TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

WISKUNDE 31

bestemd voor

BDK-III

WISKUNDE 49

bestemd voor

N-IV, E-IV, T-IV, B-IV

WISKUNDE 64

bestemd voor

W-VI

Kansrekenening en Statistiek

Voorjaarssemester 1981

Technische Hogeschool Eindhoven

Onderafdeling der Wiskunde

Wiskunde 31

bestemd voor BDK-III

Wiskunde 49

bestemd voor N-IV, E-IV, T-IV, B-IV

Wiskunde 64

bestemd voor W-VI

Kansrekening en Statistiek

Diktaatnr. 2.265 Prijs f 4,–

TECHNISCHE HOGESCHOOL EINDHOVEN

Onderafdeling der Wiskunde

WISKUNDE 31
bestemd voor BDK-III

WISKUNDE 49
bestemd voor N-IV, E-IV, T-IV, B-IV

WISKUNDE 64,
bestemd voor W-VI

KANSREKENING EN STATISTIEK

Voorjaarssemester 1981

Sept '81

Inhoudsopgave

0. Inleiding

0.1. Introductie.

De statistiek wordt overal toegepast waar numerieke waarnemings-resultaten worden verwerkt, met name in de sterrenkunde, biologie, natuur- en scheikunde, de technische wetenschappen, de sociologie en de psychologie. De kansrekening is als basis voor de statistiek onmisbaar. Verder wordt de kansrekening direct toegepast in gebieden als informatietheorie, regeltechniek, reliability, wachttijd- en voorraadproblemen en problemen in de natuurkunde, o.a. bij deeltjestellers.

Stellingen, definities en formules zijn in dit dictaat per hoofdstuk genummerd. Waar over de nummering verwarring zou kunnen ontstaan, is bij de verwijzing het bladzijnummer toegevoegd.

Het is duidelijk dat in een college van 9 of 14 maal anderhalf uur maar een heel bescheiden introductie in de kansrekening en statistiek kan worden gegeven. Voor hen die zich wat verder willen oriënteren volgt hier een lijstje van (meest engelstalige) boeken, gerangschikt in afnemende overeenkomst met dit dictaat, en een lijst van veel voorkomende engelse termen. In de genoemde boeken staan vele verwijzingen naar verdere literatuur.

Verdere hulpmiddelen bij de bestudering van de collegestof zijn de Vraagstukkenverzameling en de Statistische Tabellen.

0.2. Literatuur.

[1] Paul L. Meyer, Introductory probability and statistical applications, Addison-Wesley, 1972 (ca. ƒ 35,--).

[2] C. Mack, Essentials of statistics for Scientists and Technologists, Plenum Press, New York, 1967 (ca. ƒ 25,--).

[3] K.L. Chung, Elementary probability theory with stochastic processes, Springer, 1974 (ca. ƒ 45,--).

[4] A.J. Stam, Inleiding tot de waarschijnlijkheidsrekening, Technische Uitgeverij H. Stam, 1964 (ca. ƒ 45,--).

[5] W. Feller, An introduction to probability theory and its applications, Vol. 1 (3e editie) Wiley, 1968 (ca. ƒ 35,--).

0.3. <u>Lijst van Engelse termen.</u>

bias	*onzuiverheid*
bias(s)ed	*onzuiver*
conditional	*voorwaardelijk*
confidence interval	*betrouwbaarheidsinterval*
critical region	*kritiek gebied*
density (function)	*dichtheid(s)functie*
dependence	*afhankelijkheid*
dependent	*afhankelijk*
disjoint	*disjunct*
distribution	*verdeling*
distribution function	*verdelingsfunctie*
estimate	*schatting*
estimator	*schatter*
event	*gebeurtenis*
expectation	*verwachting*
Gaussian	*normaal (genoemd naar Gauss)*
independence	*onafhankelijkheid*
independent	*onafhankelijk*
intersection	*doorsnede (van verzamelingen)*
interval estimate	*intervalschatting (betr. interval)*
joint density	*simultane dichtheid*
joint distribution	*simultane verdeling*
joint distribution function	*simultane verdelingsfunctie*
law of large numbers	*wet van de grote aantallen*
level of significance	*betrouwbaarheidsdrempel*
moment generating function	*momentenvoortbrengende functie*
null hypothesis	*nulhypothese*
power	*onderscheidingsvermogen*
probability	*kans, waarschijnlijkheid*
probability density	*kansdichtheid*
probability space (- field)	*kansruimte (-veld)*
random	*stochastisch, aselect*
random sample	*aselecte steekproef*

random variable	*stochastische grootheid*
rectangular distribution	*homogene verdeling*
sample	*steekproef*
sample mean	*steekproefgemiddelde*
sample space	*uitkomstenruimte*
sample variance	*steekproefvariantie*
sampling with (without) replacement	*trekken met (zonder) terugleggen*
statistic	*statistische grootheid*
test (of significance)	*toets*
test statistic	*toetsingsgrootheid*
unbias(s)ed	*zuiver*
uniform distribution	*homogene verdeling*
union	*vereniging (van verzamelingen)*
variance	*variantie*

1 Kansvelden

1.1. Experimenten en kansvelden.

In de kansrekening houden we ons bezig met de studie van *modellen* van situaties (experimenten, verschijnselen), waarbij "het toeval" een rol speelt. Bij deze modellen blijft het begrip "toeval" onbesproken en dikwijls ook het mechanisme achter de verschijnselen dat tot het model leidt. Het model geeft meestal alleen aan welke gebeurtenissen kunnen optreden en wat de kansen op deze gebeurtenissen zijn. Soms kan uit de beschrijving van de beschouwde situatie een volledig model worden afgeleid, maar meestal zijn extra gegevens (bijvoorbeeld uit vroegere ervaring) of aannamen nodig om een model compleet te maken (zie Voorbeeld 4). We zullen een algemeen model introduceren aan de hand van enkele eenvoudige voorbeelden.

Voorbeeld 1 (dobbelsteen) : iemand gooit één maal met één dobbelsteen.

Vraag: wat is de kans op minstens 3 ogen?

Voorbeeld 2 (wachten op een bus) : iemand komt bij een bushalte aan en leest daar "15-minutendienst".

Vraag: wat is de kans dat hij langer dan 10 minuten moet wachten?

Voorbeeld 3 (toevalstreffer) : een zeer matig schutter mikt op een cirkelvormige schijf totdat hij deze treft.

Vraag: wat is de kans dat zijn treffer dichter bij het middelpunt van de schijf terecht komt dan bij de rand?

Voorbeeld 4 (scherpschutter) : een zeer goede schutter schiet één maal op dezelfde schijf.

Vraag: als bij voorbeeld 3.

Bij elk van de voorbeelden is een verzameling van mogelijke (toegelaten) resultaten of *uitkomsten*. Zo'n verzameling noemen we *uitkomstenruimte* en geven we aan met U. In de voorbeelden hebben we achtereenvolgens:

Vb. 1 : $U = \{1,2,3,4,5,6\}$

Vb. 2 : $U = (0,15)$

Vb. 3 : $U = \{(x,y) \mid x^2 + y^2 \leq 1\}$ (straal van de schijf gelijk aan 1 gekozen)

Vb. 4 : $U = \mathbb{R}^2 = \{(x,y) \mid -\infty < x < \infty, -\infty < y < \infty\}$

In het laatste voorbeeld laten we de mogelijkheid toe dat de scherpschutter de schijf mist (al zal de kans daarop klein zijn). Voor alle zekerheid (en ook terwille van wiskundige stroomlijn) nemen we alle punten van het platte vlak in U op.

<u>Opmerking</u>: de elementen van de uitkomstenruimte zijn in alle voorbeelden min of meer abstracte beelden van de werkelijke resultaten; zo schrijven we in Vb. 1 $\{1,2,\ldots\}$ i.p.v. {•, ⁚, ...}.

In elk van de voorbeelden bestaat het optreden van de gebeurtenis, waarvan de bijbehorende kans gevraagd wordt, daarin dat de uitkomst van het experiment in een bepaalde deelverzameling, zeg A, van U terecht komt, nl. achtereenvolgens

Vb. 1 : $A = \{3,4,5,6\}$ = "minstens 3 ogen".

Vb. 2 : $A = (10,15)$ = "meer dan 10 minuten".

Vb. 3/4 : $A = \{(x,y) \mid \sqrt{x^2 + y^2} < \frac{1}{2}\}$ = "afstand tot middelpunt kleiner dan ½".

Het ligt nu voor de hand om gebeurtenissen met de bijbehorende deelverzamelingen te identificeren, dus

gebeurtenis = deelverzameling van U.

De elementen $u \in U$, of ook wel de één-punts-deelverzamelingen $\{u\}$ van U noemen we elementaire gebeurtenissen.

Het model werkt nu als volgt: door een (dikwijls onbekend) "toevalsmechanisme" komt een element $u \in U$ als uitkomst van het experiment te voorschijn. Als u een element van A is ($u \in A$), dan zeggen we dat A optreedt. Dikwijls wordt "A treedt op" i.p.v. met $u \in A$ met alleen A aangegeven. In plaats van de verzamelingstheoretische terminologie gebruiken we dikwijls een terminologie die aansluit bij de interpretatie van verzamelingen als gebeurtenissen:

Notatie	Verzamelingen	Gebeurtenissen
$u \in A$	u is element van A	A gebeurt
A	deelverzameling A (van U)	gebeurtenis A (ook: A gebeurt)
$A \cup B$	vereniging van A en B	A of B(gebeurt); minstens één
$\cup A_j$	vereniging van A_j's	minstens één van de A_j's (gebeurt)
$A \cap B$ of AB	doorsnede van A en B	A en B(gebeuren); beide
$\cap A_j$	doorsnede van A_j's	A_j's (gebeuren) allemaal
A^*	complement (t.o.v. U)	A (gebeurt) niet; niet-A
U	hele verzameling	zékere gebeurtenis
$\emptyset$	lege verzameling	onmogelijke gebeurtenis
$A \cap B = \emptyset$	A en B zijn disjunct	A en B sluiten elkaar uit
$A \subset B$	A is bevat in B	A impliceert B (als A gebeurt dan gebeurt B ook)

Eigenschappen[1] van gebeurtenissen (verzamelingen):

1. $A \cap U = A \cup \emptyset = A$; $A \cup U = U$; $A \cap \emptyset = \emptyset$

2. $A = AB \cup AB^*$; $A \cup A^* = U$; $A \cap A^* = \emptyset$

3. $(A \cap B)^* = A^* \cup B^*$; $(A \cup B)^* = A^* \cap B^*$

4. $A \cup B = A \cup A^*B = B \cup AB^*$

5. $(A \cup B)C = AC \cup BC$; $(\overset{\infty}{\underset{1}{\cup}} A_j)C = \overset{\infty}{\underset{1}{\cup}} (A_jC)$

[1] Voor overige eigenschappen zie Wsk. 10.

Wat nog aan onze modellen ontbreekt is de *kans* op de daar optredende gebeurtenissen. Omdat gebeurtenissen deelverzamelingen zijn (van U), zal voor de kans, die we aangeven met P, moeten gelden:

P is een functie op de deelverzamelingen van U.

We introduceren nu de eigenschappen van de functie P aan de hand van Voorbeeld 3. Onze schutter schiet zo slecht, dat het "volmaakt toevallig" is in welk punt van de schijf de treffer terecht komt: de kans dat de treffer binnen een gebied (deelverzameling) A van de schijf valt, hangt niet van de plaats van A af, maar alleen van de grootte van A. We geven dit weer door

a. P(A) is evenredig met opp(A).

Omdat de schutter dóórgaat tot hij de schijf (U) raakt, is het zeker dat dit gebeurt. Deze zekerheid geven we aan met

b. $P(U) = 1$

Soms wordt zekerheid ook weergegeven met 100 (%). Omdat $\text{opp}(U) = \pi$, volgt uit a. en b. dat

c. $P(A) = \frac{1}{\pi}\,\text{opp}(A)$.

Het model voor Voorbeeld 3 is nu compleet. We kunnen nu ook de daar gestelde vraag beantwoorden:

$P(\text{treffer dichter bij middelpunt dan bij rand}) = P(\{(x,y) \mid x^2+y^2 < \frac{1}{4}\}) =$

$\frac{1}{\pi}\,\text{opp(cirkel met straal } \frac{1}{2}) = \frac{1}{\pi}\,\frac{\pi}{4} = \frac{1}{4}$.

<u>Opgave</u>: ga na dat de door c. gegeven kans de volgende eigenschappen heeft:

(i) $P(U) = 1$

(1) (ii) $P(A) \geq 0$

(iii) $P(A \cup B) = P(A) + P(B)$ <u>als</u> $A \cap B = \emptyset$;

hierbij zijn A en B gebeurtenissen, dus deelverzamelingen van U.

We geven nu zonder veel commentaar de kansen bij de overige voorbeelden en *antwoorden op de daar gestelde vragen*

Vb. 1 : $P(A) = \frac{\#(A)}{\#(U)} = \frac{1}{6}\#(A)$,

waarbij $\#(A)$ het aantal elementen van A voorstelt.

Antwoord: P(minstens 3 ogen) = $P(\{3,4,5,6\}) = \frac{1}{6}\#(\{3,4,5,6\}) = \frac{2}{3}$.

Vb. 2 : $P(A) = \frac{\ell(A)}{\ell(U)} = \frac{1}{15}\ell(A)$.

waarbij $\ell(A)$ de "lengte" van A voorstelt : $\ell(A) = \int_A du$.

Antwoord: P(meer dan 10 minuten wachten) = $P((10,15)) = \frac{1}{15}\int_{10}^{15} du = \frac{1}{3}$.

Vb. 3 : We kunnen P(A) ook schrijven als (zie c. , blz. 4).

$$P(A) = \iint_A \frac{1}{\pi}\,dxdy.$$

Vb. 4 : $$P(A) = \frac{1}{2\pi\sigma^2}\iint_A e^{-(x^2+y^2)/(2\sigma^2)}dxdy.$$

Dit model volgt natuurlijk niet zonder meer uit de beschrijving op blz. 1. Het voldoet echter aan de volgende redelijke eisen: het verandert niet bij rotatie van de schijf; de kans op treffers is het grootst in het midden van de schijf; bij geschikte keuze van σ is de kans om de schijf te missen klein: 0,001 als $\sigma = 0{,}269$. Bij deze waarde van σ is de gevraagde kans: $P(\{(x,y) \mid x^2 + y^2 < \frac{1}{4}\}) = 0{,}822$.

Opgave: ga na dat de functie P in de voorbeelden 1,2 en 4 ook aan (1) voldoet.

We geven nu, naar analogie van de behandelde voorbeelden, een formele definitie van het begrip *kansveld (of kansruimte)*, d.i. een model voor een experiment, waarbij het toeval een rol speelt.

Definitie 1: Een kansveld bestaat uit een verzameling U, uitkomstenruimte (waarvan de elementen de mogelijke uitkomsten van het beschouwde experiment voorstellen) en een functie P gedefinieerd op deelverzamelingen van U, (gebeurtenissen), die voldoet aan

(1')
(i) $P(U) = 1$

(ii) $P(A) \geq 0$ voor alle gebeurtenissen A

(iii) $P(A \cup B) = P(A) + P(B)$ als $A \cap B = \emptyset$

(iii') $P(\bigcup_1^\infty A_j) = \sum_1^\infty P(A_j)$ als $A_i \cap A_j = \emptyset$ voor $i \neq j$;

een functie die aan (1') voldoet heet een *kans*.

Uit de eigenschappen (1') kunnen een aantal andere, veel gebruikte eigenschappen worden afgeleid.

Stelling 1: een kans heeft de volgende eigenschappen:

a. $0 \leq P(A) \leq 1$

b. $P(A^*) = 1 - P(A)$

c. $P(A) = P(AB) + P(AB^*)$

d. $P(A \cup B) = P(A) + P(B) - P(AB)$

e. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$.

Hierbij zijn A,B en C willekeurige gebeurtenissen.

Opgave: bewijs Stelling 1.

1.2. Discrete kansvelden, combinatoriek

Definitie 2: een discreet kansveld is een kansveld, waarbij U eindig of aftelbaar veel elementen heeft, dus $U = \{u_1, u_2, \ldots\}$. De kans wordt vastgelegd door

(2) $$P(\{u_j\}) = p_j,$$

waarbij

(3) $$p_j \geq 0 \text{ en } \sum_1^\infty p_j = 1.$$

Gevolg: voor een discreet kansveld geldt

(4) $$P(A) = \sum_{u_j \in A} p_j \ ,$$

waarbij $\sum_{u_j \in A}$ de som voorstelt over alle waarden van j waarvoor $u_j \in A$.

Opgave: ga na dat (4) volgt uit (2) m.b.v. (1')

We komen in paragraaf 1.3 op deze discrete kansvelden terug. We bekijken nu eerst een nog specialer geval.

Definitie 3: een *symmetrisch kansveld* is een discreet kansveld, waarbij U *eindig* veel elementen heeft : $U = \{u_1, u_2, \ldots, u_N\}$, terwijl voor de p_j geldt

(5) $$p_j = \frac{1}{N} \quad (j = 1,2,\ldots,N).$$

Gevolg: in een symmetrisch kansveld geldt

(6) $$P(A) = \frac{\#(A)}{\#(U)}$$

Opmerking: in een symmetrisch kansveld geldt dus: de kans op gebeurtenis A is gelijk aan het quotient van het aantal uitkomsten dat aanleiding geeft tot A (d.w.z. tot A behoort) en het totaal aantal uitkomsten. In 1814 werd al een soortgelijke definitie gegeven door Laplace: "de kans op een gebeurtenis is de verhouding van het aantal voor die gebeurtenis gunstige gevallen tot het totale aantal gevallen, mits alle gevallen even waarschijnlijk zijn." Deze definities sluiten aan bij de intuïtieve interpretatie van het begrip kans als *fractie*: als we vaak met een dobbelsteen gooien, zal de fractie van het aantal worpen dat een zes oplevert, ongeveer gelijk zijn aan de kans op een zes. Deze uitspraak wordt later gepreciseerd in de wet van de grote aantallen (paragraaf 4.2).

In een symmetrisch kansveld komt het berekenen van kansen neer op het *tellen* van aantallen uitkomsten. Dit tellen noemt men wel *combinatoriek* en het bijbehorende onderdeel van de kansrekening combinatorische kansrekening.

<u>Voorbeeld</u>: wat is de kans op precies 9 ogen bij een worp met twee dobbelstenen ?

<u>Antwoord</u>: als model gebruiken we een symmetrisch kansveld met 36 uitkomsten $U = \{(x_1, x_2) \mid x_1 = 1,2,\ldots,6;\ x_2 = 1,2,\ldots,6\}$ en (dus) voor $P(\{(x_1, x_2)\}) = \frac{1}{36}$ voor iedere $(x_1, x_2) \in U$. Als nu A = "precies 9 ogen", dan is dus $A = \{(3,6), (4,5), (5,4), (6,3)\}$, zodat $P(A) = \frac{1}{36} \# (A) = \frac{4}{36} = \frac{1}{9}$.

<u>Opgave 1</u>: wat is de kans op minstens 9 ogen ?

<u>Opgave 2</u>: we gooien met *twee* dobbelstenen, I en II. Door welke verzameling wordt de gebeurtenis "I levert 5 ogen" voorgesteld ? Wat is de kans op deze gebeurtenis ?

We bespreken nu een aantal veel voorkomende situaties, waarbij symmetrische kansvelden als model worden gebruikt, n.l. het *trekken van een steekproef*. Hierbij gaan we uit van een verzameling van n verschillende (althans, onderscheidbare) "objecten" of "elementen" (dit kunnen in de praktijk knikkers, studenten, transistoren, ongelukken of pakjes boter zijn, e.d.). Uit deze n objecten wordt een keuze gemaakt van k stuks, die in de steekproef worden opgenomen. De uitkomstenruimte U bij dit experiment bestaat dan uit alle mogelijke k-tallen en het eerste probleem is om $\#(U)$ te berekenen. Omdat in het model van het symmetrische kansveld al deze steekproeven (k-tallen) dezelfde kans hebben, spreekt men wel van *aselecte steekproeven.* Hierbij onderscheiden we vier verschillende situaties, die ontstaan door de indeling in: trekken *met* of *zonder* terugleggen en *geordende* of *ongeordende* steekproeven.

Bij trekken *met* terugleggen wordt telkens een objekt gekozen, geregistreerd (d.w.z. in de steekproef opgenomen) en weer teruggelegd, zodat een object meer dan eens in de steekproef kan voorkomen; bij trekken *zonder* terugleggen verdwijnt het gekozen object uit de verzameling, zodat elk object maar één keer gekozen kan worden: alle elementen van de steekproef zijn dan verschillend.

Twee *geordende* steekproeven noemen we gelijk, als zij dezelfde objecten bevatten en als die objecten bovendien in *dezelfde volgorde* gekozen zijn. De objecten worden hier dus één voor één gekozen en hier is bijv. $\{a,b,c\} \neq \{a,c,b\}$. Twee *ongeordende* steekproeven noemen we gelijk als zij dezelfde elementen bevatten; we letten hierbij niet op de volgorde en de steekproefelementen kunnen hetzij één voor één of (bij trekken zonder terugleggen) alle tegelijk gekozen (gedacht) worden. Hier geldt dus $\{a,b,c\} = \{a,c,b\}$ etc.

We formuleren eerst zonder bewijs de volgende voor de hand liggende

Telregel: Als men k maal een keuze doet, waarbij het aantal mogelijkheden bij de 1^e, $2^e, \dots, k^e$ keuze resp. $n_1, n_2, \dots, n_k$ bedraagt, dan is het totaal aantal mogelijke keuzen $n_1 n_2 \dots n_k$.

Uit deze regel volgt gemakkelijk

Stelling 2: het aantal *geordende* steekproeven van de grootte k bij trekking *zonder* terugleggen uit een verzameling van n objecten is

$$n^{(k)} := n(n-1)(n-2)\dots(n-k+1) = \frac{n!}{(n-k)!} \quad ^{1)}$$

Bewijs: er zijn voor het 1^e, $2^e, \dots, k^e$ element van de steekproef resp. $n, n-1, \dots, n-k+1$ mogelijke keuzen.

Gevolg: het aantal rangschikkingen (volgorden, "permutaties") van n objecten bedraagt $n! = n(n-1)(n-2)\dots 2.1$.

Bewijs: dit aantal is gelijk aan het aantal geordende "steekproeven" van de grootte n bij trekken zonder terugleggen uit n objecten. Neem dus $k = n$ in Stelling 2.

Stelling 3: het aantal *geordende* steekproeven van de grootte k bij trekking *met* terugleggen uit een verzameling van n objecten is

$$n^k.$$

Bewijs: bij ieder van de k keuzen zijn er n mogelijkheden.

Stelling 4: het aantal *ongeordende* steekproeven van de grootte k bij trekken *zonder* terugleggen uit een verzameling van n objecten is

$$\binom{n}{k},$$

---

1) In $L := R$ wordt L door R *gedefinieerd*.

waarbij $\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{n(n-1)\ldots(n-k+1)}{k!}$

Bewijs: bij de $n(n-1)\ldots(n-k+1)$ geordende steekproeven komt ieder k-tal elementen in alle $k!$ volgorden voor (vergelijk Gevolg van Stelling 2). Omdat we bij ongeordende steekproeven al deze $k!$ k-tallen als gelijk beschouwen, moeten we het resultaat van Stelling 2 door $k!$ delen.

Gevolg 1: uit een verzameling van n elementen kunnen we op $\binom{n}{k}$ manieren k elementen, d.w.z. een deelverzameling van de grootte k, kiezen.

Gevolg 2:

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

(vergelijk Wsk. 10).

Opmerking: het geval "trekken *met* terugleggen en *ongeordende* steekproeven" ligt als model minder voor de hand en wordt hier niet behandeld.

Opgave: uit een verzameling U met n elementen kan op 2^n manieren een deelverzameling worden gekozen (U en $\emptyset$ meegeteld). Bewijs dit op twee manieren: 1^e direct m.b.v. de Telregel en 2^e m.b.v. Gevolg 1 en Gevolg 2.

Stelling 5: als uit een verzameling van N elementen, waaronder M elementen van type I en N-M van type II, een aselecte steekproef ter grootte n wordt genomen, dan geldt: de kans dat de steekproef precies k elementen van type I bevat (en dus n - k van type II)is gelijk aan (zie Stelling 2 voor notatie onder (b)).

(a) $\binom{n}{k} \left(\frac{M}{N}\right)^k \left(1 - \frac{M}{N}\right)^{n-k}$ bij trekken *met* terugleggen en *geordende* steekproeven

(b) $\dfrac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} = \binom{n}{k} \dfrac{M^{(k)}(N-M)^{(n-k)}}{N^{(n)}}$ bij trekken *zonder* terugleggen (zowel bij *geordende* als bij *ongeordende* steekproeven)

Bewijs: Zij A = "steekproef bevat precies k elementen van type I" =

de verzameling van steekproeven van n stuks met precies k elementen van type I. U ="zékere gebeurtenis" = de verzameling van *alle* steekproeven van n stuks. Volgens (6) is $P(A) = \#(A)/\#(U)$. In geval (a) is volgens Stelling 3 $\#(U) = N^n$. Om $\#(A)$ te vinden, splitsen we de keuze van de steekproef in drie achtereenvolgende keuzen : 1^e kies de k plaatsen in de steekproef waar de elementen van type I komen, 2^e kies de elementen van type I met terugleggen uit de M aanwezige elementen van type I, 3^e kies overige n - k elementen met terugleggen uit de N - M elementen van type II. Dit levert o.g.v. de Telregel

$$\#(A) = n_1 n_2 n_3,$$

waarbij $n_1 = \binom{n}{k}$ (zie Stelling 4), $n_2 = M^k$ en $n_3 = (N-M)^{n-k}$ (zie Stelling 3). Dit levert

$$\#(A) = \binom{n}{k} M^k (N-M)^{n-k}$$

en deling door N^n levert de onder (a) gevraagde kans. De kans onder (b) vinden we in het geval van geordende steekproeven op dezelfde manier en in het geval van ongeordende steekproeven direct d.m.v. de Telregel. In beide gevallen gebruiken we natuurlijk ook (6).

Voorbeeld: wat is de kans op "precies 11 goed" bij de voetbaltoto als we één kolom "op goed geluk" invullen ?

Antwoord: we nemen als model: dertien trekkingen met terugleggen uit de elementen g, f_1, f_2, waarbij g "goed" is en f_1 en f_2 "fout" zijn. Het antwoord wordt dan o.g.v. Stelling 5, (a)

$$P(11 \text{ goed}) = \binom{13}{11} \left(\frac{1}{3}\right)^{11} \left(\frac{2}{3}\right)^2 = 0{,}0002.$$

Opgave 1: Bereken P(minstens 11 goed)

Opgave 2: In een vijver zitten 200 vissen. Men vangt er 100, merkt ze en zet ze terug. Vervolgens vangt men (zonder terugleggen) 20 vissen. Wat is de kans dat onder deze 20 minder dan 10 gemerkte vissen voorkomen ?

<u>Opmerking</u>: men kan $\binom{n}{k}$ opvatten als het aantal manieren om een verzameling van n elementen te splitsen in twee groepen: één van k elementen en één van n-k. Analoog geldt

<u>Stelling 6</u>: het aantal manieren waarop een verzameling van n elementen kan worden gesplitst in k groepen met resp. $n_1, n_2, \ldots, n_k$ elementen $(n_1 + \ldots + n_k = n)$ is gelijk aan

$$\frac{n!}{n_1!\,n_2!\,\ldots\,n_k!}$$

<u>Opgave 1</u>: bewijs Stelling 6 m.b.v. de Telregel. Als er onder de getallen n_1, $n_2, \ldots, n_k$ gelijke voorkomen, ontstaat een complicatie; welke ?

<u>Opgave 2</u>: a. Op hoeveel manieren kunnen we 52 speelkaarten onder vier spelers verdelen (elke speler krijgt dertien kaarten).
b. Op hoeveel manieren kunnen we 52 kaarten in 4 stapeltjes van 13 stuks verdelen ?

1.3. <u>Reële kansvelden, stochastische grootheden.</u>

In de meeste praktische gevallen zijn we geïnteresseerd in de *getalwaarde* van grootheden die bij een experiment optreden, zoals lengte, gewicht, tijdsduur, aantal etc. We houden ons daarom voornamelijk bezig met kansvelden, waarbij de uitkomstenruimte bestaat uit *reële getallen*, paren reële getallen of, algemener, uit reële vectoren $(x_1, \ldots, x_n)$.

We bekijken eerst het geval $n = 1$ en beperken ons tot twee veel voorkomende gevallen, het zgn. *discrete geval* en het zgn. *continue geval.*

<u>Definitie 4</u>: Een *discreet reëel kansveld* is een discreet kansveld (zie Definitie 2), waarbij de uitkomstenruimte U uit eindig of aftelbaar veel *reële* getallen bestaat : $U = \{u_1, u_2, \ldots\}$, met $u_1, u_2, \ldots$ reëel.

Zo'n kansveld is een model voor een experiment waarbij een grootheid wordt waargenomen, die met kansen $p_1, p_2, \ldots$ de waarden $u_1, u_2, \ldots$ aanneemt. We noemen zo'n grootheid een *discrete toevalsgrootheid* of *discrete stochastische grootheid*. We geven stochastische grootheden aan met onderstreepte letters : $\underline{x}$, $\underline{y}$, $\underline{u}$, $\underline{v}$, $\underline{w}$, etc. In plaats van $P(\{u_j\}) = p_j$ schrijven we nu liever

$$P(\underline{u} = u_j) = p_j$$

en algemener als A een verzameling reële getallen is, $P(\underline{u} \in A)$ i.p.v. $P(A)$, dus (zie (4))

$$P(\underline{u} \in A) = \sum_{u_j \in A} p_j. \tag{7}$$

Hierbij is natuurlijk (zie Definitie 2) $p_j \geq 0$ en $\sum_1^\infty p_j = 1$. Als de verzameling U gegeven is, dan wordt door iedere rij $p_1, p_2, \ldots$ met deze eigenschappen een reëel discreet kansveld gedefinieerd en een daarbij behorende stochastische grootheid.

Een voorbeeld van de bovenbeschreven situatie is Voorbeeld 1 op blz. 1 (gooien met een dobbelsteen). Het aantal gegooide ogen is hier de stochastische grootheid : $P(\underline{u} = j) = \frac{1}{6}$ voor $j = 1,2,\ldots,6$. Verder geldt bijvoorbeeld: $P(\underline{u}$ is even$) = P(\underline{u} \in \{2,4,6\}) = \frac{1}{2}$. Bij discrete reële kansvelden is de bijbehorende stochastische grootheid dikwijls een aantal, dus : $U = \{0,1,2,\ldots\}$.

<u>Definitie 5</u>: een *continu reëel kansveld* is een kansveld waarbij $U = (-\infty,\infty)$ en waarbij de kans gegeven wordt door

$$P(A) = \int_A f(u)du. \tag{8}$$

Hier is f een functie waarvoor $f(u) \geq 0$ voor alle u, terwijl $\int_{-\infty}^{\infty} f(u)du = 1$ (beide eigenschappen volgen uit (8) en (1'), blz.6).
De functie f heet *kansdichtheid.*

Een continu reëel kansveld is model voor een situatie waarbij een toevalsgrootheid optreedt die continu kan variëren, zoals lengte of tijdsduur. We noemen een dergelijke grootheid een *continue stochastische grootheid* en geven deze weer aan met een onderstreepte letter. In plaats van P(A) schrijven we weer (vergelijk (7))

$$P(\underline{u} \in A) = \int_A f(u)du. \tag{9}$$

Iedere functie f met $f(u) \geq 0$ en $\int_{-\infty}^{\infty} f(u)du = 1$ bepaalt een continu reëel kansveld met een bijbehorende stochastische grootheid. Een voorbeeld van een continu reëel kansveld is Voorbeeld 2 op blz. 1 (wachten op een bus); hier is $f(u) = \frac{1}{15}$ voor $0 < u < 15$ en gelijk aan nul voor alle andere waarden van u. De stochastische grootheid is hier de wachttijd van het moment van aankomst tot de komst van de bus.

Opmerking 1: als speciaal geval van (9) hebben we

$$P(a \leq \underline{u} \leq b) = \int_a^b f(u)du,$$

en voor a = b dus : $P(\underline{u} = a) = \int_a^a f(u)du = 0$. Bij een continue stochastische grootheid is de kans op iedere vaste waarde gelijk aan nul.

Opmerking 2: als een stochastische grootheid alleen waarden kan aannemen in een interval (u_1,u_2) dan kunnen we dit in ons model weergeven door óf $U = (u_1,u_2)$ te kiezen i.p.v. $U = (-\infty,\infty)$ of door $f(u) = 0$ te nemen buiten (u_1,u_2). We zullen het laatste als regel kiezen. Omgekeerd kunnen we in Definitie 4 ook i.p.v. $U = \{u_1,u_2,...\}$ nemen $U = (-\infty,\infty)$, als we $P(A) = 0$ definiëren voor iedere A die geen enkele van de u_j bevat. Dit is in overeenstemming met formule (7), als we voor de "lege som" nul lezen.

Definitie 6: als f een kansdichtheid is, dan noemen we de functie F gedefinieerd door

$$(10) \qquad F(x) = \int_{-\infty}^{x} f(u)du$$

een *verdelingsfunctie*. Als f de kansdichtheid is van een stochastische grootheid $\underline{u}$, dan heeft F(x) de interpretatie

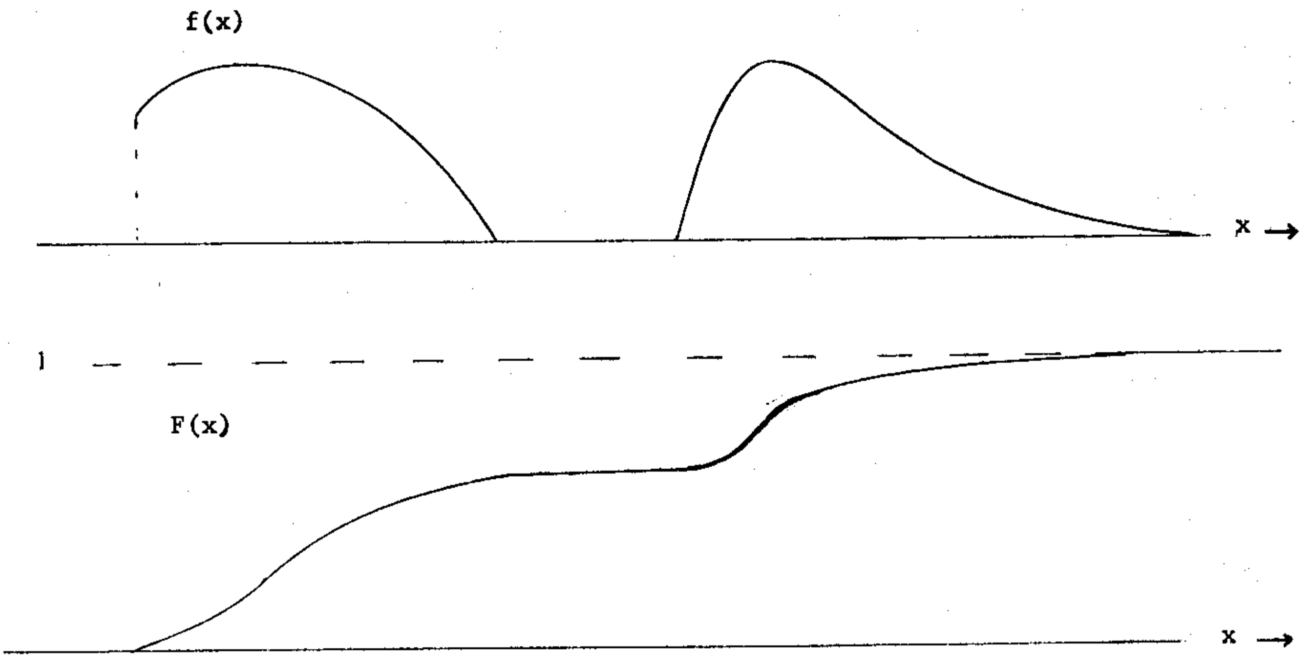
$$(11) \qquad P(\underline{u} \leq x) = F(x).$$

De reden dat men de verdelingsfunctie invoert is juist dat deze, in tegenstelling tot f(x), *een kans* voorstelt. Uit (10) volgt direct, voor waarden van x waar f continu is (zie Wiskunde 10),

$$(12) \qquad f(x) = F'(x)$$

en uit (11) (vergelijk Stelling 1, c, blz. 6)

$$(13) \qquad P(a < \underline{u} \leq b) = \int_a^b F'(u)du = F(b) - F(a)$$



$$P(a < \underline{u} \leq b) = \int_a^b f(x)dx = F(b) - F(a)$$

Ook bij discrete kansvelden kunnen we m.b.v. (11) een verdelingsfunctie invoeren (vergelijk (7))

$$F(x) = P(\underline{u} \leq x) = \sum_{u_j \leq x} p_j ;$$

dit heeft echter weinig voordelen, omdat de p_j al *kansen* zijn. Men doet het soms toch, om beide gevallen gemeenschappelijk te kunnen behandelen, en omdat men in kansen van de vorm $P(\underline{u} \leq x)$ geïnteresseerd is.

Notatie: kansdichtheden en hun verdelingsfuncties worden aangegeven met f,g,h,... resp. F,G,H,... . Dikwijls zullen we kansdichtheden en verdelingsfuncties *behorend bij stochastische grootheden* $\underline{x},\underline{u},\underline{v},\ldots$ aangeven met $f_{\underline{x}},f_{\underline{u}},f_{\underline{v}},\ldots$, resp. $F_{\underline{x}},F_{\underline{u}},F_{\underline{v}},\ldots$ Vaak gebruiken we dan corresponderende letters als variabele: $F_{\underline{x}}(x) = P(\underline{x} \leq x)$ etc.

Opgave: ga na dat een verdelingsfunctie F de volgende eigenschappen heeft:

a. $F(x)$ is niet-dalend

b. $\lim_{x \to -\infty} F(x) = 0$

c. $\lim_{x \to \infty} F(x) = 1$

Opmerking: de functie $P(A) = P(\underline{u} \in A)$ noemt men de *(kans-)verdeling* van $\underline{u}$; deze geeft aan hoe de totale kans 1 over $(-\infty,\infty)$ *verdeeld* is. Bij discrete $\underline{u}$ spreekt men over een discrete verdeling, bij continue $\underline{u}$ over een continue verdeling. De kansverdeling van $\underline{u}$ wordt bepaald door $P(\underline{u} = u_j)$ $(j = 1,2,\ldots)$, door $f_{\underline{u}}$ of door $F_{\underline{u}}$.

Voorbeelden

1. Discrete homogene verdeling

$$P(\underline{u} = x_k) = \frac{1}{n} \qquad (k = 1,2,\ldots,n).$$

2. Binomiale verdeling (als $n = 1$ ook *Bernoulli-verdeling* genoemd)

$$P(\underline{u} = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (k = 0,1,\ldots,n;\ 0 \le p \le 1)$$

(vergelijk Stelling 5 (a), p. 10).

3. Poisson - verdeling

$$P(\underline{u} = k) = e^{-\mu} \frac{\mu^k}{k!} \qquad (k = 0,1,2,\ldots;\ \mu > 0).$$

4. Geometrische verdeling

$$P(\underline{u} = k) = (1-p)^{k-1} p \qquad (k = 1,2,3,\ldots;\ 0 < p \le 1).$$

5. Homogene verdeling op (a,b)

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b \\ 0 & \text{anders} \end{cases}$$

$$F(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x < b \\ 1 & x \ge b. \end{cases}$$

6. <u>Gamma - verdeling</u> (ook wel: Erlang-verdeling)

$$f(x) = \begin{cases} 0 & x < 0 \\ \dfrac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} & x > 0 \end{cases} \qquad (n=1,2,\ldots;\ \lambda>0)$$

$$F(x) = \begin{cases} 0 & (x < 0) \\ 1 - \sum\limits_{k=0}^{n-1} e^{-\lambda x} \dfrac{(\lambda x)^k}{k!} & (x \geq 0) \end{cases}$$

7. <u>Exponentiële verdeling</u> (neem n = 1 in Vb. 6)

$$f(x) = \lambda e^{-\lambda x} \qquad (x \geq 0)$$ [1)]

$$F(x) = 1 - e^{-\lambda x} \qquad (x \geq 0)$$

8. <u>Normale verdeling</u>

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad ; \quad F(x) = \int_{-\infty}^{x} f(u)\,du \ .$$

Speciaal met $\mu = 0$ en $\sigma^2 = 1$: $\varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.
Als $\underline{x}$ een normale verdeling heeft met parameters μ en σ^2, dan schrijven we wel "$\underline{x}$ is $N(\mu,\sigma^2)$-verdeeld" of kortweg "$\underline{x} \sim N(\mu,\sigma^2)$". We definiëren nog

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

De functie Φ is getabelleerd in de Statistische Tabellen.

<u>Opgave 1</u>. Als $\underline{x} \sim N(\mu,\sigma^2)$ is, dan is $(\underline{x} - \mu)/\sigma \sim N(0,1)$; als $\underline{y} \sim N(0,1)$, dan is $\sigma\underline{y} + \mu \sim N(\mu,\sigma^2)$.

---

1) Dikwijls geven we alleen het gebied aan waar $f(x) > 0$ is.

Opgave 2: controleer dat in de Voorbeelden 1 t.m. 4 geldt

$$\sum_k P(\underline{u} = x_k) = 1 \quad \text{of} \quad \sum_k P(\underline{u} = k) = 1$$

en teken $P(\underline{u} = k)$ als functie van k.

Opgave 3: controleer (zie Wiskunde 10 en 20) dat in de voorbeelden 5 t.m. 8 geldt

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

en schets telkens de grafieken van f(x) en F(x).

Opgave 4: laat zien dat $\Phi(-x) = 1 - \Phi(x)$.

Opmerking: Reële kansvelden en de bijbehorende stochastische grootheden ontstaan soms uit meer gedetailleerde kansvelden, zoals in de volgende voorbeelden:

Voorbeeld (gooien met 2 dobbelstenen): We hebben een symmetrisch kansveld met $U = \{(x,y) \mid x = 1,2,\ldots,6;\ y = 1,2,\ldots,6\}$, waarbij x = aantal ogen met de eerste steen en y = aantal ogen met de tweede steen. We noemen het totaal aantal ogen $\underline{k}$. Blijkbaar is $\underline{k} = x + y$ en voor $k = 1,2,\ldots,13$ $P(\underline{k} = k) = P(\{(x,y) \mid x + y = k\}) = \dfrac{6 - |7-k|}{36}$ (ga na).

Voorbeeld (toevalstreffer ; zie blz. 1) : we hebben een kansveld met $U = \{(x,y) \mid x^2 + y^2 \le 1\}$ en $P(A) = \frac{1}{\pi}$ opp (A). Zij $\underline{z}$ de afstand van de treffer tot (0,0), dan geldt $\underline{z} = \sqrt{x^2 + y^2}$ en

$$F_{\underline{z}}(z) = P(\underline{z} \le z) = P(\{(x,y) \mid x^2 + y^2 \le z^2\}) = z^2 \qquad (0 \le z \le 1)$$

(ga na), en dus

$$f_{\underline{z}}(z) = 2z \qquad (0 \le z \le 1).$$

In deze voorbeelden kan de stochastische grootheid beschouwd worden als *een functie* op de uitkomstenruimte U. In veel gevallen is een dergelijk gedetailleerd model niet beschikbaar en werken we direct met de Definities 4 en 5.

2 Voorwaardelijke kans en onafhankelijkheid

2.1. Voorwaardelijke kans.

Voorbeeld: We gooien met twee dobbelstenen (zie Voorbeeld op blz.8). We kijken niet, maar iemand vertelt ons "er ligt minstens één zes bij". Wat is onder deze omstandigheden de kans op "minstens 9 ogen"? Deze "kans" is eigenlijk niet gedefinieerd, maar de volgende redenering ligt voor de hand : het aantal uitkomsten met "minstens één 6" is 11; *hieronder* zijn er 7 met minstens 9 ogen, n.l. (3,6), (4,6), (5,6), (6,6), (6,5), (6,4) en (6,3). De gevraagde kans is "dus" $\frac{7}{11}$. Deze redenering komt hierop neer: zij C = "minstens één 6" en A = "minstens 9 ogen"; noteer de gevraagde kans, d.i. de kans op A gegeven C als $P(A|C)$; voor $P(A|C)$ hebben we genomen

$$\frac{\text{aantal mog. minstens één 6} \cap \text{minstens 9 ogen}}{\text{aantal minstens een 6}}$$

(1) $$P(A|C) = \frac{\#(A \cap C)}{\#(C)}$$

Door teller en noemer van (1) te delen door $\#(U) = 36$ vinden we ($P(A \cap C) = \#(A \cap C)/36$, $P(C) = \#(C)/36$)

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

We definiëren nu algemeen

Definitie 1: Als $P(C) > 0$ is, dan definiëren we $P(A|C)$, de *voorwaardelijke kans* op A *onder de voorwaarde* (of : *gegeven*) C, door

(2) $$P(A|C) = \frac{P(A \cap C)}{P(C)}$$

Stelling 1: Een voorwaardelijke kans is een kans.

Bewijs: de bewering houdt in , dat $P(A|C)$ voor *vaste* C voldoet aan de eisen (i) t.m. (iii') van Definitie 1.1 . Ga dit na (schrijf voor het gemak even $P_C(A)$ i.p.v. $P(A|C)$).

X Stelling 2: Als $P(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) > 0$, dan geldt

(3) $$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3 | A_1 \cap A_2)\ldots P(A_n| A_1 \cap \ldots \cap A_{n-1}).$$

Opgave 1: bewijs (3) met volledige inductie (gebruik : $P(A \cap C) = P(C)P(A|C)$).

Voorbeeld 1: Uit een bak met 50 knikkers, 30 rode en 20 witte, trekt men zonder terugleggen 2 knikkers. Bereken $P(R_1 \cap W_2)$, waarbij R_1 = "1^e knikker is rood" en W_2 = "2^e knikker is wit".

Oplossing: $P(R_1 \cap W_2) = P(R_1).P(W_2 | R_1) = \frac{30}{50}.\frac{20}{49}.$

Opmerking 1: We gebruiken hier als model "kies bij iedere trekking met gelijke kansen uit de nog resterende knikkers". Dit model is equivalent met het symmetrische kansveld , waarbij U = alle geordende paren. Ook hier vinden we

$$P(R_1 \cap W_2) = \frac{\#(R_1 \cap W_2)}{\#(U)} = \frac{30.20}{50.49}.$$

Opmerking 2: In de meeste gevallen zijn de voorwaardelijke kansen *niet* de grootheden die we willen berekenen, maar (zoals in Voorbeeld 1) *modelgrootheden*, die we gebruiken om onvoorwaardelijke kansen te berekenen.

Opgave 1: bereken in de situatie van Voorbeeld 1 $P(W_1 \cap W_2 \cap R_3 \cap W_4)$.

Stelling 3: Als $A_1, A_2, \ldots, A_n$ gebeurtenissen zijn met $P(A_j) > 0$ $(j = 1,\ldots,n)$, $\bigcup_1^n A_j = U$ en $A_i \cap A_j = \emptyset$ voor $i \neq j$, dan geldt voor een willekeurige gebeurtenis B

(4) $$P(B) = \sum_{j=1}^{n} P(B|A_j)P(A_j) \qquad \text{(ook met } n = \infty)$$

Bewijs: $B = B \cap U$ (zie Eigenschap 1, blz. 3), dus $B = B \bigcup_1^n A_j = \bigcup_1^n BA_j$, met $BA_j \cap BA_i = \emptyset$ voor $i \neq j$. Definitie 1.1 (iii') levert dan

$$P(B) = \sum_1^n P(BA_j) = \sum_1^n P(B|A_j)P(A_j),$$

waarbij in de tweede overgang Definitie 1 wordt gebruikt.

Toepassing ($n = 2$): in de situatie van Voorbeeld 1 geldt

$$P(W_2) = P(W_2 \mid W_1)P(W_1) + P(W_2 \mid R_1)P(R_1) = \frac{19}{49} \cdot \frac{20}{50} + \frac{20}{49} \cdot \frac{30}{50} = \frac{20}{50} = P(W_1).$$

Opmerking: In het symmetrische kansveld vinden we ook (nu o.g.v. symmetrie)

$$P(W_2) = \frac{\text{aantal paren } (x,y) \text{ met } y = W}{\text{aantal paren}} = \frac{\text{aantal paren } (x,y) \text{ met } x = W}{\text{aantal paren}} = P(W_1).$$

Analoog geldt algemeen $P(W_n) = P(W_1)$: de kans dat de n-de knikker wit is, is gelijk aan de kans dat de eerste wit is.

Voorbeeld 1: In Voorbeeld 1.1 (wachten op de bus) kan de man zich na 5 minuten wachten afvragen wat de kans is dat hij nog minstens 5 minuten moet wachten. We noemen de wachttijd $\underline{w}$ en vinden als antwoord (ga na!)

$$P(\underline{w} \geq 10 \mid \underline{w} \geq 5) = \frac{P(\underline{w} \geq 10)}{P(\underline{w} \geq 5)} = \frac{5/15}{10/15} = \frac{1}{2}.$$

Voorbeeld 2: een veelgebruikt *model* voor de beschrijving van de levensduur $\underline{t}$ van technische apparatuur is de exponentiële verdeling (zie Voorbeelden op blz. 17). Hiervoor geldt

$$P(\underline{t} > t) = e^{-\lambda t} \qquad (t > 0),$$

dus de kans op een levensduur groter dan t is $e^{-\lambda t}$. Vragen we nu naar de kans dat de apparatuur , gegeven dat ze op tijdstip s nog "leeft", nog langer dan een tijd t ná s zal leven, dan blijkt (ga na) dat

$$P(\underline{t} > t + s \mid \underline{t} > s) = e^{-\lambda t},$$

dus de zgn. rest-levensduur op tijdstip s heeft dezelfde verdeling als de totale levensduur. Dit verschijnsel is karakteristiek voor de exponentiële verdeling en wordt soms geïnterpreteerd als het ontbreken van slijtage: als de apparatuur op tijdstip s nog werkt, dan is ze nog "als nieuw" (*in dit model*).

Opgave: Bewijs onder de veronderstellingen van Stelling 3 de zgn.

Regel van Bayes: $$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}.$$

2.2. Onafhankelijkheid

Voorbeeld: we trekken één kaart uit een spel van 52 en bekijken de volgende gebeurtenissen : Z = "zwarte kaart", S = "schoppenkaart", H = "hartenkaart" en V = "vrouw". Nu is (ga na)

$$P(S) = \frac{1}{4},\ P(S \mid Z) = \frac{1}{2},\ P(S \mid H) = 0 \text{ en } P(S \mid V) = \frac{1}{4}.$$

Blijkbaar kan $P(A \mid C)$ groter dan, kleiner dan of gelijk aan $P(A)$ zijn. Als $P(A \mid C) = P(A)$, dan heeft het optreden van C geen invloed op de kans dat A gebeurt: A is onafhankelijk van C. Nu is ook C onafhankelijk van A, want als $P(A \mid C) = P(A)$, dan is ook $P(C \mid A) = P(C)$, immers beide gelijkheden zijn equivalent met $P(A \cap C) = P(A)P(C)$ (als $P(A) > 0$ en $P(C) > 0$). We definiëren nu

Definitie 2: twee gebeurtenissen A en B heten *(onderling) onafhankelijk* (afkorting: o.o.) als

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

Definitie 3: de gebeurtenissen $A_1, A_2, \ldots, A_n$ heten o.o. als

$$P(A_i \cap A_j) = P(A_i)P(A_j) \text{ voor alle paren } A_i, A_j \text{ met } i \neq j;$$

$$P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k) \text{ voor alle drietallen } A_i, A_j, A_k$$
$$\text{met } i, j \text{ en } k \text{ verschillend }, \ldots$$

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2) \ldots P(A_n).$$

In het "experiment" van Voorbeeld 1, p. 20 worden *zonder terugleggen* twee knikkers getrokken. Dit experiment bestaat uit twee (deel-) experimenten: het trekken van de eerste knikker en het trekken van de tweede knikker. We zagen al dat

$$P(W_1) = P(W_2) = \frac{20}{50}\ .\ \text{Verder is } P(W_1 \cap W_2) = \frac{20}{50} \cdot \frac{19}{49} \neq P(W_1)P(W_2),$$

zodat W_1 en W_2 *afhankelijk* zijn.

Als we het experiment uitvoeren *met* terugleggen, dan blijken W_1 en W_2 onafhankelijk te zijn: $P(W_1 \cap W_2) = P(W_1)P(W_2) = (\frac{20}{50})^2$. Op dezelfde manier zijn dan ook W_1 en R_2, R_1 en W_2, R_1 en R_2 onafhankelijk. Alle gebeurtenissen m.b.t. het eerste experiment zijn: W_1, R_1, U en $\emptyset$, die m.b.t. het tweede experiment: W_2, R_2, U en $\emptyset$. Elke gebeurtenis m.b.t. het eerste experiment is onafhankelijk van elke gebeurtenis m.b.t. het tweede experiment. Experimenten met deze eigenschap noemen we (onderling) onafhankelijk.

Definitie 4: als n experimenten $E_1, E_2, \ldots, E_n$ gegeven zijn en als $\mathcal{E}_j$ de verzameling is van alle gebeurtenissen die alleen betrekking hebben op experiment E_j $(j = 1,2,\ldots,n)$, dan heten $E_1, E_2, \ldots, E_n$ (onderling) onafhankelijk, als voor iedere keuze van de gebeurtenissen $A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2, \ldots, A_n \in \mathcal{E}_n$, de gebeurtenissen $A_1, A_2, \ldots, A_n$ o.o. zijn (zie Definitie 3).

Opgave: ga na dat bij het voorbeeld "gooien met twee dobbelstenen"(zie blz.8) de experimenten,"gooien met de eerste steen" en "gooien met de tweede steen" onafhankelijk zijn.

Opmerking: meestal wordt onafhankelijkheid in het model *geëist*, d.w.z. men definieert het model zo, dat de beschouwde experimenten onafhankelijk zijn. Als voor elk experiment de kansen $P(A_j)$ met $A_j \in \mathcal{E}_j$ bekend zijn, dan definieert men eenvoudig $P(A_iA_j) = P(A_i)P(A_j)$, $P(A_iA_jA_k) = P(A_i)P(A_j)P(A_k)$ etc.. Dikwijls gaat het hierbij om onafhankelijke herhalingen van één experiment, zoals in het volgende voorbeeld.

Voorbeeld: we beschouwen n o.o. experimenten, elk met kans p op succes(S) en kans $1-p$ op mislukking (M). Voor elk experiment geldt dat $P(S) = p$ en $P(M) = 1-p$. De uitkomsten van het samengestelde experiment zijn rijtjes M -en en S -en :

$$U = \{(x_1,\ldots,x_n) \mid x_1 \in \{S,M\},\ldots,x_n \in \{S,M\}\}.$$

O.g.v. de onafhankelijkheid geldt nu $P(SSM\ldots SM) = P(S)P(S)P(M)\ldots P(S)P(M)$, $= p^k(1-p)^{n-k}$, voor een rijtje dat precies k S -en bevat, Als we het totale aantal successen $\underline{k}$ noemen, dan geldt (zie formule (7), p.13).

$$P(\underline{k} = k) = \sum_k P(x_1,\ldots,x_n),\text{ waarbij } \sum_k \text{ de som is over alle rijtjes die}$$

precies k keer S bevatten. Dit zijn er $\binom{n}{k}$ (zie Gevolg 1 van Stelling 1.4.)

en elk zo'n rijtje heeft kans $p^k(1-p)^{n-k}$. We vinden dus voor $\underline{k}$

$$P(\underline{k}=k) = \binom{n}{k} p^k(1-p)^{n-k} \qquad (k = 0,1,\ldots,n),$$

d.w.z. een binomiale verdeling. Deze verdeling komt in allerlei situaties voor en is daarom getabelleerd (bijv. in de Statistische Tabellen). Voor niet getabelleerde waarden van n en kleine waarden van p ($p \leq 0{,}10$), kan men de binomiale verdeling *benaderen* m.b.v. de Poissonverdeling met $\mu = np$ (zie Voorbeeld 3 op blz. 16). Hiervoor geldt

<u>Stelling 4</u>: als $p = \frac{\mu}{n}$, dan is (voor *vaste* μ en k)

$$\lim_{n\to\infty} \binom{n}{k} p^k(1-p)^{n-k} = e^{-\mu} \frac{\mu^k}{k!}.$$

<u>Bewijs</u>:

$$\binom{n}{k} p^k(1-p)^{n-k} = \left(1-\frac{\mu}{n}\right)^n \frac{\mu^k}{k!} \frac{n(n-1)\ldots(n-k+1)}{n^k} \left(1-\frac{\mu}{n}\right)^{-k}.$$

Het eerste deel van deze uitdrukking gaat voor $n\to\infty$ naar $e^{-\mu}\frac{\mu^k}{k!}$, het tweede deel gaat naar 1. Gebruik: $\lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n = e^x$.

<u>Opmerking</u>: als we heel veel o.o. experimenten beschouwen, waarbij met kleine kans een gegeven gebeurtenis ("succes") optreedt en met grote kans die gebeurtenis niet optreedt ("mislukking"), dan leidt een soortgelijke toepassing van Stelling 4 tot het zgn. *Poisson-proces*, dat het aantal gebeurtenissen (bijv. aanvragen voor telefoongesprekken, schademeldingen, emissies van een radio-actief deeltje etc.) beschrijft dat in de loop van een tijd t optreedt. Voor dit aantal, $\underline{k}(t)$, geldt dan

$$P(\underline{k}(t) = k) = e^{-\mu t} \frac{(\mu t)^k}{k!}. \tag{5}$$

De aantallen gebeurtenissen die in disjuncte tijdsintervallen optreden zijn onderling onafhankelijk (zie volgende paragraaf voor definitie). In de Appendix vindt u een wat uitgebreidere behandeling van het Poisson-proces.

2.3. <u>Stochastische vectoren, onafhankelijke stochastische grootheden</u>

We willen dikwijls twee (of meer) stochastische grootheden tegelijk (simultaan) bekijken, zoals lengte en breedte bij metingen of concentraties van verschillende stoffen in één oplossing e.d. Daarom definiëren we analoog aan de kansverdeling van één stochastische grootheid de (*simultane*) kansverdeling van twee (en meer) stochastische grootheden.

<u>Definitie 5</u>: een paar stochastische grootheden $(\underline{x},\underline{y})$ heeft een *discrete* (simultane) *verdeling*, als geldt

$$(6) \qquad P((\underline{x},\underline{y}) \in A) = \sum_{(x_i,y_j)\in A}\sum p_{ij},$$

waarbij $p_{ij} = P(\underline{x} = x_i,\ \underline{y} = y_j) \geq 0$ en $\sum_i \sum_j p_{ij} = 1$.

Vergelijk (7) op blz. 13. We schrijven $P(\underline{x} = x_i,\ \underline{y} = y_j)$ i.p.v. $P(\underline{x} = x_i$ én $\underline{y} = y_j)$.

<u>Definitie 6</u>: een paar stochastische grootheden $(\underline{x},\underline{y})$ heeft een *continue* (simultane) verdeling, als geldt

$$(7) \qquad P((\underline{x},\underline{y}) \in A) = \iint_A f(x,y)\,dxdy,$$

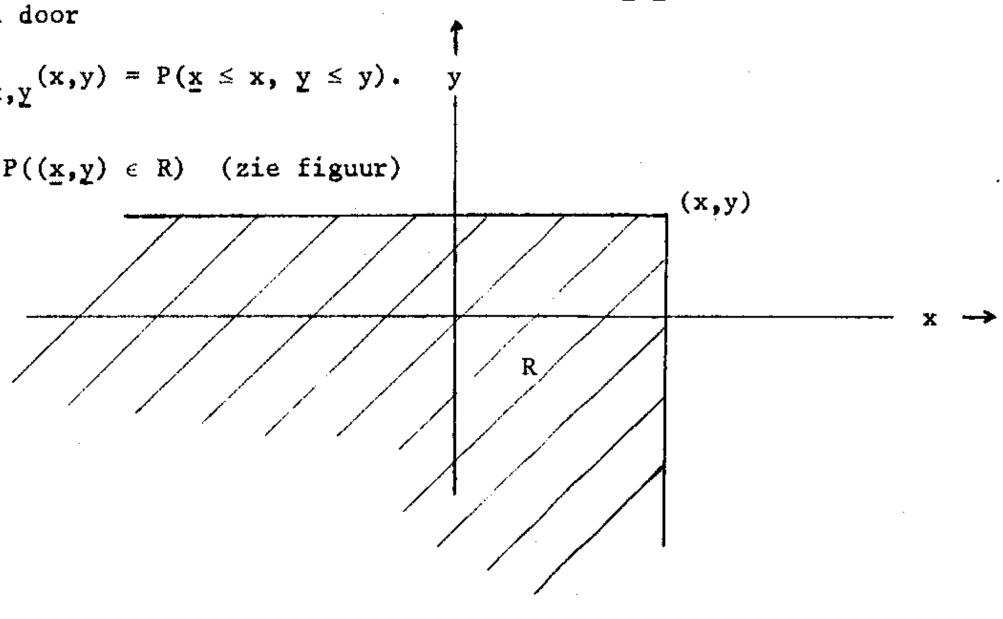
waarbij $f(x,y) \geq 0$ en $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ dxdy = 1.$

De functie f heet de (simultane) *kansdichtheid* van $\underline{x}$ en $\underline{y}$. We schrijven ook wel $f_{\underline{x},\underline{y}}$ i.p.v. f.

<u>Definitie 7</u>: De(simultane) verdelingsfunctie F of $F_{\underline{x},\underline{y}}$ van $\underline{x}$ en $\underline{y}$ wordt gedefinieerd door

$$F_{\underline{x},\underline{y}}(x,y) = P(\underline{x} \leq x,\ \underline{y} \leq y).$$

$F_{\underline{x},\underline{y}}(x,y) = P((\underline{x},\underline{y}) \in R)$ (zie figuur)



We gebruiken de verdelingsfunctie alleen in het continue geval. Dan geldt

$$(8) \qquad F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)du\, dv$$

en (als f(x,y) continu is)

$$(9) \qquad f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$$

De belangrijkste eigenschappen van de verdelingsfunctie geven we in

Stelling 5.

(i) $F_{\underline{x},\underline{y}}(x,\infty) = F_{\underline{x}}(x)$

(ii) $F_{\underline{x},\underline{y}}(\infty,y) = F_{\underline{y}}(y)$

(iii) $F_{\underline{x},\underline{y}}(\infty,\infty) = 1$

(iv) $F_{\underline{x},\underline{y}}(-\infty, y) = F_{\underline{x},\underline{y}}(x, -\infty) = 0.$

Hierbij is $F(x,\infty) = \lim_{y\to\infty} F(x,y)$ etc.

Opgave: Ga na dat stelling 5 jüist is door telkens $F_{\underline{x},\underline{y}}$ te interpreteren als een kans. Zo is $F_{\underline{x},\underline{y}}(x,\infty) = P(\underline{x} \le x,\ \underline{y} \le \infty) = P(\underline{x} \le x) = F_{\underline{x}}(x)$.

Gevolg: uit (i) en (ii) volgt in het continue geval

$$(i') \qquad f_{\underline{x}}(x) = \int_{-\infty}^{\infty} f_{\underline{x},\underline{y}}(x,y)dy$$

$$(ii') \qquad f_{\underline{y}}(y) = \int_{-\infty}^{\infty} f_{\underline{x},\underline{y}}(x,y)dx$$

De verdelingsfuncties $F_{\underline{x}}$ en $F_{\underline{y}}$ van $\underline{x}$ resp. $\underline{y}$ alléén (in (i) en (ii)) noemt men de *marginale verdelingsfuncties* van $\underline{x}$ en $\underline{y}$. De kansdichtheden $f_{\underline{x}}$ en $f_{\underline{y}}$ (zie (i') en (ii')) heten de *marginale kansdichtheden* . Analoog aan (i') geldt voor een paar discreet-verdeelde grootheden $\underline{x}$ en $\underline{y}$

(i") $P(\underline{x} = x_k) = \sum_j P(\underline{x} = x_k, \underline{y} = y_j)$.

Ook hier noemt men de verdeling van $\underline{x}$ alleen de marginale verdeling van $\underline{x}$.

Op geheel analoge wijze definieert men de (simultane) verdelingsfunctie en de (simultane) kansdichtheid van meer dan twee stochastische grootheden. We gaan hierop niet nader in. Formules (6) en (7) worden nu resp.

(6') $$P((\underline{x}_1,\dots,\underline{x}_n) \in A) = \sum_{(x_1,\dots,x_n) \in A} \sum P(\underline{x}_1 = x_1,\dots,\underline{x}_n = x_n)$$

(7') $$P((\underline{x}_1,\dots,\underline{x}_n) \in A) = \int_{(x_1,\dots,x_n) \in A} \int f_{\underline{x}_1,\dots,\underline{x}_n}(x_1,\dots,x_n)\, dx_1 \dots dx_n.$$

In het vervolg zullen we dikwijls te maken hebben met onderling onafhankelijke stochastische grootheden. We geven daarvan nu de definitie:

Definitie 8: Twee stochastische grootheden $\underline{x}$ en $\underline{y}$ heten (onderling) *onafhankelijk* (afkorting o.o.) , als

a. $$P(\underline{x} = x_i,\ \underline{y} = y_j) = P(\underline{x} = x_i)P(\underline{y} = y_j) \qquad (i = 1,2,\dots;\ j = 1,2,\dots)$$

in het discrete geval.

b. $$f_{\underline{x},\underline{y}}(x,y) = f_{\underline{x}}(x).f_{\underline{y}}(y) \qquad (-\infty < x < \infty;\ -\infty < y < \infty)$$

in het continue geval.

Gevolg: $\underline{x}$ en $\underline{y}$ zijn alleen dan o.o. als

$$P(\underline{x} \in A,\ \underline{y} \in B) = P(\underline{x} \in A)P(\underline{y} \in B) \qquad \text{(alle A en B)}$$

of als

$$F_{\underline{x},\underline{y}}(x,y) = F_{\underline{x}}(x).F_{\underline{y}}(y) \qquad \text{(alle x en y)}.$$

Analoog heten $\underline{x}_1,\dots,\underline{x}_n$ o.o. als resp.

$$P(\underline{x}_1 = x_1, \ldots, \underline{x}_n = x_n) = P(\underline{x}_1 = x_1) \ldots P(\underline{x}_n = x_n) \qquad (\text{alle } (x_1, \ldots, x_n))$$

$$f_{\underline{x}_1, \ldots, \underline{x}_n}(x_1, \ldots, x_n) = f_{\underline{x}_1}(x_1) \ldots f_{\underline{x}_n}(x_n) \qquad (\text{alle } (x_1, \ldots, x_n))$$

Voorbeelden.

1. Homogene verdeling op een cirkelschijf

$$f(x,y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \le 1 \\ 0 & \text{anders} \end{cases}$$

2. Homogene verdeling op een vierkant

$$f(x,y) = \begin{cases} 1 & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{anders} \end{cases}$$

3. Normale verdeling ($\underline{x}$ en $\underline{y}$ o.o.)

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2 + y^2)/(2\sigma^2)}$$

4. Normale verdeling ($\underline{x}$ en $\underline{y}$ afhankelijk)

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}(x^2 - 2\rho xy + y^2)/(1-\rho^2)} \qquad (0 < |\rho| < 1)$$

5. Exponentiële verdeling ($\underline{x}$ en $\underline{y}$ o.o.)

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & 0 < x < \infty,\ 0 < y < \infty \\ 0 & \text{anders} \end{cases}$$

Opgave 1: ga na dat de marginale kansdichtheid van $\underline{x}$ (en dus van $\underline{y}$) in de voorbeelden 3 en 4 normaal zijn met parameters (zie Voorbeeld 8, blz. 17) $\mu = 0$ en $\sigma = 1$ resp. $\mu = 0$ en $\sigma = 1$.

Opgave 2: ga na dat de stochastische grootheden met dichtheden zoals in bovenstaande voorbeelden *afhankelijk* zijn in de gevallen 1 en 4 en *onafhankelijk* in de gevallen 2,3 en 5. Zie Definitie 8.

2.4. Functies van stochastische grootheden

In veel gevallen zijn we geïnteresseerd in *functies* van de waargenomen grootheden.

Voorbeeld 1: de diameter $\underline{x}$ van een bepaald type stalen kogels heeft de volgende kansdichtheid (homogene verdeling op (1,2))

$$f_{\underline{x}}(x) = \begin{cases} 1 & 1 < x < 2 \\ 0 & \text{anders} \end{cases}$$

Wat is de kansdichtheid van het *volume* $\underline{v}$ van de kogels?

Antwoord: we weten dat $\underline{v} = \frac{1}{6}\pi\,\underline{x}^3$. Om $f_{\underline{v}}$ te vinden berekenen we eerst $F_{\underline{v}}$:

$$F_{\underline{v}}(v) = P(\underline{v} \le v) = P(\tfrac{1}{6}\pi\,\underline{x}^3 \le v) = P(\underline{x} \le (\tfrac{6v}{\pi})^{1/3}) = F_{\underline{x}}((\tfrac{6v}{\pi})^{1/3}) =$$

$$= \frac{(\frac{6v}{\pi})^{1/3} - 1}{2 - 1} \qquad (\tfrac{\pi}{6} < v < \tfrac{4\pi}{3}).$$

Omdat $f_{\underline{v}}(v) = F'_{\underline{v}}(v)$ vinden we

$$f_{\underline{v}}(v) = \frac{1}{3}\,(\frac{6}{\pi})^{1/3}\, v^{-2/3} \qquad (\tfrac{\pi}{6} < v < \tfrac{4\pi}{3}).$$

Het is duidelijk dat $\frac{\pi}{6} < \underline{v} < \frac{4\pi}{3}$ en het is nuttig dit te constateren alvorens met het rekenwerk te beginnen. Voor waarden van v buiten $(\frac{\pi}{6}, \frac{4\pi}{3})$ is $f_{\underline{v}}(v)$ natuurlijk nul.

Dikwijls zijn we geïnteresseerd in functies van meer dan één stochastische grootheid.

Voorbeeld 2: twee gloeilampen hebben een brandduur (in maanden) $\underline{x}$ resp. $\underline{y}$ waarbij $\underline{x}$ en $\underline{y}$ o.o. zijn en exponentieel verdeeld:

$$P(\underline{x} \le x) = P(\underline{y} \le x) = 1 - e^{-\lambda x} \qquad (x \ge 0).$$

Wat is de kans dat de totale branduur van beide lampen samen (de tweede lamp wordt ingedraaid als de eerste uitvalt) minder dan 3 maanden bedraagt ?

Antwoord: de gevraagde kans is (zie formule(7) op blz. 25)

$$P(\underline{x}+\underline{y} < 3) = \iint\limits_{x+y\le 3} \lambda^2 e^{-\lambda x} e^{-\lambda y} dxdy = \int_0^3 \lambda e^{-\lambda x} \left\{ \int_0^{3-x} \lambda e^{-\lambda y} dy \right\} dx =$$

$$= \int_0^3 \lambda e^{-\lambda x}(1 - e^{-\lambda(3-x)})dx = \int_0^3 (\lambda e^{-\lambda x} - \lambda e^{-3\lambda})dx$$

$$= 1 - e^{-3\lambda} - 3\lambda e^{-3\lambda}. \qquad \text{(zie voorbeeld 6, blz. 17)}$$

Voor $\lambda = 1$ vinden we bijvoorbeeld $P(\underline{x}+\underline{y} < 3) = 1 - 4e^{-3} = 0{,}801$.

Opgave 1: als $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ o.o. zijn en alle dezelfde verdelingsfunctie F hebben, dan geldt voor $\underline{M} := \max(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$ resp. $\underline{m} := \min(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$

$$F_{\underline{M}}(x) = \{F(x)\}^n$$ $f_M = n\{F(x)\}^{n-1} \cdot f(x)$ parallel.

$$F_{\underline{m}}(x) = 1 - \{1 - F(x)\}^n.$$ $F_m = n\{1 - F(x)\}^{n-1} \cdot f(x)$ serie

Waarom ? Bereken ook $f_{\underline{M}}$ en $f_{\underline{m}}$.

Opgave 2: Bereken $P(\underline{x}+\underline{y} > 0)$ als $(\underline{x},\underline{y})$ homogeen verdeeld is op een cirkelschijf (zie Voorbeelden , 1, blz.28).

Opgave 3: Bereken $F_{\underline{x}+\underline{y}}$ en $f_{\underline{x}+\underline{y}}$, als $(\underline{x},\underline{y})$ homogeen verdeeld is over het eenheidsvierkant (Voorbeelden, 2).

Opgave 4: $\underline{x}$ en $\underline{y}$ zijn $N(0,1)$ en o.o. (zie blz. 17). Bereken $f_{\underline{x}+\underline{y}}$.

3. Verwachting

De kansverdeling (gegeven door $P(\underline{x} = x_i)$ voor alle i of door $f_{\underline{x}}(x)$ voor alle x) geeft een volledige beschrijving van het gedrag van $\underline{x}$. Dikwijls neemt men echter genoegen met een onvolledige beschrijving d.m.v. één of twee karakteristieke getallen, zoals men in de mechanica soms alleen de plaats van het zwaartepunt van een lichaam opgeeft of van een inkomensverdeling alleen het gemiddelde inkomen. Wij gaan uit van het begrip gemiddelde: als N getallen $x_1, x_2, \ldots, x_N$ gegeven zijn, dan is het gemiddelde $\bar{x}$ van die getallen

$$\bar{x} := \frac{1}{N} \sum_{1}^{N} x_j \ . \tag{1}$$

Komen de getallen in groepen voor: k_1 maal x_1, k_2 maal $x_2, \ldots, k_n$ maal x_n, met $k_1 + k_2 + \ldots + k_n = N$, dan wordt het gemiddelde

$$\bar{x} = \frac{1}{N} \sum_{1}^{n} x_j k_j \ , \tag{1'}$$

of als we de fracties k_j/N aangeven met f_j, zodat $\sum_{1}^{n} f_j = 1$,

$$\bar{x} = \sum_{1}^{n} x_j f_j \ . \tag{1''}$$

Analoog (hier treedt weer de analogie kans-fractie op) geven we voor discrete stochastische grootheden

Definitie 1: als $\underline{x}$ een discrete stochastische grootheid is, met $P(\underline{x} = x_j) = p_j$ $(j = 1,2,\ldots)$ dan definiëren we $E\,\underline{x}$, de *verwachting* van $\underline{x}$ door

$$E\,\underline{x} = \sum_{j=1}^{\infty} x_j p_j \ . \tag{2}$$

Voor continue stochastische grootheden hebben we analoog

Definitie 2: als $\underline{x}$ een continue stochastische grootheid is met kansdichtheid $f_{\underline{x}}$, dan definiëren we $E\,\underline{x}$, de verwachting van $\underline{x}$, door

$$(3) \qquad E\,\underline{x} = \int_{-\infty}^{\infty} x\, f_{\underline{x}}(x)dx \,.$$

Voor een functie van een stochastische grootheid $\underline{x}$ definiëren we

Definitie 3:

$$(4) \qquad E\,h(\underline{x}) = \begin{cases} \sum_{j=1}^{\infty} h(x_j)P(\underline{x} = x_j) & \text{(discrete geval)} \\ \int_{-\infty}^{\infty} h(x)f_{\underline{x}}(x)dx & \text{(continue geval)} \,. \end{cases}$$

Opmerking: Eigenlijk was $E\,h(\underline{x})$ al gedefinieerd: als we schrijven $h(\underline{x}) = \underline{y}$, dan kunnen we in het continue geval $f_{\underline{y}}$ uitrekenen en in het discrete geval $P(\underline{y} = y_j)$ en vervolgens $E\,\underline{y}$ berekenen volgens (3) of (2). Omdat deze methode altijd hetzelfde oplevert als (4) is Definitie 3 toegestaan. Een soortgelijke opmerking is van toepassing op de volgende definitie.

Definitie 4: als $\underline{x}$ en $\underline{y}$ stochastische grootheden zijn, dan is

$$(5) \qquad E\,h(\underline{x},\underline{y}) = \begin{cases} \sum_{i} \sum_{j} h(x_i,y_j)P(\underline{x} = x_i,\ \underline{y} = y_j) & \text{(discrete geval)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y)\ f_{\underline{x},\underline{y}}(x,y)dxdy & \text{(continue geval)} \,. \end{cases}$$

Opgave 1: laat zien dat in Voorbeeld 1 (blz. 29) geldt

$$E\,\underline{v} = \int_{\pi/6}^{4\pi/3} v f_{\underline{v}}(v)dv = \frac{\pi}{6} \int_{1}^{2} x^3 f_{\underline{x}}(x)dx \,.$$

Opgave 2: zij in Voorbeeld 2 (blz. 29) $\underline{z} = \underline{x} + \underline{y}$.

Bereken $f_{\underline{z}}(z)$ en laat zien dat $E\,\underline{z} = \int_0^\infty z f_{\underline{z}}(z)dz = E\,\underline{x} + E\,\underline{y}$.

Veel gebruikte speciale gevallen zijn de volgende.

Definitie 5:

$$\mu_{\underline{x}}^{(\ell)} := E\,\underline{x}^{\ell}$$

heet het ℓ-de *moment* van $\underline{x}$. In plaats van $\mu_{\underline{x}}^{(1)}$ schrijven we dikwijls $\mu_{\underline{x}}$ of alleen μ.

$$\text{var}\ \underline{x} := E\,(\underline{x} - \mu_{\underline{x}})^2$$

heet de *variantie* van $\underline{x}$. We schrijven ook wel $\sigma_{\underline{x}}^2$, $\sigma^2(\underline{x})$ of alleen σ^2. $\sigma_{\underline{x}}$ heet *standaardafwijking* of *standaarddeviatie* (soms ook "spreiding").

$$\text{cov}(\underline{x},\underline{y}) := E\,(\underline{x} - \mu_{\underline{x}})(\underline{y} - \mu_{\underline{y}})$$

heet de *covariantie* van $(\underline{x},\underline{y})$.

$$\rho(\underline{x},\underline{y}) := \frac{\text{cov}(\underline{x},\underline{y})}{\sigma_{\underline{x}}\,\sigma_{\underline{y}}}$$

heet de *correlatiecoefficient* van $(\underline{x},\underline{y})$.

Stelling 1: (i) $E(g(\underline{x}) + h(\underline{x})) = Eg(\underline{x}) + Eh(\underline{x})$ en $E(a\underline{x} + b) = aE\underline{x} + b$; speciaal $E\,a\underline{x} = a\,E\,\underline{x}$ en $E\,b = b$.

(ii) $\text{var}(a\underline{x} + b) = a^2\ \text{var}\ \underline{x}$.

(iii) $\text{var}\ \underline{x} = E\,\underline{x}^2 - (E\underline{x})^2$.

Bewijs: (i) volgt direct uit (4) en (2) of (3).

(ii) $$\text{var}(a\underline{x} + b) = E\,\{a\underline{x} + b - E\,(a\underline{x} + b)\}^2 = E\,(a\underline{x} - a\,E\,\underline{x})^2$$
$$= E\,a^2(\underline{x} - E\,\underline{x})^2 = a^2\,E\,(\underline{x} - E\,\underline{x})^2 = a^2\ \text{var}\ \underline{x}.$$

(iii) $$E\,(\underline{x} - \mu)^2 = E\,(\underline{x}^2 - 2\mu\underline{x} + \mu^2) = E\,\underline{x}^2 - 2\mu\,E\,\underline{x} + \mu^2$$
$$= E\,\underline{x}^2 - \mu^2 = E\,\underline{x}^2 - (E\,\underline{x})^2.$$

μ

Gevolg: $\sigma(a\underline{x}) = |a|\,\sigma(\underline{x})$ en dus speciaal $\sigma(-\underline{x}) = \sigma(\underline{x})$.

Opmerking 1: de verwachting $\mu_{\underline{x}}$ geeft de plaats aan van het "zwaartepunt" van de kansverdeling van $\underline{x}$. De variantie $\sigma^2_{\underline{x}}$ is de verwachting van de kwadratische afwijking van dit zwaartepunt: een kleine variantie wijst op een grote concentratie van de kansverdeling in de omgeving van $\mu_{\underline{x}}$ (als $\sigma^2_{\underline{x}} = 0$, dan is zelfs $\underline{x}$ altijd gelijk aan een constante: $P(\underline{x} = c) = 1$ met natuurlijk $\mu_{\underline{x}} = c$), een grote variantie wijst op een verspreid liggen van de "kansmassa".

Opmerking 2: men gebruikt soms andere constanten dan $\mu_{\underline{x}}$ om de globale ligging van de kansverdeling aan te geven: 1e. de "*modus*", d.i. de plaats waar $f_{\underline{x}}(x)$ zijn top heeft (als er precies één top is), 2e. de "*mediaan*", $m_{\underline{x}}$, waarvoor geldt

$$P(\underline{x} \le m_{\underline{x}}) = P(\underline{x} \ge m_{\underline{x}}) = \tfrac{1}{2}\,.$$

Men spreekt dan wel over "het mediane inkomen", d.i. het inkomen waar 50% van de mensen "onder zit" (of: "boven zit").

Voorbeelden (zie blz. 16)

1. Diskrete homogene verdeling:

   a) $E\underline{u} = \bar{x} := \frac{1}{n}\sum_{1}^{n} x_k$; $\operatorname{var}\underline{u} = \frac{1}{n}\sum_{1}^{n}(x_k - \bar{x})^2 = \frac{1}{n}\sum_{1}^{n} x_k^2 - \bar{x}^2$.

   b) speciaal als $x_k = k$ $(k = 1,2,\dots,n)$; $E\underline{u} = \frac{n+1}{2}$, $\operatorname{var}\underline{u} = \frac{n^2-1}{12}$.

2. Binomiale verdeling : $E\,\underline{u} = np$, $\operatorname{var}\underline{u} = np(1-p)$

3. Poisson verdeling : $E\,\underline{u} = \mu$, $\operatorname{var}\underline{u} = \mu$

4. Geometrische verdeling : $E\,\underline{u} = \frac{1}{p}$, $\operatorname{var}\underline{u} = \frac{1-p}{p^2}$

5. Homogene verdeling op (a,b): $E\,\underline{x} = \frac{a+b}{2}$, $\operatorname{var}\underline{x} = \frac{(b-a)^2}{12}$ (verg. 1)

6. Gamma-verdeling : $E\,\underline{x} = \frac{n}{\lambda}$, $\operatorname{var}\underline{x} = \frac{n}{\lambda^2}$

7. <u>Exponentiële verdeling</u> : $E\underline{x} = \frac{1}{\lambda}$, $\text{var } \underline{x} = \frac{1}{\lambda^2}$

8. <u>Normale verdeling</u> : $E\underline{x} = \mu$, $\text{var } \underline{x} = \sigma^2$.

<u>Opgave 1</u>: als $\underline{x}$ normaal verdeeld is met $E\underline{x} = \mu$ en $\text{var } \underline{x} = \sigma^2$, dan is $\underline{y} := \frac{\underline{x} - \mu}{\sigma}$ normaal verdeeld met $E\underline{y} = 0$ en $\text{var } \underline{y} = 1$.

<u>Opgave 2</u>: schets de grafiek van de normale kansdichtheid

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

voor de waarden $(1,1)$, $(2,1)$, $(1,\frac{1}{4})$ en $(1,4)$ van (μ,σ^2).

De betekenis van de variantie komt ook tot uiting in de volgende (beroemde) ongelijkheid.

<u>Stelling 2</u> (ongelijkheid van Chebyshev):

$$P(|\underline{x} - \mu| \geq c) \leq \frac{\text{var } \underline{x}}{c^2} \qquad (c > 0) \; . \tag{6}$$

<u>Bewijs</u>: in het continue geval geldt

$$\text{var } \underline{x} = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \geq \int_{|x-\mu|\geq c} (x-\mu)^2 f(x)dx \geq c^2 \int_{|x-\mu|\geq c} f(x)dx =$$

$$= c^2 P(|\underline{x} - \mu| \geq c) \qquad \text{(zie formule (9), blz. 13)} \; ,$$

waaruit de ongelijkheid direct volgt. In het discrete geval gaat het bewijs geheel analoog.

<u>Opgave 3</u>: Verifieer de uitkomsten in de voorbeelden 1 t/m 8 en controleer de ongelijkheid (6) in deze gevallen.

Voor functies van stochastische grootheden hebben we

Stelling 3:

1. $E(\underline{x}+\underline{y}) = E\underline{x} + E\underline{y}$

1a. $E\sum_1^n \underline{x}_j = \sum_1^n E\underline{x}_j$

2. $\mathrm{cov}(\underline{x},\underline{y}) = E\underline{x}\,\underline{y} - E\underline{x}\; E\underline{y}$

3. $\mathrm{var}(\underline{x}\pm\underline{y}) = \mathrm{var}\,\underline{x} + \mathrm{var}\,\underline{y} \pm 2\mathrm{cov}(\underline{x},\underline{y})$

3a. $\mathrm{var}(\sum_1^n \underline{x}_j) = \sum_1^n \mathrm{var}\,\underline{x}_j + 2\sum_{i<j} \mathrm{cov}(\underline{x}_i,\underline{x}_j)$.

Bewijs: deze eigenschappen volgen vrijwel direct uit de Definities 4 en 5, bijv.

$$\mathrm{var}(\underline{x}+\underline{y}) = E\{\underline{x}+\underline{y} - E(\underline{x}+\underline{y})\}^2 = E\{(\underline{x}-E\underline{x}) + (\underline{y}-E\underline{y})\}^2 =$$

$$= E(\underline{x}-E\underline{x})^2 + E(\underline{y}-E\underline{y})^2 + 2E(\underline{x}-E\underline{x})(\underline{y}-E\underline{y}) = \mathrm{var}\,\underline{x} + \mathrm{var}\,\underline{y} + 2\mathrm{cov}(\underline{x},\underline{y}).$$

Voor *onafhankelijke stochastische grootheden* geldt bovendien

Stelling 4: als $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n$ *onderling onafhankelijk* zijn, dan geldt

1. $E\underline{x}_1\,\underline{x}_2 = E\underline{x}_1\; E\underline{x}_2$

1a. $E\prod_1^n \underline{x}_j = \prod_1^n E\underline{x}_j$

2. $\mathrm{cov}(\underline{x}_1,\underline{x}_2) = \rho(\underline{x}_1,\underline{x}_2) = 0$

3. $\mathrm{var}(\underline{x}_1+\underline{x}_2) = \mathrm{var}\,\underline{x}_1 + \mathrm{var}\,\underline{x}_2$

3a. $\mathrm{var}\sum_1^n \underline{x}_j = \sum_1^n \mathrm{var}\,\underline{x}_j$

3b. $\mathrm{var}\sum_1^n \alpha_j\underline{x}_j = \sum_1^n \alpha_j^2\, \mathrm{var}\,\underline{x}_j$.

Bewijs: dit volgt direct uit de definitie van o.o. stochastische grootheden (zie Definitie 8, blz. 27 en analogon op blz. 28), bijvoorbeeld

$$E\,\underline{x}_1\underline{x}_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2\, f_{\underline{x}_1,\underline{x}_2}(x_1,x_2)dx_1dx_2$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2\, f_{\underline{x}_1}(x_1)f_{\underline{x}_2}(x_2)dx_1dx_2$$

$$= \int_{-\infty}^{\infty} x_1\, f_{\underline{x}_1}(x_1)dx_1 \int_{-\infty}^{\infty} x_2\, f_{\underline{x}_2}(x_2)dx_2 = E\,\underline{x}_1\;E\,\underline{x}_2\;.$$

Zie verder Stelling 3 en Stelling 1.

Opgave: laat zien dat uit de voorgaande stellingen volgt dat voor het gemiddelde, $\bar{\underline{x}} := \frac{1}{n}\sum_1^n \underline{x}_j$, van n onafhankelijke stochastische grootheden $\underline{x}_1,\ldots,\underline{x}_n$ met $E\,\underline{x}_j = \mu$ en var $\underline{x}_j = \sigma^2$ geldt

(7)
$$E\,\bar{\underline{x}} = \mu$$
$$\mathrm{var}\;\bar{\underline{x}} = \sigma^2/n\;.$$

Voorbeeld (samengestelde verdeling): laat $\underline{n}$, $\underline{x}_1,\underline{x}_2,\ldots$ onderling onafhankelijke stochastische grootheden zijn, waarvan $\underline{n}$ alleen de waarden 0,1,2,... kan aannemen, terwijl de $\underline{x}_j$ alle dezelfde discrete verdeling als een gegeven grootheid $\underline{x}$ hebben. We definiëren

$$\underline{z} = \underline{x}_1 + \underline{x}_2 + \ldots + \underline{x}_{\underline{n}}$$

(met $\underline{x}_1 + \ldots + \underline{x}_{\underline{n}} = 0$ als $\underline{n} = 0$) en vragen naar $E\,\underline{z}$ en var $\underline{z}$.

Antwoord: ook $\underline{z}$ heeft een discrete verdeling en volgens Definitie 1 (of Definitie 3) geldt, als $z_1,z_2,\ldots$ de waarden zijn die $\underline{z}$ kan aannemen,

$$E\,\underline{z} = \sum_j z_j\; P(\underline{z} = z_j),$$

Nu is (zie (i"), blz. 27)

$$P(\underline{z} = z_j) = \sum_{n=0}^{\infty} P(\underline{n} = n,\ \underline{z} = z_j) = \sum_{n=0}^{\infty} P(\underline{n} = n,\ \underline{x}_1 + \ldots + \underline{x}_n = z_j)$$

$$= \sum_{n=0}^{\infty} P(\underline{n} = n,\ \underline{x}_1 + \ldots + \underline{x}_n = z_j)\ .$$

Omdat $\underline{n}$, $\underline{x}_1, \underline{x}_2, \ldots$ o.o. zijn, zijn ook $\underline{n}$ en $\underline{x}_1 + \ldots + \underline{x}_n$ o.o., en dus geldt

$$P(\underline{n} = n, \underline{x}_1 + \ldots + \underline{x}_n = z_j) = P(\underline{n} = n)P(\underline{x}_1 + \ldots + \underline{x}_n = z_j)\ ,$$

zodat tenslotte (zie Stelling 3)

$$E\,\underline{z} = \sum_j \sum_n z_j\ P(\underline{n} = n)P(\underline{x}_1 + \ldots + \underline{x}_n = z_j) =$$

$$= \sum_{n=0}^{\infty} P(\underline{n} = n)\ E\,(\underline{x}_1 + \ldots + \underline{x}_n) = \sum_{n=0}^{\infty} P(\underline{n} = n)n\,E\,\underline{x} = E\,\underline{n}\,E\,\underline{x}\ .$$

Analoog vinden we (vergelijk iii, blz. 33, Stelling 3, 1a en Stelling 4, 3a)

$$E\,\underline{z}^2 = \sum_n \sum_j P(\underline{n} = n)P(\underline{x}_1 + \ldots + \underline{x}_n = z_j)z_j^2 =$$

$$= \sum_{n=0}^{\infty} P(\underline{n} = n)\ E\,(\underline{x}_1 + \ldots + \underline{x}_n)^2 =$$

$$= \sum_{n=0}^{\infty} P(\underline{n} = n)\ [\mathrm{var}(\underline{x}_1 + \ldots + \underline{x}_n) + \{E\,(\underline{x}_1 + \ldots + \underline{x}_n)\}^2] =$$

$$= \sum_{n=0}^{\infty} P(\underline{n} = n)\ [n\ \mathrm{var}\ \underline{x} + n^2(E\,\underline{x})^2] = E\,\underline{n}\ \mathrm{var}\ \underline{x} + E\,\underline{n}^2(E\,\underline{x})^2\ ,$$

zodat (met nogmaals (iii), blz. 33 en met $E\,\underline{z} = E\,\underline{n}\ E\,\underline{x}$)

$$\mathrm{var}\ \underline{z} = E\,\underline{n}\ \mathrm{var}\ \underline{x} + \mathrm{var}\ \underline{n}(E\,\underline{x})^2\ .$$

Voor continu-verdeelde $\underline{x}$ gelden *dezelfde* formules. De bewijzen zijn analoog maar passen niet helemaal in het kader van dit college.

Opmerking: covariantie en correlatiecoëfficiënt hebben te maken met de mate van afhankelijkheid van twee stochastische grootheden: als $\mathrm{cov}(\underline{x},\underline{y})$ groot is (d.w.z. $\rho(\underline{x},\underline{y}) \approx 1$) dan gaan grote waarden van $\underline{x}$ gepaard met grote waarden van $\underline{y}$, en kleine waarden van $\underline{x}$ met kleine waarden van $\underline{y}$. Als $\mathrm{cov}(\underline{x},\underline{y})$ klein is (d.w.z. $\rho(\underline{x},\underline{y}) \approx -1$), dan gaan grote (kleine) waarden van $\underline{x}$ gepaard met kleine (grote) waarden van $\underline{y}$. Als $\underline{x}$ en $\underline{y}$ o.o. zijn dan is $\rho(\underline{x},\underline{y}) = 0$. Het omgekeerde is waar als $(\underline{x},\underline{y})$ normaal verdeeld is, maar geldt niet algemeen (zie Opgave 3). In Voorbeeld 4 op blz. 28 stelt ρ de correlatiecoëfficiënt van $\underline{x}$ en $\underline{y}$ voor; als $\rho = 0$ is zijn $\underline{x}$ en $\underline{y}$ inderdaad o.o. (waarom?). De lijnen met $f_{\underline{x},\underline{y}}(x,y)$ = constant zijn ellipsen, die langwerpiger worden naar mate $|\rho|$ dichter bij 1 komt. De lange as van deze ellipsen heeft richtingscoëfficiënt 1 als $\rho > 0$ en richtingscoëfficient -1 als $\rho < 0$. We vermelden nog dat steeds geldt dat $|\rho(\underline{x},\underline{y})| \leq 1$.

Opgave 1:

$$f_{\underline{x},\underline{y}}(x,y) = \begin{cases} a^2 e^{-ax-ay} & x > 0,\ y > 0 \\ 0 & \text{anders}. \end{cases}$$

Bereken $E\,\underline{x}$, var $\underline{x}$, $E\,\underline{x}\,\underline{y}$ en $\rho(\underline{x},\underline{y})$.

Opgave 2:

$$f_{\underline{x},\underline{y}}(x,y) = \begin{cases} \tfrac{1}{2}(a^2 e^{-ax-ay} + b^2 e^{-bx-by}) & x \geq 0,\ y \geq 0 \\ 0 & \text{anders}. \end{cases}$$

Bereken $\rho(\underline{x},\underline{y})$.

Opgave 3: bereken $\rho(\underline{x},\underline{y})$ als $(\underline{x},\underline{y})$ homogeen verdeeld is op de eenheidscirkelschijf (zie Voorbeeld 1, blz. 28).

4. Sommen van o.o. stochastische grootheden, limietstellingen

4.1. Sommen van o.o. stochastische grootheden

Sommen van stochastische grootheden komen in veel toepassingen voor, hetzij omdat de sommen zelf van belang zijn, hetzij omdat men geïnteresseerd is in het gemiddelde van de waargenomen grootheden.

Voorbeeld 1: men doet n metingen van een natuurconstante c. Door meetfouten en "toevallige" omstandigheden kan men de meetresultaten beschouwen als o.o. stochastische grootheden, $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$, alle met dezelfde verdeling en met $E\,\underline{x}_j = c$, zodat ook (zie (6), blz. 37) $E\,\bar{\underline{x}} = c$.

Voorbeeld 2: bij n onafhankelijke experimenten elk met kans p op succes telt men het totaal aantal successen $\underline{k}$. We kunnen $\underline{k}$ schrijven als

$$\underline{k} = \underline{x}_1 + \ldots + \underline{x}_n ,$$

waarbij $\underline{x}_1, \ldots, \underline{x}_n$ o.o. zijn en waarbij $\underline{x}_j$ het aantal successen is bij het j-de experiment: $P(\underline{x}_j = 1) = p$ en $P(\underline{x}_j = 0) = 1 - p$. We weten al (zie blz.24) dat $\underline{k}$ een binomiale verdeling heeft met $E\,\underline{k} = np$ en var $\underline{k} = np(1-p)$ (zie ook Stellingen 3 en 4, blz. 36).

In beide voorbeelden zijn we geinteresseerd in het gedrag voor grote waarden van n; in Voorbeeld 1, omdat we hopen dat meer waarnemingen tot een nauwkeuriger resultaat zullen leiden en in Voorbeeld 2, omdat het ondoenlijk is om de binomiale verdeling voor grote n te tabelleren, zodat we geïnteresseerd zijn in een *benadering*.

Voordat we $\underline{x}_1 + \ldots + \underline{x}_n$ voor grote n beschouwen, kijken we eerst naar wat eenvoudige gevallen. In het discrete geval hebben we

Stelling 1: als $\underline{k}$ en $\underline{\ell}$ onafhankelijke stochastische grootheden zijn, die alleen de waarden 0,1,2,... kunnen aannemen, met $P(\underline{k} = k) = p_k$ $(k = 0,1,\ldots)$ en $P(\underline{\ell} = \ell) = q_\ell$ $(\ell = 0,1,\ldots)$, dan geldt

$$(1) \qquad P(\underline{k} + \underline{\ell} = n) = \sum_{k=0}^{n} p_k q_{n-k} = \sum_{\ell=0}^{n} q_\ell p_{n-\ell} .$$

Bewijs:

$P(\underline{k} + \underline{\ell} = n) = P(\underline{k} = 0,\ \underline{\ell} = n$ òf $\underline{k} = 1,\ \underline{\ell} = n-1$ òf ... òf $\underline{k} = n,\ \underline{\ell} = 0) =$

$= P(\underline{k} = 0,\ \underline{\ell} = n) + P(\underline{k} = 1,\ \underline{\ell} = n-1) + \dots + P(\underline{k} = n,\ \underline{\ell} = 0) =$

$$= \sum_{k=0}^{n} p_k q_{n-k} = \sum_{\ell=0}^{n} q_\ell p_{n-\ell} \ .$$

Hierbij gebruiken we de onafhankelijkheid van $\underline{k}$ en $\underline{\ell}$: $P(\underline{k} = k,\ \underline{\ell} = n-k) = p_k q_{n-k}$. De laatste overgang ontstaat door substitutie van $n-k = \ell$.

Opgave: a. als $\underline{k}_1$ en $\underline{k}_2$ o.o. zijn en binomiaal verdeeld met parameters (n_1,p) resp. (n_2,p), dan is $\underline{k}_1 + \underline{k}_2$ ook binomiaal verdeeld met parameters $(n_1 + n_2,p)$. Gebruik Stelling 1 (het kan ook met behulp van de interpretatie in Voorbeeld 2, blz. 40).

b. Als $\underline{k}_1$ en $\underline{k}_2$ o.o. zijn en Poisson-verdeeld met verwachting resp. μ_1 en μ_2 dan is $\underline{k}_1 + \underline{k}_2$ ook Poisson-verdeeld met verwachting (uiteraard) gelijk aan $\mu_1 + \mu_2$.

Voor continue stochastische grootheden geldt analoog

Stelling 2: als $\underline{x}$ en $\underline{y}$ o.o. zijn met kansdichtheid resp. $f_{\underline{x}}$ en $f_{\underline{y}}$, dan geldt

$$f_{\underline{x}+\underline{y}}(z) = \int_{-\infty}^{\infty} f_{\underline{x}}(z-y) f_{\underline{y}}(y)\,dy = \int_{-\infty}^{\infty} f_{\underline{x}}(x) f_{\underline{y}}(z-x)\,dx \ . \tag{2}$$

Bewijs: we beginnen weer met de *verdelingsfunctie*. Zij $\underline{z} = \underline{x} + \underline{y}$, dan is (zie (7), blz. 25 en Definitie 8.b, blz. 27).

$$F_{\underline{z}}(z) = P(\underline{x} + \underline{y} \le z) = \iint\limits_{x+y\le z} f_{\underline{x}}(x) f_{\underline{y}}(y)\,dx\,dy = \int_{-\infty}^{\infty} f_{\underline{x}}(x) \left\{ \int_{-\infty}^{z-x} f_{\underline{y}}(y)\,dy \right\} dx =$$

$$= \int_{-\infty}^{\infty} f_{\underline{x}}(x) F_{\underline{y}}(z-x)\,dx \ .$$

Differentiatie naar z (onder het integraalteken) levert de laatste integraal in (2). Integratie in de andere volgorde geeft de eerste integraal.

Opmerking: als $\underline{x}$ en $\underline{y}$ beide positief zijn (zodat $f_{\underline{x}}(x)$ en $f_{\underline{y}}(x)$ nul zijn voor $x < 0$) dan gaat (2) over in

$$(2') \qquad f_{\underline{x}+\underline{y}}(z) = \int_0^z f_{\underline{x}}(z-y)f_{\underline{y}}(y)dy = \int_0^z f_{\underline{x}}(x)f_{\underline{y}}(z-x)dx \ ;$$

de integralen in (2) en (2') worden *convolutie-integralen* genoemd.

Opgave: $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n$ zijn o.o. en exponentieel verdeeld met $E\,\underline{x}_j = \frac{1}{\lambda}$ (zie blz. 17). Bewijs met inductie dat $\underline{x}_1 + \ldots + \underline{x}_n$ een Gamma-verdeling heeft (blz. 17), d.w.z.

$$f_{\underline{x}_1+\ldots+\underline{x}_n}(z) = \frac{\lambda^n}{(n-1)!} z^{n-1}e^{-\lambda z} \qquad (z > 0) \ .$$

4.2. Limietstellingen

We beschouwen nu $\underline{x}_1 + \ldots + \underline{x}_n$ voor grote n. Voor het gemiddelde (zie Voorbeeld 1, blz. 40) geldt

Stelling 3 (wet van de grote aantallen): als $\underline{x}_1,\underline{x}_2,\ldots$ onderling onafhankelijk zijn met $E\,\underline{x}_j = \mu$ en var $\underline{x}_j = \sigma^2$ $(j = 1,2,\ldots)$ en als $\bar{\underline{x}}_n = \frac{1}{n}\sum_1^n \underline{x}_j$, dan geldt voor iedere $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\bar{\underline{x}}_n - \mu| < \varepsilon) = 1 \ ,$$

d.w.z. het gemiddelde van $\underline{x}_1,\ldots,\underline{x}_n$ ligt voor grote n met grote waarschijnlijkheid vlak bij de verwachting.

Bewijs: uit de ongelijkheid van Chebyshev (blz. 35) en formule (6) (blz. 37) voor $E\,\bar{\underline{x}}$ en var $\bar{\underline{x}}$ volgt direct

$$P(|\bar{\underline{x}}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \ ,$$

dus (zie Stelling 1, blz. 6)

$$P(|\bar{\underline{x}}_n - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \to 1$$

voor $n \to \infty$ en iedere vaste $\varepsilon > 0$.

Opmerking: de betekenis van de wet van de grote aantallen is ook de volgende: als men een experiment waarbij een gebeurtenis A kan optreden n maal onafhankelijk uitvoert, dan is volgens deze wet de *fractie* van het aantal keren dat A optreedt voor grote n een goede benadering voor P(A). Immers als $p = P(A)$ en $\underline{k}$ het aantal malen dat A optreedt, dan is (verg. Voorbeeld 2, blz. 40) $\underline{k} = \underline{x}_1 + \ldots + \underline{x}_n$, met $P(\underline{x}_j = 1) = p$ en $P(\underline{x}_j = 0) = 1 - p$ en $E\,\underline{x}_j = p$. Stelling 3 zegt nu dat $P(|\underline{k}/n - p| < \varepsilon) \to 1$ voor $n \to \infty$, of ruwweg: $\underline{k}/n \approx p$. We kunnen dus P(A) goed schatten door $\underline{k}/n$.

Om de kansverdeling van $\underline{x}_1 + \ldots + \underline{x}_n$ te benaderen hebben we de zgn. "*Centrale limietstelling*" (afgekort C.L.S.).

Stelling 4 (C.L.S.): als $\underline{x}_1, \underline{x}_2, \ldots$ o.o. stochastische grootheden zijn met dezelfde kansverdeling en met $E\,\underline{x}_j = \mu$, var $\underline{x}_j = \sigma^2$, dan geldt

$$\lim_{n\to\infty} P\left(\frac{\underline{x}_1 + \ldots + \underline{x}_n - n\mu}{\sigma\sqrt{n}} \le z\right) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du\ .$$

Kansverdelingen van sommen van o.o. stochastische grootheden kunnen zeer verscheiden en zeer ingewikkeld zijn. De betekenis van de C.L.S. is dat we al deze kansverdelingen met één enkele kansverdeling, de normale, kunnen benaderen.

Voorbeeld 1: als $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ o.o. zijn en exponentieel verdeeld met $E\,\underline{x}_j = \frac{1}{\lambda}$, dan heeft $\underline{z} = \underline{x}_1 + \ldots + \underline{x}_n$ de kansdichtheid (zie blz. 17)

$$(3) \qquad f_{\underline{z}}(z) = \frac{\lambda^n}{(n-1)!} z^{n-1} e^{-\lambda z} \qquad (z > 0)\ .$$

De C.L.S. zegt nu dat voor grote n (zie ook blz. 34)

$$F_{\underline{z}}(z) = \int_0^z f_{\underline{z}}(x)dx = P(\underline{x}_1 + \ldots + \underline{x}_n \le z) = P\left(\frac{\underline{x}_1 + \ldots + \underline{x}_n - n/\lambda}{\sqrt{n}/\lambda} \le \frac{z - n/\lambda}{\sqrt{n}/\lambda}\right)$$

$$\approx \Phi\left(\frac{z - n/\lambda}{\sqrt{n}/\lambda}\right)\ .$$

Als we, uitgaande van (3), $F_{\underline{z}}(z)$ willen tabelleren, hebben we een tabel met 2 of 3 "ingangen" nodig; door deze benadering te gebruiken hebben we een tabel van Φ met maar 1 ingang nodig.

<u>Voorbeeld 2</u>: wat is bij 200 onafhankelijke experimenten elk met kans $\frac{1}{3}$ op succes de kans op hoogstens 55 successen?

<u>Antwoord</u>: het aantal successen $\underline{k} = \underline{x}_1 + \ldots + \underline{x}_{200}$ met $E\underline{x}_j = p = \frac{1}{3}$, var $\underline{x}_j = p(1-p) = \frac{2}{9}$. Op grond van de C.L.S. geldt dus

$$(4) \qquad P(\underline{k} \le k) = P\left(\frac{\underline{k} - np}{\sqrt{np(1-p)}} \le \frac{k - np}{\sqrt{np(1-p)}}\right) \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right) .$$

Voor $n = 200$, $p = \frac{1}{3}$ en $k = 55$ vinden we dan

$$P(\underline{k} \le 55) = P\left(\frac{\underline{k} - 200/3}{\sqrt{200.2/9}} \le \frac{55 - 200/3}{20/3}\right) \approx \Phi(-1{,}75) = 1 - \Phi(1{,}75) = 0{,}0401.$$

<u>Opmerking 1</u>: Om de binomiale verdeling met succes m.b.v. de normale verdeling te kunnen benaderen moet n niet te klein zijn en $p(1-p)$ niet te klein. We zullen hier de volgende regel hanteren: benader met de normale verdeling als $np \ge 5$ én $n(1-p) \ge 5$; benader (vergelijk blz. 24) als $p \le 0{,}1$ of als $1 - p \le 0{,}1$ met de Poisson-verdeling (resp. met $\mu = np$ en $\mu = n(1-p)$, en natuurlijk zolang de tabel 4.2 dit toelaat). Deze regel dekt de meest voorkomende gevallen, maar niet alle; de regel is niet eenduidig. De Poisson-verdeling kan op analoge wijze m.b.v. de normale verdeling worden benaderd als μ groot is ($\mu > 10$):

$$P(\underline{k} \le k) = P\left(\frac{\underline{k} - \mu}{\sqrt{\mu}} \le \frac{k - \mu}{\sqrt{\mu}}\right) \approx \Phi\left(\frac{k - \mu}{\sqrt{\mu}}\right)$$

o.g.v. de C.L.S., omdat $\underline{k}$ geschreven kan worden als

$$\underline{k} = \underline{k}_1 + \underline{k}_2 + \ldots + \underline{k}_n ,$$

waarbij de $\underline{k}_j$ o.o. zijn en Poisson-verdeeld met $E\underline{k}_j = \mu/n$ (vergelijk opgave b. blz. 41). *Waar mogelijk gebruiken we een exacte tabel.*

<u>Opnerking 2.</u> De C.L.S. doet geen uitspraak over de grootte van de fout bij deze benadering. In het algemeen zijn deze fouten niet groter dan "van de orde" $\frac{1}{\sqrt{n}}$. In de meeste praktische gevallen is de benadering redelijk goed. Soms worden correcties op de normale benadering aangebracht, die er op neerkomen dat men $P(\underline{k} = k)$ benadert door

$$\Phi\left(\frac{k + \frac{1}{2} - E\underline{k}}{\sigma\underline{k}}\right) - \Phi\left(\frac{k - \frac{1}{2} - E\underline{k}}{\sigma\underline{k}}\right) .$$

4.3. Momentenvoortbrengende functies

Een handig hulpmiddel bij de bestudering van sommen van o.o. stochastische grootheden is de *momentenvoortbrengende functie*. In deze paragraaf geven we daarvan een aantal eenvoudige toepassingen en tenslotte geven we met behulp van deze functies een schets van het bewijs van de C.L.S.

Definitie 1: de momentenvoortbrengende functie $\varphi_{\underline{x}}(s)$ van $\underline{x}$ is gedefinieerd door

$$\varphi_{\underline{x}}(s) = E\, e^{s\underline{x}} ,$$

dus

$$\varphi_{\underline{x}}(s) = \begin{cases} \sum_{1}^{\infty} e^{sx_j} P(\underline{x} = x_j) & \text{(discrete geval)} \\ \\ \int_{-\infty}^{\infty} e^{sx} f_{\underline{x}}(x)dx & \text{(continue geval)} . \end{cases}$$

Twee eenvoudige eigenschappen van $\varphi_{\underline{x}}$ geven we in

Stelling 5:

(5) $\qquad \varphi_{a\underline{x}+b}(s) = e^{bs}\, \varphi_{\underline{x}}(as)$

(6) $\qquad \varphi_{\underline{x}+\underline{y}}(s) = \varphi_{\underline{x}}(s)\, \varphi_{\underline{y}}(s) \qquad$ *als* $\underline{x}$ en $\underline{y}$ o.o. zijn.

Eigenschap (5) volgt direct uit de definitie, eigenschap (6) volgt uit het feit dat $e^{s(\underline{x}+\underline{y})} = e^{s\underline{x}}\, e^{s\underline{y}}$, waarbij $e^{s\underline{x}}$ en $e^{s\underline{y}}$ o.o. zijn omdat $\underline{x}$ en $\underline{y}$ o.o. zijn. Gebruik nu Stelling 4.1., blz. 36.

Gevolg: als $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ o.o. zijn, dan geldt

(7) $\qquad \varphi_{\underline{x}_1 + \ldots + \underline{x}_n}(s) = \varphi_{\underline{x}_1}(s) \ldots \varphi_{\underline{x}_n}(s)$.

Voorbeelden.

1. Binomiale verdeling : $\varphi_{\underline{k}}(s) = (1-p+pe^s)^n$

2. Poisson-verdeling : $\varphi_{\underline{k}}(s) = e^{\mu(e^s-1)}$

3. Exponentiële verdeling : $\varphi_{\underline{x}}(s) = \frac{\lambda}{\lambda - s}$

4. Gamma-verdeling : $\varphi_{\underline{x}}(s) = (\frac{\lambda}{\lambda - s})^n$

5. Normale verdeling : $\varphi_{\underline{x}}(s) = e^{\mu s+\sigma^2 s^2/2}$.

Stelling 6: als $\underline{x}_1,\ldots,\underline{x}_n$ o.o. en normaal verdeeld zijn met $E\,\underline{x}_j = \mu_j$ en var $\underline{x}_j = \sigma_j^2$, dan is $\sum_1^n \alpha_j \underline{x}_j$ *normaal verdeeld* met (uiteraard)

$$E\,(\sum_1^n \alpha_j \underline{x}_j) = \sum_1^n \alpha_j \mu_j\,,\quad \mathrm{var}(\sum_1^n \alpha_j \underline{x}_j) = \sum_1^n \alpha_j^2 \sigma_j^2\ .$$

Bewijs: gebruik gevolg van Stelling 5 en Stelling 5.

We geven nu (zonder bewijs) twee fundamentele eigenschappen van momentenvoortbrengende functies; de eerste hiervan hebben we al stilzwijgend gebruikt.

Stelling 7:

1. *Eenduidigheid*: $\varphi_{\underline{x}}$ bepaalt de kansverdeling van $\underline{x}$, d.w.z. als $\varphi_{\underline{x}} = \varphi_{\underline{y}}$ dan is $F_{\underline{x}} = F_{\underline{y}}$ ($f_{\underline{x}} = f_{\underline{y}}$ etc.).

2. Zogenaamde *continuiteitsstelling:* als $\underline{z},\underline{z}_1,\underline{z}_2,\ldots$ stochastische grootheden zijn, dan geldt (voor $n \to \infty$)

(8) $$\varphi_{\underline{z}_n}(s) \to \varphi_{\underline{z}}(s) \Leftrightarrow F_{\underline{z}_n}(z) \to F_{\underline{z}}(z)\ .$$

<u>Opgave 1</u>: maak opgave 1, blz. 35 met behulp van Stelling 5.

<u>Opgave 2</u>: maak de opgave op blz. 41 met behulp van (6). Generaliseer het resultaat tot het geval van N o.o. grootheden.

Door differentiatie (onder som- of integraalteken; zie Definitie 1) en door reeksontwikkeling van $e^{s\underline{x}}$ $\left(e^{sx} = \sum_0^\infty \frac{(sx)^n}{n!}\right)$ vinden we nog (geen bewijs):

<u>Stelling 8</u>: 1. $E\,\underline{x}^\ell = \varphi_{\underline{x}}^{(\ell)}(0)$,

en speciaal

$$E\,\underline{x} = \varphi_{\underline{x}}'(0) \ , \quad E\,\underline{x}^2 = \varphi_{\underline{x}}''(0) \ ;$$

$$2. \ \varphi_{\underline{x}}(s) = \sum_{\ell=0}^{\infty} E\,\underline{x}^\ell \frac{s^\ell}{\ell!} \ .$$

Eigenschap 2 geeft de momentenvoortbrengende functie haar naam.

<u>Gevolg</u>: voor kleine waarden van s geldt (we verwaarlozen hogere machten van s)

$$\text{(9)} \qquad \varphi_{\underline{x}}(s) \approx 1 + s\,E\,\underline{x} + \tfrac{1}{2}s^2\,E\,\underline{x}^2 \ .$$

We geven nu met behulp van (9), (7) en (5) een schets van het bewijs van de C.L.S. (Stelling 4, p. 43). We hebben achtereenvolgens, omdat $E\,\frac{\underline{x}_j-\mu}{\sigma} = 0$, $E\left(\frac{\underline{x}_j-\mu}{\sigma}\right)^2 = \operatorname{var}\frac{\underline{x}_j-\mu}{\sigma} = 1$ (zie Stelling 1, blz. 33)

$$\varphi_{\frac{\underline{x}_j-\mu}{\sigma}}(s) \approx 1 + \tfrac{1}{2}s^2 \qquad \text{(s klein)}$$

$$\varphi_{\frac{\underline{x}_j-\mu}{\sigma\sqrt{n}}}(s) \approx 1 + \frac{1}{2n}s^2 \qquad \text{(s vast, n groot)}$$

$$\varphi_{\underline{U}_n}(s) \approx \left(1 + \frac{s^2}{2n}\right)^n \to e^{\frac{1}{2}s^2} \qquad (n \to \infty) \ ,$$

waarbij

$$\underline{U}_n = \frac{\underline{x}_1 + \ldots + \underline{x}_n - n\mu}{\sigma\sqrt{n}} = \sum_{j=1}^{n} \frac{\underline{x}_j - \mu}{\sigma\sqrt{n}} \ .$$

Op grond van Stelling 7 geldt dus

$$F_{\underline{U}_n}(z) \to \Phi(z) \ .$$

5. Statistiek

5.1. Onbekende parameters en steekproeven

Tot nu toe hebben we steeds kansen uitgerekend op grond van volledig gespecificeerde modellen. In de praktijk hebben we dikwijls te maken met onvolledig bekende modellen en hebben we *waarnemingen* nodig om het model te completeren. Dit is het werkterrein van de *statistiek*.

Voorbeeld 1: Een fabrikant van kogellagers overweegt de koop van een partij van 100.000 kogels. Een kogel is bruikbaar als hij een diameter heeft tussen de 4,90 en 5,10 millimeter en de fabrikant wil weten hoeveel van de 100.000 kogels bruikbaar zijn. Uit vroegere ervaring is bekend dat de diameter, die enigszins van het toeval afhangt, een normale verdeling heeft met een standaardafwijking $\sigma = 0{,}05$. De verwachte diameter μ is niet bij iedere partij dezelfde. De vraag van de fabrikant komt neer op: wat is de kans dat een kogel een diameter heeft tussen 4,90 en 5,10 millimeter; deze kans is (verg. Opmerking op blz. 43) ongeveer gelijk aan de *fractie* bruikbare kogels en vermenigvuldiging met 100.000 geeft het gevraagde aantal. Om de gevraagde kans te kunnen uitrekenen moeten we eerst (d.m.v. een steekproef) iets over de onbekende parameter μ te weten zien te komen.

Voorbeeld 2: een radiomonteur heeft 10.000 transistoren nodig. Omdat er meestal een vrij grote fractie van de transistoren defect is, zal hij er meer dan 10.000 moeten kopen. Hij wil er zoveel kopen dat hij een kans van 0,999 heeft om minstens 10.000 niet-defecte transistoren te hebben. Om het aantal dat hij moet kopen te kunnen berekenen, zal hij eerst een goede indruk moeten hebben van de fractie defecte transistoren in de partij waarvan hij koopt. Om deze fractie p (= kans dat een exemplaar defect is) te leren kennen neemt hij een steekproef van 200 stuks.

De practische statistiek houdt zich bezig met het nemen (of voorbereiden) van beslissingen op grond van waarnemingen *(steekproeven)*; de theoretische statistiek verschaft hiervoor de methoden. Wij zullen ons hier bezighouden met methoden, die o.g.v. steekproeven tot *conclusies* leiden over de *kansverdelingen* die het gedrag van een *"populatie"* bepalen. Onder de populatie verstaan we de verzameling van alle exemplaren of elementen (dikwijls zijn dit metingen of waarnemingen of zelfs hypothetische of toekomstige waarnemingen) waarop het onderzoek betrekking heeft en waarop de conclusies van toepassing zijn. Het is niet altijd precies duidelijk welke elementen tot de populatie behoren. In de zojuist genoemde voorbeelden bestaat de populatie uit resp. 100.000 kogels en een minder precies vastgelegde verzameling transistoren, waarvan er een (nog onbekend) aantal gekocht moet worden. De kansverdelingen, die op één parameter na (resp μ en p) bekend zijn, zijn de kansverdeling van de diameter van een willekeurig gekozen kogel en de kans dat een willekeurig gekozen transistor defect is.

Er zijn verschillende redenen om alleen een steekproef te onderzoeken en niet de hele populatie. Dikwijls zou het laatste gewoon te veel werk zijn; soms bestaat de populatie deels uit waarnemingen die nog gedaan moeten worden en wil men daarover juist voorspellingen doen; in andere gevallen worden de elementen van de steekproef vernield; levensduur- en treksterktebepalingen, valproeven e.d.. We geven nu een formele definitie van het begrip *aselecte steekproef:*

<u>Definitie 1</u>: een aselecte steekproef van de grootte n uit een populatie met verdelingsfunctie F is een *rij onafhankelijke stochastische grootheden,* $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$, met $F_{\underline{x}_k} = F$ voor $k = 1,2,\ldots,n$. (In speciale gevallen spreekt men ook over een steekproef uit een populatie met kansdichtheid f, of met discrete verdeling $P(\underline{x}_k = u_j) = p_j$). Als $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ dezelfde verdeling hebben als een gegeven stochastische grootheid $\underline{x}$, dan spreekt men van een steekproef van de stochastische grootheid $\underline{x}$.

Meestal hangt F, en dus f (of p_j) nog af van een onbekende parameter, die we aangeven met θ (in de voorbeelden $\theta = \mu = E\underline{x}$ resp. $\theta = p$), zodat $F = F(x;\theta)$, $f = f(x;\theta)$ en $p_j = p_j(\theta)$. Op grond van de steekproef willen we conclusies trekken over de waarde van θ. Voor een gegeven populatie heeft θ een vaste (onbekende) waarde. De verzameling van waarden die θ zou kunnen hebben, de *parameterruimte* geven we aan met Ω. In voorbeeld 1 kunnen we nemen $\Omega = (0,\infty)$, in voorbeeld 2: $\Omega = [0,1]$.

<u>Opmerking 1</u>: steekproeven zijn het resultaat van onafhankelijke metingen, of waarnemingen bij onafhankelijke experimenten. Voorbeelden hiervan zijn trekken met terugleggen uit een eindige populatie(vergelijk blz. 22) en (bij goede benadering) trekken zonder terugleggen uit een zeer grote populatie (groot t.o.v. de steekproefomvang). In de meeste gevallen zullen we er zonder meer van uitgaan dat we te maken hebben met o.o. waarnemingen,

zonder te vermelden hoe die tot stand zijn gekomen. Het voorvoegsel *"aselect"* duidt aan dat we de elementen van de steekproef *niet selecteren*, bijv. naar grootte of andere kenmerken, *zodat de samenstelling van de steekproef in overeenstemming is met de samenstelling van de populatie.*

Opmerking 2: *nadat* het experiment is uitgevoerd is een *realisatie* van de steekproef ontstaan. We hebben dan een rijtje getallen $x_1, x_2, \ldots, x_n$ (geen stochastische grootheden), dat we de *waargenomen* steekproef of "de waarnemingen" zullen noemen. We kunnen niet spreken over de kans dat dit rijtje getallen bepaalde eigenschappen bezit, bijv. $(x_1, \ldots, x_n) \in A$, waarbij A een deel is van $\mathbb{R}^n$. Wel is het o.g.v. de wet v.d. grote aantallen zó dat, als $P((\underline{x}_1, \ldots, \underline{x}_n) \in A) = p$, herhaald nemen van een steekproef bij een fractie van (ongeveer) p van de steekproeven $(x_1, \ldots, x_n) \in A$ zal zijn.

We behandelen achtereenvolgens drie statistische methoden, waarbij we steeds uitgaan van een steekproef van $\underline{x}$ met $f_{\underline{x}}(x) = f(x;\theta)$ (of $P(\underline{x} = x_j) = p_j(\theta)$, $j = 1,2,\ldots$, of algemeen : met $F_{\underline{x}}(x) = F(x;\theta)$) waarbij θ een onbekende parameter is. De drie methoden zijn

1. Schatten van een parameter: hierbij wordt o.g.v. de waarnemingen één (aannemelijke) waarde van θ bepaald. Zo'n waarde noemen we een *puntschatting.*

2. Toetsen van een hypothese: hierbij gaan we na of de aanname (hypothese) dat θ een bepaalde opgegeven waarde θ_0 heeft al of niet aanvaardbaar is, gezien de waargenomen steekproef.

3. Methode van het betrouwbaarheidsinterval: hierbij wordt een interval aangegeven, waarvan het, gezien de waarnemingen, aannemelijk is dat het de onbekende parameterwaarde zal bevatten. Men noemt zo'n interval een *betrouwbaarheidsinterval* of een *intervalschatting.* Deze methode kan beschouwd worden als een combinatie van 1 en 2.

5.2. Schatten van een parameter

In deze paragraaf beperken we ons tot een aantal definities en een paar voorbeelden. We geven geen algemene methode om parameters te schatten, omdat hiervoor de tijd ontbreekt.

Definitie 2: als $\underline{x}_1,\dots,\underline{x}_n$ een aselecte steekproef is van $\underline{x}$ met $F_{\underline{x}}(x) = F(x;\theta)$, dan heet een stochastische grootheid $\underline{t} := t(\underline{x}_1,\dots,\underline{x}_n)$, die een functie is van de steekproef $\underline{x}_1,\dots,\underline{x}_n$, maar *niet* van θ een *steekproefgrootheid* of een *statistische* grootheid. Als $\underline{t}$ gebruikt wordt om θ te schatten, dan heet $\underline{t}$ een *schatter* (of: een *puntschatter*) voor θ.

Opmerking 1: na realisatie van de steekproef is $t := t(x_1,\dots,x_n)$ een *getal*. Dit getal noemen we een *schatting* (puntschatting) voor θ.

Opmerking 2: de functie t hangt ook van n af. Als dit een rol speelt, dan schrijven we t_n (en analoog $\underline{t}_n$) in plaats van t (en $\underline{t}$).

Voordat we een paar voorbeelden van schatters bespreken, definiëren we een paar eigenschappen die wenselijk zijn voor schatters.

Definitie 3: een schatter $\underline{t}_n = t_n(\underline{x}_1,\dots,\underline{x}_n)$ voor θ heet *zuiver* als

$$E\underline{t}_n = \theta \text{ (voor alle } \theta \in \Omega) \ .$$

$\underline{t}_n$ heet asymptotisch zuiver als $E\underline{t}_n \to \theta$ voor $n\to\infty$ (voor alle $\theta \in \Omega$). Men noemt $E\underline{t}_n - \theta$ de *onzuiverheid*.

Definitie 4: als $\underline{t}$ een schatter is voor θ, dan heet

$$\sqrt{E(\underline{t} - \theta)^2}$$

de *onnauwkeurigheid* van $\underline{t}$. Als de onnauwkeurigheid van $\underline{t}_1$ kleiner is dan die van $\underline{t}_2$, dan heet $\underline{t}_1$ nauwkeuriger dan $\underline{t}_2$.

Opmerking: $E(\underline{t} - \Theta)^2 = \text{var } \underline{t} + (E\underline{t} - \Theta)^2$, dus als $\underline{t}$ zuiver is, dan is $\sqrt{E(\underline{t} - \Theta)^2} = \sigma(\underline{t})$.

We geven nu een paar eigenschappen van een voor de hand liggende schatter voor $E\underline{x}$, n.l. het *steekproefgemiddelde* $\bar{\underline{x}} = \frac{1}{n} \sum_1^n \underline{x}_j$.

Stelling 1: $\bar{\underline{x}}$ is een zuivere schatter voor $E\underline{x}$.

Bewijs: zie blz. 37.

Stelling 2: van alle zuivere lineaire schatters voor $E\underline{x}$, dat zijn schatters van de vorm

$$\underline{t} = \sum_1^n \alpha_j \underline{x}_j$$

met $\sum_1^n \alpha_j = 1$, is $\bar{\underline{x}}$ de nauwkeurigste.

<u>Bewijs</u>: omdat $\underline{t}$ zuiver is, moeten we bewijzen dat var $\underline{t}$ minimaal is voor $\underline{t} = \bar{\underline{x}}$, d.w.z. als $\alpha_1 = \alpha_2 = \ldots = \alpha_n = \frac{1}{n}$. We schrijven $E\underline{x} = \mu$, var $\underline{x} = \sigma^2$ en $\delta_j = \alpha_j - \frac{1}{n}$, zodat $\sum_1 \delta_j = 0$. Dan geldt (zie Stelling 4, 3^b, blz.36)

$$\text{var } \underline{t} = \sum_1^n \alpha_j^2 \text{ var } \underline{x}_j = \sigma^2 \sum_1^n (\tfrac{1}{n} + \delta_j)^2 = \frac{\sigma^2}{n} + \sigma^2 \sum_1^n \delta_j^2,$$

zodat var $\underline{t}$ minimaal is als $\delta_1 = \delta_2 = \ldots = \delta_n = 0$, d.w.z. als $\underline{t} = \bar{\underline{x}}$.

De variantie van $\underline{x}$ schatten we met de zgn. *steekproefvariantie* $\underline{s}^2$:

$$\underline{s}^2 = \frac{1}{n-1} \sum_1^n (\underline{x}_j - \bar{\underline{x}})^2.$$

<u>Stelling 3</u>: de steekproefvariantie $\underline{s}^2$ is een zuivere schatter voor var $\underline{x}$.

<u>Bewijs</u>: zij $E\underline{x} = \mu$ en var $\underline{x} = \sigma^2$, dan hebben we

$$(n-1)E\underline{s}^2 = E \sum_1^n (\underline{x}_j - \bar{\underline{x}})^2 = E \sum_1^n \{(\underline{x}_j - \mu) - (\bar{\underline{x}} - \mu)\}^2 =$$

$$= \sum_1^n E(\underline{x}_j - \mu)^2 + \sum_1^n E(\bar{\underline{x}} - \mu)^2 - 2E(\bar{\underline{x}} - \mu) \sum_1^n (\underline{x}_j - \mu)$$

$$= n\sigma^2 + n \text{ var } \bar{\underline{x}} - 2nE(\bar{\underline{x}} - \mu)^2 = n\sigma^2 + \sigma^2 - 2\sigma^2 = (n-1)\sigma^2,$$

immers var $\bar{\underline{x}} = \sigma^2/n$ en $\sum_1^n (\underline{x}_j - \mu) = n(\bar{\underline{x}} - \mu)$ (zie p. 34, voorb. 1a en p. 37).

<u>Opmerking 1</u>: *als* $E\underline{x} = \mu$ *bekend* is, dan is

$$\underline{s}_0^2 = \frac{1}{n} \sum_1^n (\underline{x}_j - \mu)^2$$

een zuivere schatter voor var $\underline{x}$.

Opmerking 2: omdat $\bar{\underline{x}}$ een goede schatter is voor $E\underline{x}$, is ook (speciaal geval hiervan) $\underline{k}/n$ een goede schatter voor de succeskans p, als $\underline{k}$ het aantal successen is bij n o.o. experimenten met kans p op succes (zie blz. 40)

Opgave: $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n$ is een aselecte steekproef uit een populatie met een homogene verdeling op $(0,\theta)$ (zie blz.16). Laat zien dat $\frac{n+1}{n}\max(\underline{x}_1,\ldots,\underline{x}_n)$ een zuivere schatter is voor θ (zie ook blz. 30, Opgave 1).

5.3. Toetsen van een hypothese.

Als een kippenhouder een zeer groot ei vindt, dat bij preciese weging 200,107 gram blijkt te wegen is hij zeer verbaasd. Niet omdat het ei dit preciese gewicht heeft (een ei van precies 60,107 gram is ook een uitzondering), maar omdat het *zwaarder* is dan, zeg, 200 gram. Hij zal zich wellicht afvragen of er niet een gans of zelfs een zwaan op bezoek geweest is. Op dezelfde manier is men verbaasd als men met een dobbelsteen in 100 worpen 51 zessen gooit: de kans op 51 zessen is klein, maar dat geldt (in iets mindere mate ook voor 21 zessen (kans <0,06);ook hier komt de verbazing voort uit het feit dat er zoveel , *meer dan* 50 zessen zijn gegooid. De kans hierop is (in tegenstelling tot de kans op meer dan 20 zessen) zeer klein: $<10^{-20}$.

In het geval van de dobbelsteen zal men bij 51 (of meer) zessen sterk betwijfelen of de kans (per worp) op een zes wel 1/6 is. Dit soort twijfel is de kern van de toetsingstheorie: als de *kans op overschrijding* (het zij naar boven of naar beneden) van de waarde van een waargenomen grootheid, de zgn. *overschrijdingskans, bij gebruik van een bepaald model,* erg klein is, dan geloven we niet dat dit model de werkelijkheid weergeeft. In het voorbeeld van de dobbelsteen wil dat zeggen dat we niet geloven dat de kans op een zes gelijk is aan 1/6. Ook bij 35 of zelfs bij 25 zessen zal een (ervaren) dobbelaar (of een statisticus) twijfelen aan de dobbelsteen. Bij welke waarde we moeten gaan twijfelen is een kwestie van kansrekening: wat is de kans op overschrijding van een gegeven waarde, en een kwestie van smaak: welke kans vinden we nog "klein".

Hierboven komen twee methoden naar voren om na te gaan, te *toetsen*, of de hypothese $p := P(\text{zes}) = 1/6$ aanvaardbaar is: we kunnen eerst 100 keer gooien en dan uitrekenen wat de kans is om het gevonden aantal zessen (naar boven of naar beneden) te overschrijden, óf we kunnen eerst twee grenzen voor dit aantal vastleggen (door berekening en aan de hand van een gegeven criterium) waarbinnen we de hypothese $p = 1/6$ nog accepteren. We zullen beide methoden bespreken aan de hand van een situatie als Voorbeeld 1 op blz. 49.

Voorbeeld A. *(toets voor de verwachting bij een normale verdeling):*

Laat $x_1, x_2, \ldots, x_n$ waarnemingen zijn van een normaal verdeelde grootheid $\underline{x}$ met $E\underline{x} = \mu$ en var $\underline{x} = \sigma^2$, waarbij σ^2 bekend is en μ de onbekende parameter. We willen toetsen of de hypothese $\mu = \mu_0$ (in Voorbeeld 1 is $\mu_0 = 5$), gezien de waarnemingen, houdbaar is. Omdat we iets willen zeggen over $\mu = E\underline{x}$ en omdat $\bar{\underline{x}}$ een goede schatter is voor μ (zie Stelling 1 en 2, blz. 52, 53), ligt het voor de hand om de waargenomen waarde $\bar{x}$ van $\bar{\underline{x}}$ te toetsen aan de waarde μ_0 (we noemen $\bar{\underline{x}}$ in dit verband een *toetsingsgrootheid)*. We zullen de hypothese $\mu = \mu_0$ verwerpen, als $\bar{x}$ "veel" van μ_0 afwijkt, of liever: als een van de *overschrijdingskansen*, d.w.z. de kansen om de gevonden waarde $\bar{x}$ naar boven resp. naar beneden te overtreffen "klein" is. Dit "veel" en "klein" specificeren we als volgt: we verwerpen μ_0 als $\bar{x}$ zo groot is, dat voor de *rechter overschrijdingskans* geldt

(1) $\qquad P(\bar{\underline{x}} \geq \bar{x};\ \mu_0) \leq \alpha/2,$

of als $\bar{x}$ zo klein is dat voor de *linker overschrijdingskans* geldt

(2) $\qquad P(\bar{\underline{x}} \leq \bar{x};\ \mu_0) \leq \alpha/2.$

Hierbij betekent $P(\bar{\underline{x}} \geq \bar{x};\ \mu_0)$ de kans op "$\bar{\underline{x}} \geq \bar{x}$", uitgerekend met een normale verdeling met *verwachting* μ_0 (en variantie σ^2) voor $\underline{x}_1, \ldots \underline{x}_n$. Men schrijft ook wel $P_{\mu_0}(\bar{\underline{x}} \geq \bar{x})$. Hoe klein α genomen moet worden is een kwestie van smaak: wanneer is een kans "klein"; traditionele waarden zijn $\alpha = 0{,}05$ en $\alpha = 0{,}01$. We schrijven in (1) en (2) $\alpha/2$, omdat we de kans op verwerpen van de juiste waarde, dus de som van linker- en rechter overschrijdingskans $\leq \alpha$ willen hebben; om redenen van symmetrie kiezen we de bovengrenzen voor linker- en rechteroverschrijdingskans gelijk.

Opgave: niet beide kansen in (1) en (2) kunnen kleiner dan $\alpha/2$ zijn; waarom niet ?

Als we de toets één keer uitvoeren (zoals op een tentamen), dan is het voldoende om de overschrijdingskansen (1) en (2) uit te rekenen en te kijken of één van beide misschien $\leq \alpha/2$ is. Als we de toets vaker uitvoeren (of als dit op een tentamen nadrukkelijk wordt gevraagd) dan is het verstandiger (en handiger) om van tevoren na te gaan welke waarden van $\bar{x}$ aan (1) of (2) voldoen; dat zijn alle $\bar{x}$ - waarden met $\bar{x} \geq r$, waarbij voor r geldt

(3) $\qquad P(\underline{\bar{x}} \geq r;\ \mu_0) = \alpha/2$

en alle $\bar{x}$ - waarden met $\bar{x} \leq \ell$, waarbij ℓ voldoet aan

(4) $\qquad P(\underline{\bar{x}} \leq \ell;\ \mu_0) = \alpha/2.$

De getallen ℓ en r heten *kritieke waarden* en de verzameling van alle $\bar{x}$ - waarden die tot verwerping van μ_0 leiden heet het *kritieke gebied*, dat we aangeven met K, dus

(5) $\qquad K = \{\bar{x} \mid \bar{x} \leq \ell \text{ of } \bar{x} \geq r\}.$

Bij gegeven α berekenen we ℓ en r als volgt: onder de hypothese $\mu = \mu_0$ zijn $\underline{x}_1, \ldots, \underline{x}_n$ $N(\mu_0, \sigma^2)$ en dus (zie Stelling 6, blz. 46) $\underline{\bar{x}}$ is $N(\mu_0, \sigma^2/n)$ (zie voor notatie blz. 17) zodat (zie Opgave 1, blz. 35) $(\underline{\bar{x}} - \mu_0)/(\sigma/\sqrt{n})$ $N(0,1)$ is. Nu geldt dus

$$\alpha/2 = P(\underline{\bar{x}} \geq r;\ \mu_0) = P\left(\frac{\underline{\bar{x}} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{r - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{r - \mu_0}{\sigma}\sqrt{n}\right).$$

Als we i.h.a u_γ definiëren door

$$1 - \Phi(u_\gamma) = \gamma,$$

dan is dus $\dfrac{r - \mu_0}{\sigma}\sqrt{n} = u_{\alpha/2}$ of

(6) $\qquad r = \mu_0 + \sigma u_{\alpha/2}/\sqrt{n},$

waarbij we $u_{\alpha/2}$ kunnen opzoeken in een tabel van de normale verdelingsfunctie. Op dezelfde manier vinden we (ga na)

(7) $\qquad \ell = \mu_0 - \sigma u_{\alpha/2}/\sqrt{n}$

We verwerpen dus μ_0 als $\bar{x} \in K$, d.w.z. als $\bar{x} \leq \ell$ of $\bar{x} \geq r$. De conclusie is dan $\mu \neq \mu_0$. Bij niet-verwerpen zijn er verschillende mogelijkheden: men aanvaardt $\mu = \mu_0$ of men zet het onderzoek voort.

<u>Opgave 1</u>: de fabrikant van Voorbeeld 1 (blz. 49) neemt een steekproef van 50 stuks om te toetsen of $\mu = 5{,}00$. Hij vindt $\bar{x} = 5{,}02$ en gebruikt $\alpha = 0{,}01$. Kan hij concluderen dat $\mu \neq 5{,}00$? (Er is gegeven dat $\sigma(\underline{x}_j) = 0{,}05$ voor $j = 1,2,\ldots,50$).

Bij het uitvoeren van bovenstaande toets kunnen twee soorten "fouten" optreden:

Fout van de eerste soort: μ_0 verwerpen terwijl $\mu = \mu_0$.
De kans hierop is (per definitie , zie (3), (4) en (5))

$$P(\bar{\underline{x}} \in K;\ \mu_0) = P(\bar{\underline{x}} \leq \ell;\ \mu_0) + P(\bar{\underline{x}} \geq r;\ \mu_0) = \alpha.$$

Men noemt α de *onbetrouwbaarheid* van de toets.

Fout van de tweede soort: μ_0 niet verwerpen terwijl $\mu \neq \mu_0$.
De kans hierop hangt natuurlijk af van de waarde van μ. We geven deze kans aan met $\beta^*(\mu)$. Hiervoor geldt

$$\beta^*(\mu) = P(\bar{\underline{x}} \notin K;\ \mu) = P(\ell < \bar{\underline{x}} < r;\ \mu).$$

Men noemt $\beta(\mu) := 1 - \beta^*(\mu)$ het *onderscheidingsvermogen* van de toets, dus

$$\beta(\mu) = P(\bar{\underline{x}} \in K;\ \mu)$$

<u>Opgave 2</u>: ga na dat $\beta(\mu) = \Phi(\frac{\ell - \mu}{\sigma/\sqrt{n}}) + 1 - \Phi(\frac{r - \mu}{\sigma/\sqrt{n}})$.

<u>Opgave 3</u>: kies $\alpha = 0{,}0456$ (zodat $u_{\alpha/2} = 2$), $n = 100$ en $\sigma^2 = 1$. Teken een plaatje van $\beta(\mu)$ als functie van μ . Laat zien dat $\beta(\mu)$ minimaal is voor $\mu = \mu_0$ (wat is $\beta(\mu_0)$?). Doe hetzelfde voor $n = 400$.

<u>Opmerking</u>: in het geval dat $\underline{x}$ de diameter is van een kogel die in een lager moet passen, zullen we $\mu = \mu_0$ verwerpen als $\bar{x}$ groot én als $\bar{x}$ klein is: zowel te grote als te kleine kogels zijn onbruikbaar. Is $\underline{x}$ bijv. het gewicht van pakjes boter die we kunnen kopen, met een verondersteld gewicht van $\mu_0 = 250$ gram, dan zullen bij toetsing van μ_0 *alleen kleine waarden* van $\bar{x}$ tot verwerping leiden: we maken geen bezwaar tegen een gemiddeld gewicht van méér dan 250 gram. We verwerpen nu μ_0 als $\bar{x}$ zó is dat

$$P(\underline{\bar{x}} \leq \bar{x};\ \mu_0) \leq \alpha.$$

We noemen de toets nu *eenzijdig* (in dit geval *links-eenzijdig)*. Het bijbehorende kritieke gebied, K', is nu ook (links-) eenzijdig:

$$K' = \{\bar{x} \mid \bar{x} \leq \ell'\}.$$

met ℓ' zó dat

$$(4')\qquad P(\underline{\bar{x}} \leq \ell';\ \mu_0) = \alpha;$$

dus de onbetrouwbaarheid van de toets is weer α. We verwerpen nu μ_0 als $\bar{x} \in K'$, d.w.z. als $\bar{x} \leq \ell'$; we concluderen dan $\mu < \mu_0$. Analoog geldt bij een rechtseenzijdige toets (als $\underline{x}$ bijv. de hoeveelheid verontreiniging is per cm^3 lucht, die gemiddeld niet boven μ_0 mag komen) voor het rechtseenzijdige kritieke gebied K":

$$K'' = \{\bar{x} \mid \bar{x} \geq r''\}$$

met

$$(3'')\qquad P(\underline{\bar{x}} \geq r'';\ \mu_0) = \alpha.$$

We verwerpen μ_0 als $\bar{x} \in K''$ en concluderen dan $\mu > \mu_0$. We berekenen ℓ' en r'' op dezelfde manier als ℓ en r; we vinden

$$(7')\qquad \ell' = \mu_0 - \sigma u_\alpha / \sqrt{n}$$

$$(6'')\qquad r'' = \mu_0 + \sigma u_\alpha / \sqrt{n}.$$

We geven nu in iets algemenere vorm het schema voor het toetsen van een hypothese.

1. We gaan uit van n waarnemingen $x_1, x_2, \ldots, x_n$ van een grootheid $\underline{x}$ met verdelingsfunctie $F_{\underline{x}}(x) = F(x;\theta)$ met θ onbekend

2. We willen de hypothese $H_0 : \theta = \theta_0$ (ook wel *nul-hypothese* genoemd) toetsen tegen de *alternatieve hypothese* $H_a : \theta \neq \theta_0$ (bij tweezijdige toetsing) of tegen $\theta < \theta_0$, resp. $\theta > \theta_0$ (bij links-, resp. rechts- eenzijdige toetsing).

3. We gebruiken hiervoor een *toetsingsgrootheid* $\underline{t} = t(\underline{x}_1,\ldots,\underline{x}_n)$ (meestal een goede schatter voor θ of een functie daarvan) en we verwerpen H_0 "ten gunste van" H_a als één van de overschrijdingskansen

$$P(\underline{t} \geq t;\ H_0)$$

$$P(\underline{t} \leq t;\ H_0)$$

voldoende klein is, d.w.z. één van beide $\leq \alpha/2$ bij tweezijdige toetsing, $P(\underline{t} \geq t;\ H_0) \leq \alpha$ bij rechts-eenzijdige toetsing of $P(\underline{t} \leq t;\ H_0) \leq \alpha$ bij links-eenzijdige toetsing.

4. In plaats van direct met de overschrijdingskansen te werken, kunnen we alle waarden van t waarvoor de betreffende overschrijdingskans voldoende klein is (dus $\leq \alpha/2$ resp. $\leq \alpha$) bij elkaar nemen tot een *kritiek gebied* K (voor de eenzijdige gevallen K' resp. K"). De kritieke gebieden K, K' en K" hebben dikwijls de vorm $K = \{t \mid t \leq t_1 \text{ of } t \geq t_2\}$, resp. $K' = \{t \mid t \leq t_1'\}$ en $K'' = \{t \mid t \geq t_2''\}$.

5. $P(\underline{t} \in K;\ H_0) = \alpha$ heet de *onbetrouwbaarheid* van de toets en $\beta(\theta) = P(\underline{t} \in K;\ \theta)$ heet het *onderscheidingsvermogen*.

<u>Opmerking</u>: bij de eenzijdige toetsen schrijft men ook wel $H_0 : \theta \geq \theta_0$ (i.p.v. $\theta = \theta_0$) als $H_a : \theta < \theta_0$ en $H_0 : \theta \leq \theta_0$ (i.p.v. $\theta = \theta_0$) als $H_a : \theta > \theta_0$.

<u>Voorbeeld B</u> *(toets voor de kans op succes bij een Bernoulli-verdeling)*:
We bespreken dit voorbeeld (zie ook Voorbeeld 2, p. 49) aan de hand van bovenstaand schema.

1. We hebben waarnemingen $x_1, x_2, \ldots, x_n$ van $\underline{x}$ met $P(\underline{x} = 1) = p$ en $P(\underline{x} = 0) = 1 - p$, waarbij p onbekend is.

2. We toetsen $H_0 : p = p_0$ tegen $H_a : p \neq p_0$

3. Als toetsingsgrootheid kiezen we $\underline{k} = \underline{x}_1 + \ldots + \underline{x}_n$, zodat $\underline{k}$ binomiaal verdeeld is (zie blz. 16). We schrijven $k = x_1 + \ldots + x_n$ en we verwerpen p_0 als

$$P(\underline{k} \leq k;\ p_0) \leq \alpha/2$$

of

$$P(\underline{k} \geq k;\ p_0) \leq \alpha/2.$$

4. Bij de bepaling van het kritieke gebied treedt een moeilijkheid op : de vergelijking

(8) $$P(\underline{k} \leq k_1;\ p_0) = \alpha/2$$

heeft i.h.a. geen oplossing, omdat $\sum_{k=0}^{k_1} \binom{n}{k} p_0^{\,k}(1-p_0)^{n-k}$ sprongsgewijze verandert met k_1. We definiëren nu K (in overeenstemming met 4.,blz. 59) door

$$K = \{k \mid k \leq k_1 \text{ of } k \geq k_2\},$$

waarbij k_1 zo groot mogelijk wordt gekozen onder de voorwaarde dat

(9) $$P(\underline{k} \leq k_1;\ p_0) \leq \alpha/2$$

en k_2 zo klein mogelijk onder de voorwaarde dat

(10) $$P(\underline{k} \geq k_2;\ p_0) \leq \alpha/2$$

De waarden van k_1 en k_2 kunnen worden opgezocht in een tabel van de binomiale verdeling (zie bijv. Statistische Tabellen).
Een gevolg van de onder 4 gesignaleerde moeilijkheid is dat de *onbetrouwbaarheid* $P(\underline{k} \in K;\ p_0)$ i.h.a. *kleiner dan* α is, omdat in (9) en (10) i.h.a. het ongelijkteken zal gelden. Men noemt in dit verband α wel de onbetrouwbaarheidsdrempel: onbetrouwbaarheid $\leq \alpha$.

Opmerking: we kunnen ook hier (desgewenst) éénzijdig toetsen. We krijgen dan bij de links-eenzijdige toets: verwerp p_0 ten gunste van $p < p_0$ als

$$P(\underline{k} \leq k;\ p_0) \leq \alpha$$

of als $k \in K'$ met $K' = \{k \mid k \leq k_1'\}$ waarbij k_1' zo groot mogelijk is, maar met $P(\underline{k} \leq k_1';\ p_0) \leq \alpha$. Bij de rechts-eenzijdige toets: verwerp p_0, ten gunste van $p > p_0$, als

$$P(\underline{k} \geq k;\ p_0) \leq \alpha$$

of als $k \in K''$ met $K'' = \{k \mid k \geq k_2''\}$, waarbij k_2'' zo klein mogelijk is, maar met $P(\underline{k} \geq k_2'';\ p_0) \leq \alpha$.

Opgave: als de radiomonteur in Voorbeeld 2 (blz. 49) in een steekproef van 20 stuks 4 defecten vindt, kan hij dan de bewering van de verkoper dat de partij hoogstens 10% defecten bevat verwerpen? Kies $\alpha = 0{,}05$ en toets rechts-éénzijdig (waarom ?).

Als de steekproef groter is dan 20 stuks dan kunnen we de benodigde kansen niet in de Statistische Tabellen opzoeken. We gebruiken dan meestal (o.g.v. de C.L.S.) een benadering met de normale verdeling (soms, als $p \leq 0{,}10$, een benadering met de Poisson-verdeling; zie Stelling 4, blz. 24). Volgens de C.L.S. geldt (zie Stelling 4, blz. 43 en Voorbeeld 2, blz. 44) voor de overschrijdingskansen:

$$P(\underline{k} \leq k;\, p_0) \approx \Phi\left(\frac{k - np_0}{\sqrt{np_0(1 - p_0)}}\right)$$

resp.

$$P(\underline{k} \geq k;\, p_0) \approx 1 - \Phi\left(\frac{k - np_0}{\sqrt{np_0(1 - p_0)}}\right).$$

De kritieke grenzen k_1 en k_2 vinden we dan (bij benadering) uit (verg. (3), (4), (9) en (10))

$$\Phi\left(\frac{k_1 - n\,p_0}{\sqrt{np_0(1 - p_0)}}\right) = \alpha/2$$

resp.

$$1 - \Phi\left(\frac{k_2 - np_0}{\sqrt{np_0(1 - p_0)}}\right) = \alpha/2,$$

zodat (weer bij benadering)

$$(11) \qquad \begin{cases} k_1 = np_0 - u_{\alpha/2}\sqrt{np_0(1 - p_0)} \\ \\ k_2 = np_0 + u_{\alpha/2}\sqrt{np_0(1 - p_0)} \end{cases}$$

(verg. met (6) en (7) en zie blz.56 voor de definitie van u_γ).

<u>Opgave</u>: als de radiomonteur van Voorbeeld 2 (blz. 49) een steekproef neemt van 100 stuks en 16 defecten vindt, kan hij dan de bewering "hoogstens 10% defect" verwerpen ? Neem $\alpha = 0,05$, zie ook Opgave op blz. 61.

5.4. Betrouwbaarheidsintervallen.

We gaan nog even terug naar de situatie van blz. 54, waar bij 100 worpen met een dobbelsteen 51 zessen werden gegooid. Het is inmiddels duidelijk dat op grond van dit gegeven de hypothese $p := P(\text{zes}) = 1/6$ verworpen moet worden (voor elke redelijke waarde van α). Hadden we niet met een dobbelsteen maar met een munt gegooid, dan hadden we op grond van de waarneming 51 keer kruis de hypothese $p' := P(\text{kruis}) = 1/2$ zeker niet verworpen. De hypothese $p'' := P(\text{succes}) = 1/3$ zou bij 51 successen in 100 experimenten met kans p'' op succes weer wél verworpen worden (ga na). We geven nu twee equivalente definities van het begrip *betrouwbaarheidsinterval*.

<u>Definitie 5</u>: een betrouwbaarheidsinterval (of betrouwbaarheidsgebied) voor een onbekende parameter θ is de verzameling van *alle parameterwaarden* θ_0 *die* op grond van een waargenomen steekproef $x_1, x_2, \ldots, x_n$ van $\underline{x}$ met $F_{\underline{x}}(x) = F(x;\theta)$ met een gegeven toets *niet worden verworpen*. Als α de onbetrouwbaarheid van de toets is, dan noemen we $1 - \alpha$ de *betrouwbaarheid* van het betrouwbaarheidsinterval.

Zo'n betrouwbaarheidsinterval bestaat dus uit alle parameterwaarden, die , gezien de waarnemingen, redelijkerwijs de werkelijke waarde zouden kunnen zijn. Als de toets gebaseerd is op een toetsingsgrootheid $\underline{t}$ die een goede schatter is voor θ, dan zal de puntschatting t bevat zijn in het betrouwbaarheidsinterval, dat men ook wel *intervalschatting* noemt. De eindpunten θ_1 en θ_2 van het betrouwbaarheidsinterval zijn natuurlijk *functies van de waarnemingen* (zie volgende definitie).Verder is het duidelijk uit Definitie 5 dat een betrouwbaarheidsinterval gebruikt kan worden om de toets uit te voeren: verwerp $H_0 : \theta = \theta_0$ als $\theta_0 \notin (\theta_1, \theta_2)$. In de meeste gevallen is dit echter omslachtig.

<u>Definitie 5'</u>:een betrouwbaarheidsinterval met betrouwbaarheid $1-\alpha$ voor een onbekende parameter θ is een interval (θ_1, θ_2), waarbij $\theta_1 = \theta_1(x_1, \ldots, x_n)$

en $\theta_2 = \theta_2(x_1,\dots,x_n)$ functies zijn van een steekproef $\underline{x}_1,\dots,\underline{x}_n$ van $\underline{x}$ met $F_{\underline{x}}(x) = F(x;\theta)$. Hierbij geldt: als $\underline{x}_1,\dots,\underline{x}_n$ een steekproef is van $\underline{x}$ en als $\underline{\theta}_1 = \theta_1(\underline{x}_1,\dots,\underline{x}_n)$ en $\underline{\theta}_2 = \theta_2(\underline{x}_1,\dots,\underline{x}_n)$, dan is

$$(12) \qquad P(\underline{\theta}_1 < \theta < \underline{\theta}_2 \ ; \ \theta) = 1 - \alpha \qquad (\text{of} > 1 - \alpha) \ .$$

De betekenis van (12) is, dat bij vaak herhalen van het experiment het betrouwbaarheidsinterval in een fractie van (ongeveer) $1 - \alpha$ van de gevallen de werkelijke (onbekende) waarde van θ zal bevatten (wet van de grote aantallen, blz. 42 en Opmerking 2 op blz. 51).

Wij zullen steeds uitgaan van Definitie 5 en in een voorbeeld laten zien dat dan ook aan Definitie 5' voldaan is.

<u>Voorbeeld A'</u> *(betrouwbaarheidsinterval voor μ bij normale verdeling):*

We gaan weer uit van n waarnemingen $x_1,\dots,x_n$ en gebruiken de toets gebaseerd op het steekproefgemiddelde $\underline{\bar{x}}$, zoals beschreven op blz. 55. Welke waarden van μ (we schrijven nu μ i.p.v. μ_0 omdat μ variabel is) worden op grond van de waarde $\bar{x}$ niet verworpen ? D.w.z. voor welke μ geldt $\bar{x} \notin K = K(\mu)$ (zie (5), (6) en (7)), of: voor welke μ geldt ($\ell = \ell(\mu)$ en $r = r(\mu)$)

$$\ell(\mu) < \bar{x} < r(\mu) \ ?$$

Invullen van ℓ en r (uit (6) en (7)) geeft

$$(13) \qquad \mu - \sigma u_{\alpha/2}/\sqrt{n} < \bar{x} < \mu + \sigma u_{\alpha/2}/\sqrt{n}.$$

Als we deze ongelijkheden oplossen naar μ , dan zijn alle waarden van μ die bij toetsing niet worden verworpen, diegene die voldoen aan

$$(14) \qquad \bar{x} - \sigma u_{\alpha/2}/\sqrt{n} < \mu < \bar{x} + \sigma u_{\alpha/2}/\sqrt{n}.$$

Het betrouwbaarheidsinterval is dus (μ_1,μ_2) met

$$(15) \qquad \mu_1 = \bar{x} - \sigma u_{\alpha/2}/\sqrt{n}, \quad \mu_2 = \bar{x} + \sigma u_{\alpha/2}/\sqrt{n} \ ,$$

waarbij, zoals te verwachten was, voor de schatter $\bar{x}$ geldt dat $\bar{x} \in (\mu_1,\mu_2)$. Uit (15) volgt (verg. definitie 5')

(15') $\qquad \underline{\mu}_1 = \underline{\bar{x}} - \sigma u_{\alpha/2}/\sqrt{n}, \quad \underline{\mu}_2 = \underline{\bar{x}} + \sigma u_{\alpha/2}/\sqrt{n}.$

Hoe zit het nu met de geldigheid van Definitie 5' ? We hebben achtereenvolgens wegens (15') en de equivalentie van (14) en (13)

$$P(\underline{\mu}_1 < \mu < \underline{\mu}_2 \; ; \mu) = P(\underline{\bar{x}} - \sigma u_{\alpha/2}/\sqrt{n} < \mu < \underline{\bar{x}} + \sigma u_{\alpha/2}/\sqrt{n}) =$$

$$= P(\mu - \sigma u_{\alpha/2}/\sqrt{n} < \underline{\bar{x}} < \mu + \sigma u_{\alpha/2}/\sqrt{n}) = P(- u_{\alpha/2} < \frac{\underline{\bar{x}} - \mu}{\sigma/\sqrt{n}} < u_{\alpha/2}) =$$

$$= \Phi(u_{\alpha/2}) - \Phi(- u_{\alpha/2}) = \Phi(u_{\alpha/2}) - \{ 1 - \Phi(u_{\alpha/2}) \} = 2\,\Phi(u_{\alpha/2}) - 1 = 1 - \alpha.$$

zodat inderdaad aan Definitie 5' is voldaan.

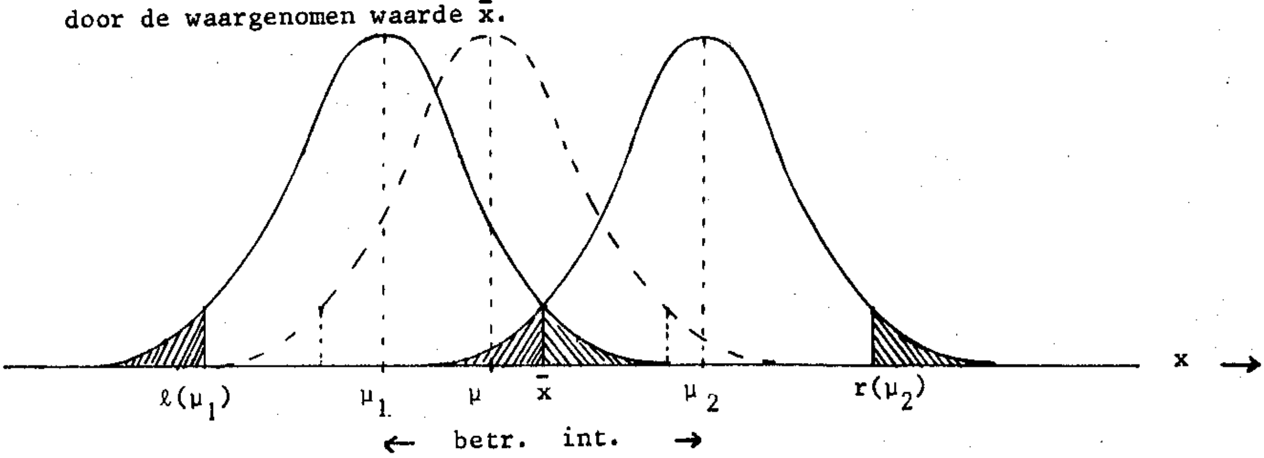
We kunnen μ_1 en μ_2 ook vinden m.b.v. de overschrijdingskansen:
μ_1 wordt nog juist verworpen als

(16) $\qquad P(\underline{\bar{x}} \geq \bar{x}; \mu_1) = \alpha/2,$

d.w.z. als $\frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}} = u_{\alpha/2}$, dus als $\mu_1 = \bar{x} - \sigma u_{\alpha/2}/\sqrt{n}$ (zie (15)). Evenzo wordt μ_2 nog juist verworpen als

(17) $\qquad P(\underline{\bar{x}} \leq \bar{x} \; ; \mu_2) = \alpha/2,$

d.w.z. als $\frac{\bar{x} - \mu_2}{\sigma/\sqrt{n}} = - u_{\alpha/2}$, dus als $\mu_2 = \bar{x} + \sigma u_{\alpha/2}/\sqrt{n}$. Bij deze situaties hoort het onderstaande plaatje, waar de grafieken zijn getekend van $f_{\underline{\bar{x}}}(x;\mu_1)$, $f_{\underline{\bar{x}}}(x;\mu_2)$ en $f_{\underline{\bar{x}}}(x;\mu)$ voor een $\mu \in (\mu_1,\mu_2)$; μ_1 en μ_2 zijn bepaald door de waargenomen waarde $\bar{x}$.



Opmerking: ook bij de eenzijdige toetsen horen (eenzijdige) betrouwbaarheidsintervallen. Bij de links-eenzijdige toets worden grote μ- waarden verworpen. Het bijbehorende betrouwbaarheidsinterval bevat dus kleine μ- waarden, d.w.z. is ook links-eenzijdig; het is van de vorm $(-\infty,\mu'_2)$ met

$$\mu'_2 = \bar{x} + \sigma u_\alpha/\sqrt{n}.$$

Immers, μ'_2 wordt nog juist verworpen als $P(\underline{\bar{x}} \le \bar{x};\mu'_2) = \alpha$, dus als

$$\frac{\bar{x} - \mu'_2}{\sigma/\sqrt{n}} = - u_\alpha .$$

Evenzo hoort bij de rechts-eenzijdige toets een rechts-eenzijdig betrouwbaarheidsinterval (μ'_1 ,∞) met (ga na)

$$\mu'_1 = \bar{x} - \sigma u_\alpha/\sqrt{n}$$

Opgave: ga na dat de betrouwbaarheidsintervallen nauwer worden bij toenemende n (meer informatie) en wijder worden bij afnemende α (hogere eisen aan betrouwbaarheid).

Voorbeeld B' *(betrouwbaarheidsinterval voor* p *bij Bernoulli-verdeling):*
We hebben n waarnemingen $x_1,\ldots,x_n$ van $\underline{x}$ uit een Bernoulli-verdeling met $P(\underline{x} = 1) = p$ en $P(\underline{x} = 0) = 1 - p$, met p onbekend. We behandelen weer het tweezijdige geval, d.w.z. we zoeken naar alle waarden van p die op grond van de waarde $k = x_1 +\ldots+ x_n$ niet worden verworpen (zie de toets op blz.59 e.v.). We doen dit m.b.v. de overschrijdingskansen: we zoeken p_1 en p_2 met $p_1 < p_2$ zó dat p_1 en p_2 nog juist worden verworpen en dus ook alle p met $p < p_1$ of $p > p_2$. Alle p met $p_1 < p < p_2$ vormen dan het gezochte betrouwbaarheidsinterval. We vinden $p_1 = p_1(k)$ uit (verg. (16))

(18) $\qquad P(\underline{k} \ge k;\ p_1) = \alpha/2$

en $p_2 = p_2(k)$ uit (verg. (17))

(19) $\qquad P(\underline{k} \le k;\ p_2) = \alpha/2.$

De vergelijkingen (18) en (19) hebben (in tegenstelling tot (8), blz.60) wel een oplossing (tenzij k = 0 of k = n; waarom?):

$$\sum_{j=k}^{n} \binom{n}{j} p_1^{\,j}(1 - p_1)^{n-j}$$

varieert continu met p_1 en kan iedere waarde tussen 0 en 1 aannemen, (hetzelfde geldt bij (19)). Het is echter niet mogelijk om p_1 en p_2 expliciet op te lossen. We vinden p_1 en p_2 uit een tabel van de binomiale verdeling (zie Statistische Tabellen) meestal m.b.v. interpolatie, omdat maar een beperkt aantal waarden van p in de tabellen voorkomt.

Voorbeeld: zij $n = 20$, $k = 7$ en $\alpha = 0{,}10$. We moeten dan p_1 zó kiezen dat

$$\sum_{j=7}^{20} \binom{20}{j} p_1^{\,j}(1 - p_1)^{20-j} = 0{,}05.$$

De tabel levert (na interpolatie tussen 0,15 en 0,20) $p_1 = 0{,}1717$.
Voor p_2 hebben we

$$P(\underline{k} \le 7;\, p_2) = \sum_{k=0}^{7} \binom{20}{k} p_2^{\,k}(1 - p_2)^{20-k} = 0{,}05$$

of (als men niet beschikt over tabellen voor waarden van $p > 1/2$)

$$\sum_{k=13}^{20} \binom{20}{k}(1 - p_2)^k p_2^{\,20-k} = 1 - \sum_{k=0}^{12} \binom{20}{k}(1 - p_2)^k p_2^{\,20-k} = 0{,}05.$$

We vinden nu (na interpolatie tussen $1 - p_2 = 0{,}40$ en $1 - p_2 = 0{,}45$) $1 - p_2 = 0{,}4392$, dus $p_2 = 0{,}5608$. Het gevraagde betrouwbaarheidsinterval is dus (afgerond) $0{,}17 < p < 0{,}56$. Dit interval bevat natuurlijk weer de puntschatting $\frac{k}{n} = \frac{7}{20} = 0{,}35$ (vergelijk formule (14), p. 63).

Voor grote waarden van n (en niet te kleine p) gebruiken we weer een benadering m.b.v. de normale verdeling. De relaties (18) en (19) gaan dan over in

$$(18') \qquad P(\underline{k} \ge k;\, p_1) \approx 1 - \Phi\left(\frac{k - np_1}{\sqrt{np_1(1 - p_1)}}\right) = \alpha/2$$

resp.

$$(19') \qquad P(\underline{k} \le k;\, p_2) \approx \Phi\left(\frac{k - np_2}{\sqrt{np_2(1 - p_2)}}\right) = \alpha/2,$$

zodat bij benadering $\frac{k - np_1}{\sqrt{np_1(1 - p_1)}} = u_{\alpha/2}$ en $\frac{k - np_2}{\sqrt{np_2(1 - p_2)}} = - u_{\alpha/2}$.

Blijkbaar zijn p_1 en p_2 (bij benadering) de wortels van de vierkants-vergelijking $(k - np)^2 = u^2_{\alpha/2}\, np(1 - p)$, of

(20) $$(n^2 + nu^2_{\alpha/2})p^2 - (2kn + nu^2_{\alpha/2})p + k^2 = 0.$$

We kunnen (20) ook schrijven als

(20') $$(p - \tilde{p})^2 = (u^2/n)\{\, \tilde{p}(1 - \tilde{p}) - \frac{u^2}{4(n + u^2)} \},$$

waarbij we u^2 hebben geschreven i.p.v. $u^2_{\alpha/2}$ en waarin $\tilde{p} = (p_1 + p_2)/2$, dus

$$\tilde{p} = \text{"} - \frac{b}{2a} \text{"} = \frac{k + u^2/2}{n + u^2}.$$

Voor niet te kleine n vinden we dan bij (verdere) benadering

(21) $$(p - \tilde{p})^2 \approx (u^2/n)\,\tilde{p}(1 - \tilde{p})$$

en dus

(22) $$p_{1,2} \approx \tilde{p} \pm u \sqrt{\tilde{p}(1 - \tilde{p})/n}.$$

Voor zéér grote n geldt bij nog verdere benadering

(23) $$p_{1,2} \approx \frac{k}{n} \pm u \sqrt{\frac{k}{n}(1 - \frac{k}{n})/n}.$$

Opgave: Bereken voor $n = 96$, $k = 18$ en $u = u_{\alpha/2} = 2$ de (benaderde) waarden van p_1 en p_2 : (a) direct uit (20), (b) volgens (22) en volgens (23). (De exacte waarden zijn $p_1 = 0{,}1132$, $p_2 = 0{,}2823$).

Opmerking 1: bij de eenzijdige toetsen kunnen we op analoge manier de eenzijdige betrouwbaarheidsintervallen (zie Opmerking, blz. 65) normaal benaderen. We vinden dan intervallen $[0,p_2')$ en $(p_1', 1]$, waarbij p_1' en p_2' resp. de kleinste en de grootste oplossing zijn van de vergelijking

$$(k - np')^2 = u_\alpha^2 np'(1 - p').$$

Het is duidelijk dat $p_2' < p_2$ en $p_1' > p_1$.

Opmerking 2: De betrouwbaarheidsintervallen (p_1,p_2), $[0,p_2')$ en $(p_1',1]$ voldoen aan Definitie 5'. Een bewijs hiervoor laten we achterwege.

Opgave 1: geef een betrouwbaarheidsinterval voor het aantal kogels dat de fabrikant uit Voorbeeld 1(blz.49) kan gebruiken, als hij in een steekproef van 100 stuks een gemiddelde diameter vindt van 5,043 (neem $\alpha = 0{,}01$).

Opgave 2: Hoeveel transistoren moet de radiomonteur inkopen als hij in een steekproef van 200 stuks 25 defecte exemplaren vindt ($\alpha = 0{,}001$) ? Zie Voorbeeld 2, blz. 49.

Appendix: Poisson-proces en Gamma-verdeling

In de laatste opgave van paragraaf 4.1, blz. 42, wordt gevraagd de volgende stelling te bewijzen:

Stelling 1: als $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ onafhankelijk zijn en exponentieel verdeeld met $E\underline{x}_j = 1/\lambda$ $(j = 1,2,\ldots,n)$, dan geldt voor $\underline{z}_n := \underline{x}_1 + \ldots + \underline{x}_n$

$$(1) \qquad f_{\underline{z}_n}(z) = \frac{\lambda^n}{(n-1)!} z^{n-1} e^{-\lambda z} \qquad (z > 0),$$

d.w.z. $\underline{z}_n$ heeft een Gamma-verdeling met parameters λ en n (zie blz. 17).

In deze appendix bewijzen we Stelling 1 met behulp van het Poisson-proces en geven we een wat algemenere definitie van de Gamma-verdeling.

Op blz. 24 hebben we het Poisson-proces ingevoerd op de volgende manier:

Definitie 1: het aantal gebeurtenissen $\underline{k}(t)$ dat optreedt in een interval $(0,t]$ is een Poisson-proces met *intensiteit* λ, als geldt

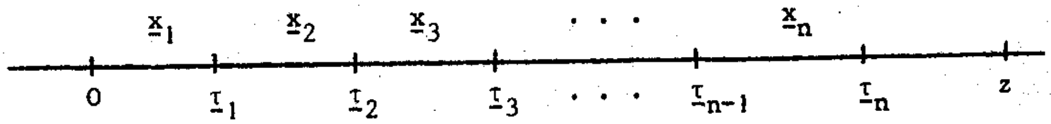
1. Voor iedere a en b met $0 < a < b$ heeft $\underline{k}(b) - \underline{k}(a)$ een Poisson-verdeling met verwachting $\lambda(b-a)$.
2. Voor iedere n en ieder n-tal $0 < t_1 < t_2 < \ldots < t_n$ zijn de grootheden $\underline{k}(t_1), \underline{k}(t_2) - \underline{k}(t_1), \ldots, \underline{k}(t_n) - \underline{k}(t_{n-1})$ onafhankelijk.

Uit deze definitie volgt dat de tijd $\underline{x}$ tussen het optreden van twee opeenvolgende gebeurtenissen exponentieel verdeeld is met $E\underline{x} = 1/\lambda$, immers: $P(\underline{x} > x) = P$ (geen gebeurtenis in interval van lengte x) $= P(\underline{k}(x) = 0) = e^{-\lambda x}$. Er geldt echter meer: het Poisson-proces kan ook worden gedefinieerd door

Definitie 2: een Poisson-proces met intensiteit λ is een proces, waarbij gebeurtenissen optreden op tijdstippen $\underline{\tau}_1, \underline{\tau}_2, \ldots$, zó dat de grootheden $\underline{\tau}_1, \underline{\tau}_2 - \underline{\tau}_1, \underline{\tau}_3 - \underline{\tau}_2, \ldots$ o.o. zijn en exponentieel verdeeld met verwachting $1/\lambda$.

Het Poisson-proces beschrijft praktijksituaties als het aankomen van klanten, het binnenkomen van schade-claims, de aanvragen van telefoongesprekken e.d. We bewijzen niet dat de Definities 1 en 2 equivalent zijn. Vergelijk ook Voorbeeld 2, blz. 21 en Opmerking op blz. 24.

We bewijzen nu Stelling 1 met behulp van deze definities (zie figuur):



We schrijven $\underline{x}_j = \underline{\tau}_j - \underline{\tau}_{j-1}$ $(j = 1,2, \ldots ; \underline{\tau}_0 = 0)$, dan zijn de beweringen

$$" \underline{x}_1 + \ldots + \underline{x}_n \le z " \quad \text{en} \quad " \underline{k}(z) \ge n "$$

equivalent.

Dus geldt, met $\underline{z}_n = \underline{x}_1 + \ldots + \underline{x}_n$ en voor $z > 0$

$$F_{\underline{z}_n}(z) = P(\underline{x}_1 + \ldots + \underline{x}_n \le z) = P(\underline{k}(z) \ge n) = \sum_{k=n}^{\infty} e^{-\lambda z} \frac{(\lambda z)^k}{k!}$$

(vergelijk Voorbeeld 6, blz. 17). Differentiatie levert

$$f_{\underline{z}_n}(z) = \sum_{k=n}^{\infty} e^{-\lambda z} \left(\frac{\lambda^k z^{k-1}}{(k-1)!} - \frac{\lambda^{k+1} z^k}{k!}\right) = \frac{\lambda^n}{(n-1)!} z^{n-1} e^{-\lambda z} .$$

Een derde methode om Stelling 1 te bewijzen is het gebruik van momenten-voortbrengende functies. We bewijzen formule (1) m.b.v. Stelling 7.1 op blz. 46 :

$$\int_0^{\infty} e^{zs} \frac{\lambda^n}{(n-1)!} z^{n-1} e^{-\lambda z} dz = \frac{\lambda^n}{(\lambda - s)^n} \int_0^{\infty} e^{-u} u^{n-1}/(n-1)! = \left(\frac{\lambda}{\lambda - s}\right)^n =$$

$$E\, e^{(\underline{x}_1 + \ldots + \underline{x}_n)s} = E\, e^{\underline{z}_n s}$$ (zie voorbeeld 3, blz. 46 en (7) blz. 45).

We geven nu een wat algemenere definitie van de Gamma-verdeling.

Definitie 3: voor reële $\beta > 0$ wordt de Gamma-functie, $\Gamma(\beta)$, gedefinieerd door

$$\Gamma(\beta) = \int_0^{\infty} u^{\beta-1} e^{-u} du .$$

Opgave 1: bewijs dat $\Gamma(\beta+1) = \beta\Gamma(\beta)$.

Opgave 2: bewijs met inductie dat $\Gamma(n+1) = n!$ $\quad (n = 0,1,2,\ldots)$.

Opgave 3: bewijs dat $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ (gebruik het feit dat $\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$ een kansdichtheid is; vergelijk Voorbeeld 8, blz. 17).

Definitie 4: $\underline{z}$ heeft een Gamma-verdeling met parameters λ en β, als

$$f_{\underline{z}}(z) = \gamma(z;\lambda;\beta) := \frac{\lambda^\beta}{\Gamma(\beta)}\, z^{\beta-1} e^{-\lambda z} \qquad (z > 0)\ ,$$

dus als

$$F_{\underline{z}}(z) = \Gamma(z;\lambda;\beta) := \frac{\lambda^\beta}{\Gamma(\beta)} \int_0^z x^{\beta-1} e^{-\lambda x} dx \qquad (z > 0)\ .$$

We schrijven "$\underline{z}$ is $\Gamma(\lambda;\beta)$-verdeeld".

Opgave 4: als $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_m$ o.o. zijn en resp. $\Gamma(\lambda;\beta_1), \ldots, \Gamma(\lambda;\beta_m)$ verdeeld, dan is

$$P(\underline{z}_1 + \ldots + \underline{z}_m \le z) = \Gamma(z;\lambda;\beta_1 + \ldots + \beta_m).$$

Opgave 5: Laat zien dat voor $t > 0$ geldt dat

$$\Gamma(z;\lambda;\beta) = \Gamma(z/t;\lambda t;\beta)\ .$$

Om $\Gamma(z;\lambda;\beta)$ te berekenen hebben we dus voldoende aan een tabel van $\Gamma(z;1;\beta)$ (λ is een schaalparameter). In de Statistische Tabellen worden om technische redenen alleen de waarden van $\frac{\lambda x}{r}$ gegeven, waarvoor $\Gamma(\frac{\lambda x}{r}; r; r) = 1-\alpha$ voor $\alpha = 0{,}100$, $\alpha = 0{,}050$, $\alpha = 0{,}025$, $\alpha = 0{,}010$, $\alpha = 0{,}005$ en $\alpha = 0{,}001$ en voor waarden van r met $\frac{1}{\sqrt{r}} = 0{,}00\ (0{,}05)\ 1{,}40$.

Een andere mogelijkheid om $\Gamma(z;\lambda;\beta)$ op te zoeken is een tabel van de χ^2-verdeling. Een grootheid $\underline{\chi}^2_m$ heeft een chi-kwadraat-verdeling met m "*vrijheidsgraden*" (een χ^2_m-verdeling) als $\underline{\chi}^2_m = \underline{u}_1^2 + \ldots + \underline{u}_m^2$, waarbij $\underline{u}_1, \ldots, \underline{u}_m$ o.o.

zijn en N(0,1)-verdeeld. Voor de verdeling van $\underline{u}_1^2$ geldt

$$F_{\underline{u}_1^2}(x) = P(\underline{u}_1^2 \leq x) = P(-\sqrt{x} \leq \underline{u}_1 \leq \sqrt{x}) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-y^2/2} dy =$$

$$= \sqrt{\frac{2}{\pi}} \int_0^{\sqrt{x}} e^{-y^2/2} dy ,$$

zodat (differentiëren) $f_{\underline{u}_1^2}(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-x/2} = \gamma(x;\frac{1}{2},\frac{1}{2})$ en dus (zie Definitie 4) $F_{\underline{u}_1^2}(x) = \Gamma(x;\frac{1}{2};\frac{1}{2})$, zodat (vergelijk Opgave 4)

$$P(\underline{\chi}_m^2 \leq x) = \Gamma(x;\tfrac{1}{2};\tfrac{m}{2}) .$$

In **de Statistische Tabellen** zijn de waarden van x getabelleerd waarvoor geldt $P(\underline{\chi}_\nu^2 \leq x) = \gamma$ voor een aantal waarden van γ en voor $\nu = 1,2,\ldots,30$ (voor grotere waarden van ν gebruiken we de centrale limietstelling).

In de praktijk wordt de Gamma-verdeling gebruikt om (bijvoorbeeld) de vraag naar een bepaald artikel te beschrijven. Als deze vraag $\underline{z}$ per gegeven periode $\Gamma(\lambda;\beta)$-verdeeld is, dan is de vraag $\underline{z}_t$ over een t maal zo lange periode (in dit model) $\Gamma(\lambda;t\beta)$-verdeeld. Voor $\underline{z}$ en $\underline{z}_t$ geldt nu (ga na!) $E\,\underline{z} = \beta/\lambda$, var $\underline{z} = \beta/\lambda^2$ en $E\,\underline{z}_t = \beta t/\lambda$ en var $\underline{z}_t = \beta t/\lambda^2$ (vergelijk Opgave 5). In dit model geldt dat de gevraagde hoeveelheden in disjuncte perioden onafhankelijk zijn en $\Gamma(\lambda)$-verdeeld met een β-parameter die evenredig is met de lengte van de periode: als $t = t_1 + .. + t_m$ en als $\underline{z}_t = \underline{z}_{t_1} + \ldots + \underline{z}_{t_m}$, waarbij de $\underline{z}_{t_j}$ onafhankelijk zijn en $\Gamma(\lambda,\beta t_j)$-verdeeld, dan heeft $\underline{z}_t$ een $\Gamma(\lambda,\beta t)$-verdeling.
<u>Voorbeeld</u>: de vraag per dag, $\underline{x}$, naar benzine aan een gegeven tankstation is Γ-verdeeld met $E\,\underline{x} = 4500$ liter en $\sigma(\underline{x}) = 2500$ liter. De pomphouder wil benzine voor één week in voorraad nemen. Hoeveel moet hij nemen om een kans van 0,95 te hebben dat hij aan de vraag kan voldoen?

<u>Oplossing</u>: $\underline{x}$ is $\Gamma(\lambda;\beta)$-verdeeld met $\beta/\lambda = 4500$ en $\sqrt{\beta}/\lambda = 2500$, zodat $\beta = 3,24$ en $\lambda = 0,000720$. De vraag per week, $\underline{y}$, is $\Gamma(\lambda;r)$-verdeeld met $r = 7\beta = 22,68$. Als we de gezochte voorraad V noemen, dan volgt V uit de vergelijking

$$0,95 = P(\underline{y} \leq V) = \Gamma(V;\lambda;r) .$$

Om de tabel van de Γ-verdeling in de Statistische Tabellen te kunnen gebruiken transformeren we dit tot $\Gamma(\frac{\lambda V}{r};r;r)$ met $\frac{1}{\sqrt{r}} = \frac{1}{\sqrt{22,68}} = 0,210$. De tabel levert (na lineaire interpolatie) $\lambda V/r = 1,37$ of $V = 43155$ (liter).