature (bijv. als het gaat om aantallen). De paren met $X_i = Y_i$ bevatten geen informatie over het al of niet waar zijn van H_0 . Daarom bekijken we de verdeling van K , gegeven dat precies n' der paren, zeg $(X_1, Y_1), \dots, (X_{n'}, Y_{n'})$, ongelijke paren zijn. Er geldt

$$P_{H_0}(K = j | X_i \neq Y_i, i = 1, \dots, n') = \binom{n'}{j} \left(\frac{1}{2}\right)^{n'}.$$

Met andere woorden, het komt er op neer dat we n vervangen door n' (dit is het aantal paren met $X_i \neq Y_i$) en de paren met $X_i = Y_i$ buiten beschouwing laten. De toets verloopt verder als boven beschreven, met n vervangen door n' . Als n' groot kunnen we de normale benadering gebruiken: Verwerp H_0 als

$$\left| \frac{K - n'(1/2)}{\sqrt{n'(1/2)(1 - (1/2))}} \right| > u_{\alpha_0/2}.$$

Eventueel past men nog de continuïteitscorrectie toe.

VOORBEELD. We bekijken het voorbeeld van Paragraaf 2.9. Hier komen geen gelijke paren voor, dus $n = n' (= 4)$. Verder zien we dat de waargenomen waarde van K is $k = 3$. De overschrijdingskansen voor deze waarde zijn

$$k_r(3) = \binom{4}{3} \left(\frac{1}{2}\right)^4 + \binom{4}{4} \left(\frac{1}{2}\right)^4 = \frac{5}{16},$$

$$k_l(3) = P_{H_0}(K \leq 3) = 1 - P_{H_0}(K \geq 4) = \frac{15}{16},$$

en $k_{2z}(3) = 2 \min(5/16, 15/16) = 10/16 > 0.05$. Dus bij onbetrouwbaarheidsdrempel $\alpha_0 = 0.05$ kunnen we H_0 niet verwerpen. Dit was te verwachten, want we konden al geen uitspraak doen als we normaliteit veronderstelden (zie Paragraaf 2.9), dus laat staan als we deze veronderstelling achterwege laten. Merk op dat $k_r(j)$ minimaal $1/16$ is (voor $j = 4$ wordt dit minimum aangenomen). Omdat $1/16 > 0.025$ kunnen we H_0 nooit verwerpen met onbetrouwbaarheid hoogstens 5%. Het aantal waarnemingen is gewoon te klein om een uitspraak te doen. We moeten minimaal 6 waarnemingen hebben. Als we dan vinden dat *alle* X_i groter dan Y_i zijn, is de overschrijdingskans $(1/2)^6 = 1/64 < 0.025$, dus dan kunnen we H_0 verwerpen.

Symmetrietoets van Wilcoxon

TOETSINGSGROOTHEID VAN DE SYMMETRIETOETS. We bekijken nog steeds de situatie van o.o. paren van o.o. stochastische grootheden (X_i, Y_i) . Voor het gemak veronderstellen we dat X_i en Y_i continu verdeeld zijn. Noem

- $Z_i = X_i - Y_i$, $i = 1, \dots, n$,

- R_i = het rangnummer van Z_i als we $|Z_1|, \dots, |Z_n|$ rangschikken naar opklimmende volgorde,

- $T = \sum_{Z_i > 0} R_i - \sum_{Z_i < 0} R_i$,

- $W = \sum_{Z_i > 0} R_i$.

Als toetsingsgrootheid gebruiken we nu T of W . Omdat er een 1-1-duidelijk verband tussen deze twee bestaat (n.l. $T = 2W - n(n+1)/2$) maakt het niet uit welke je gebruikt.

VOORBEELD.

x	5.5	3.6	6.0	10.6	0.0	10.0
y	1.0	1.0	7.0	5.0	4.0	8.0
z	4.9	2.6	-1.0	5.6	-4.0	2.0
r	5	3	1	6	4	2
teken	+	+	-	+	-	+

$n = 6$, $t = 5 + 3 + 6 + 2 - 5 = 11$, $w = 16$, $t = 2w - 21$.

VERDELING VAN W ONDER H_0 . We kunnen schrijven

$$W = \sum_{i=1}^n iV_i,$$

waarbij V_1, \dots, V_n o.o. zijn met onder H_0 ,

$$P_{H_0}(V_i = 1) = P_{H_0}(V_i = 0) = 1/2.$$

De verdeling van W (en dus die van T) onder H_0 is daarom in principe bekend. Er bestaan tabellen van. Als n groot is kunnen we de benadering met de normale verdeling gebruiken. Daartoe moet men eerst W standaardiseren. We zien dat

$$E_{H_0} W = \sum_{i=1}^n i E_{H_0}(V_i) = \sum_{i=1}^n i(1/2) = n(n+1)/4,$$

en

$$\text{var}_{H_0}(W) = \sum_{i=1}^n i^2 \text{var}_{H_0}(V_i) = \sum_{i=1}^n i^2(1/4) = n(n+1)(2n+1)/24.$$

De toetsingsgrootheid

$$\frac{W - E_{H_0}(W)}{\sqrt{\text{var}_{H_0}(W)}} = \frac{T}{\sqrt{n(n+1)(2n+1)/6}}$$

is voor n groot ongeveer standaard normaal verdeeld. We verwerpen daarom H_0 als

$$\left| \frac{T}{\sqrt{n(n+1)(2n+1)/6}} \right| > u_{\alpha_0/2}.$$

2.12.2. Twee-steekproeventoets van Wilcoxon

We beschikken over twee o.o. steekproeven uit F resp. G :

- X_1, \dots, X_m o.o. met $P(X_i \leq x) = F(x)$, $i = 1, \dots, m$,

- Y_1, \dots, Y_n o.o. met $P(Y_j \leq y) = G(y)$, $j = 1, \dots, n$.

(Denk bijv. aan de situatie van Paragraaf 2.6.)

In Paragraaf 2.10 veronderstelden we dat F en G verdelingsfuncties waren van een normale verdeling. In deze paragraaf zullen we alleen aannemen dat F en G continu zijn. Dit is slechts voor het gemak. Als F en G continu zijn, dan kunnen er geen gelijke waarden tussen de X_i en Y_j voorkomen en dit maakt het makkelijker om de onderstaande toetsingsgrootheid te beschrijven. Wat we n.l. willen toetsen is

$$H_0 : F = G.$$

M.a.w., de nulhypothese is dat X_i en Y_j voor alle i en j dezelfde verdeling bezitten.

TOETSINGSGROOTHEID. Noem

- R_i = het rangnummer van X_i als we de rij $\{X_1, \dots, X_m, Y_1, \dots, Y_n\}$ rangschikken naar opklimmende grootte,

- $U = \sum_{i=1}^m R_i$,

- $W = 2 \times$ het aantal paren (X_i, Y_j) met $X_i > Y_j$.

Als X_i vaak groter dan Y_j is, dan zullen de rangnummers van de X_i ook groot zijn. Dus dan is U groot. Ook W is dan groot. Er geldt dat

$$U = \frac{W}{2} + \frac{m(m+1)}{2},$$

(ga na). Analoog, als X_i meestal kleiner dan Y_j is, dan zijn U en W klein. We verwerpen daarom H_0 voor extreme (d.w.z. grote of kleine) waarden van U . Een toets gebaseerd op U is equivalent met een toets gebaseerd op W omdat er een 1-1-duidig verband tussen U en W is.

VOORBEELD. Stel $m = 5$, $n = 4$, en dat de waargenomen waarden zijn:

$$x_1 = 36, x_2 = 9, x_3 = 7, x_4 = 100, x_5 = 3,$$

$$y_1 = 5, y_2 = 37, y_3 = 11, y_4 = 12.$$

Dan

$$r_1 = 7, r_2 = 4, r_3 = 3, r_4 = 9, r_5 = 1,$$

want

$$x_5 < y_1 < x_3 < x_2 < y_3 < y_4 < x_1 < y_2 < x_4.$$

Dus

$$u = 7 + 4 + 3 + 9 + 1 = 24,$$

en

$$w = 2 \times (4 + 3 + 1 + 1) = 18.$$

Inderdaad zien we dat $u = w/2 + m(m+1)/2 = 9 + 15$.

VERDELING VAN U ONDER H_0 . Als H_0 waar is, dan kunnen we U beschouwen als de totale uitkomst van een steekproef van grootte m , zonder terugleggen, uit de getallen $\{1, 2, \dots, m+n\}$. D.w.z. onder H_0 is R_1, \dots, R_m te zien als een steekproef zonder terugleggen uit $\{1, 2, \dots, N\}$, waarbij $N = m+n$. De verdeling onder de nulhypothese is daarom in principe bekend. Er bestaan tabellen van.

Voor grote waarden van m en n kan men de benadering met de standaard normale verdeling gebruiken. Dan moeten we U (of W) eerst standaardiseren onder H_0 . Er geldt

$$P_{H_0}(R_i = k) = 1/N, \quad i = 1, \dots, m, \quad k = 1, \dots, N$$

(zie Paragraaf 1.6 over urnmodellen: we hebben te maken met de situatie van blz. 12, met $R = 1$, want het aantal elementen in $\{1, \dots, N\}$ dat het kenmerk bezit gelijk aan k te zijn is één). Dus

$$E_{H_0}(R_i) = \sum_{k=1}^N P_{H_0}(R_i = k)k = (1/N) \sum_{k=1}^N k = (N+1)/2,$$

en

$$E_{H_0}(U) = \sum_{i=1}^m E_{H_0}(R_i) = m(N+1)/2.$$

Omdat $W = 2U - m(m+1)$ is

$$E_{H_0}(W) = m(N+1) - m(m+1) = mn.$$

Dit laatste volgt ook uit het feit dat er mn paren (X_i, Y_j) zijn, dus onder H_0 is, in verwachting, van de helft van deze paren de X_i groter Y_j . Zonder bewijs vermelden we dat

$$\text{var}_{H_0}(U) = \frac{mn(m+n+1)}{12},$$

zodat

$$\text{var}_{H_0}(W) = \frac{mn(m+n+1)}{3}.$$

Als we nu gebruik maken van de benadering met de standaard normale verdeling, dan wordt de toets: verwerp H_0 als

$$\left| \frac{U - E_{H_0}(U)}{\sqrt{\text{var}_{H_0}(U)}} \right| \left(= \frac{W - E_{H_0}(W)}{\sqrt{\text{var}_{H_0}(W)}} \right) = \left| \frac{W - mn}{\sqrt{mn(m+n+1)/3}} \right| > u_{\alpha_0/2}.$$

Overzicht van enkele begrippen en resultaten

-voorwaardelijke kans: als $P(B) \neq 0$, $P(A|B) = P(A \cap B)/P(B)$. (blz. 5)

-Gebeurtenissen A en B zijn *onafhankelijk* als $P(A \cap B) = P(A)P(B)$. (blz. 9)

-verdelingsfunctie: $F(x) = P(X \leq x)$. (blz. 14, 19)

-Als X continu, dan $dF(x)/dx = f(x)$, $f(x)$ *dichtheid* X . Er geldt: $\int f(x)dx = 1$. (blz. 19)

-Als (X, Y) continu, dan $f_X(x) = \int f_{X,Y}(x, y)dy$ en $f_Y(y) = \int f_{X,Y}(x, y)dx$, waarbij $f_{X,Y}$ de *simultane* dichtheid voorstelt, en f_X resp. f_Y de *marginale* dichtheid van X resp. Y . (blz. 25)

-Stochastische grootheden X en Y heten *onderling onafhankelijk* als

discreet: $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$ voor alle x_i en y_j

continu: $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ voor alle x en y . (blz. 26)

-*verwachting*:

discreet: $E(X) = \sum P(X = x_k)x_k$. (blz. 29)

continu: $E(X) = \int f(x)x dx$. (blz. 30)

Eigenschap: $E(aX + bY + c) = aE(X) + bE(Y) + c$. (blz. 30)

-*variantie*: $\text{var}(X) = E(X^2) - (EX)^2$. De *standaardafwijking* is $\sigma(X) = \sqrt{\text{var}(X)}$. (blz. 31)

Eigenschap: $\text{var}(aX + b) = a^2\text{var}(X)$. (blz. 32)

-*covariantie*: $\text{cov}(X, Y) = E(XY) - (EX)(EY)$. (blz. 34)

Eigenschap: $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$. (blz. 35)

-*correlatie*: $\rho(X, Y) = \text{cov}(X, Y)/(\sigma(X)\sigma(Y))$. (blz. 36)

-*centrale limietstelling*: X_1, \dots, X_n o.o. en identiek verdeeld met verwachting μ en variantie σ^2 , dan is voor n groot $\sqrt{n}(\bar{X} - \mu)/\sigma$ ongeveer standaard normaal verdeeld. (blz. 41)

Speciaal geval: als X binomiaal verdeeld met parameters n en p , dan is voor n groot X ongeveer $N(np, np(1-p))$ -verdeeld. (blz. 42)

-*mean square error*: $\text{MSE}_\theta(T) = E_\theta(T - \theta)^2$. (blz. 48)

-*zuiverheid*: T zuivere schatter θ als $E_\theta(T) = \theta$ voor alle mogelijke waarden van θ . (blz. 48)

Eigenschap: als T zuiver schatter θ , dan $\text{MSE}_\theta(T) = \text{var}_\theta(T)$. (blz. 48)

-*schatter μ en σ^2* : stel X_1, \dots, X_n zijn o.o. met verwachting μ en variantie σ^2 , dan \bar{X} zuivere schatter μ met variantie $\text{var}(\bar{X}) = \sigma^2/n$, en $S^2 = (1/(n-1)) \sum_{i=1}^n (X_i - \bar{X})^2$ zuivere schatter σ^2 . (blz. 51)

-*empirische verdelingsfunctie*: $\mathbf{F}_n(x) = (1/n) \times \{\text{aantal } X_i \leq x\}$. (blz. 53)

-*histogram*: $\hat{\mathbf{f}}(x) = (\mathbf{F}_n(x_i) - \mathbf{F}_n(x_{i-1})) / (x_i - x_{i-1})$ voor $x \in (x_{i-1}, x_i]$. (blz. 55)

-*log-aannemelijkheidsfunctie* voor het geval dat X_1, \dots, X_n o.o. en identiek verdeeld zijn (met realisaties x_1, \dots, x_n):

discreet: $\log L(\theta) = \sum \log P_\theta(X_i = x_i)$. (blz. 58)

continu: $\log L(\theta) = \sum \log f_\theta(x_i)$. (blz. 60)

-*meest-aannemelijke schatting*: $\hat{\theta}$ maximaliseert $\log L(\theta)$. (blz. 58, 60)

-kleinste-kwadraten schattingen: $\hat{\alpha}$ en $\hat{\beta}$ minimaliseren de kwadratensom $S(\alpha, \beta) = \sum (y_i - \alpha - \beta x_i)^2$, d.w.z. $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$ en $\hat{\beta} = (\sum (x_i - \bar{x}) y_i) / (\sum (x_i - \bar{x})^2)$. (blz. 63)

-onbetrouwbaarheid: $P(H_0 \text{ verwerpen} | H_0 \text{ waar})$. (blz. 66)

-onderscheidend vermogen: $P(H_0 \text{ verwerpen} | H_0 \text{ niet waar})$. (blz. 69, 71)

-overschrijdingskansen, bijv. als X binomiaal verdeeld is met parameters n en p :

$$k_r(x) = P_{p_0}(X \geq x),$$

$$k_l(x) = P_{p_0}(X \leq x),$$

$$k_{2z}(x) = 2 \min(k_r(x), k_l(x)),$$

met x de waargenomen waarde van X .

$$\text{Verwerp } H_0 : p \leq p_0 \text{ als } k_r(x) \leq \alpha_0.$$

$$\text{Verwerp } H_0 : p \geq p_0 \text{ als } k_l(x) \leq \alpha_0.$$

$$\text{Verwerp } H_0 : p = p_0 \text{ als } k_{2z}(x) \leq \alpha_0. \text{ (blz. 82)}$$

-toets voor de normale verdeling:

σ bekend: verwerp $H_0 : \mu = \mu_0$ als $|\bar{X} - \mu_0| > (\sigma/\sqrt{n})u_{\alpha_0/2}$. Een $(1 - \alpha_0)$ -betrouwbaarheidsinterval voor μ is $\bar{X} \pm (\sigma/\sqrt{n})u_{\alpha_0/2}$. (blz. 71, 72)

σ onbekend: verwerp $H_0 : \mu = \mu_0$ als $|\bar{X} - \mu_0| > (S/\sqrt{n})t_{n-1, \alpha_0/2}$. Een $(1 - \alpha_0)$ -betrouwbaarheidsinterval voor μ is $\bar{X} \pm (S/\sqrt{n})t_{n-1, \alpha_0/2}$. Hier is $S^2 = (1/(n-1)) \sum (X_i - \bar{X})^2$. (blz. 74, 75)

-studenttoets voor paren: verwerp $H_0 : \Delta = 0$ als $\sqrt{n}|\bar{X} - \bar{Y}|/S > t_{n-1, \alpha_0/2}$. Hier is $S^2 = (1/(n-1)) \sum (X_i - Y_i - (\bar{X} - \bar{Y}))^2$. (blz. 76)

-twee-steekproeuntoets: verwerp $H_0 : \mu_1 = \mu_2$ als $\sqrt{(m+n)/(mn)}|\bar{X} - \bar{Y}|/S > t_{m+n-2, \alpha_0/2}$. Hier is $S^2 = (1/(m+n-2))(\sum (X_i - \bar{X})^2 + \sum (Y_j - \bar{Y})^2)$. (blz. 78)

-toets voor het vergelijken van twee kansen:

$$k_r(a) = P_{H_0}(A \geq a | R = r),$$

$$k_l(a) = P_{H_0}(A \leq a | R = r),$$

$$k_{2z}(a) = 2 \min(k_r(a), k_l(a)).$$

Verwerp $H_0 : p_1 = p_2$ als $k_{2z}(a) \leq \alpha_0$, met a de waargenomen waarde van A . (blz. 80)

-tekentoets: $K =$ het aantal paren met $X_i > Y_i$. (blz. 81)

-symmetrietoets van Wilcoxon: $W = \sum_{X_i > Y_i} (\text{rangnummer van } |X_i - Y_i|)$. (blz. 83)

-twee-steekproeuntoets van Wilcoxon: $W = 2 \times$ het aantal paren met $X_i > Y_j$. (blz. 84)