MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 155

Cursus

Toegepaste Statistiek II

door

Ph. van Elteren en J. Kriens

I. Algemene opmerkingen over waarschijnlijkheidsverdelingen

1954

Hoofdstuk 1

Algemene opmerkingen over waarschijnlijkheids verdelingen.

1.1 Inleiding

In het begin van het eerste deel van deze cursus hebben wij ons bezig gehouden met een pot, die rode en groene erwten bevatte. Laat N het totale aantal erwten in een dergelijke pot zijn, en k het aantal rode erwten, dan hebben wij methoden geleerd om de fractie $p = {}^{k}/_{N}$ der rode erwten in de pot te schatten, door het nemen van steekproeven. Het bleek, dat als men de erwten afzonderlijk trekt, na iedere trekking de getrokken erwt weer in de pot werpt en vervolgens goed schudt of roert (om te voldoen aan de eisen van gelijkwaardigheid en onafhankelijkheid der trekkingen, die in het eerste deel geformuleerd zijn), de fractie rode erwten in een lange reeks trekkingen meestal weinig van p afwijkt. Zo'n reeks trekkingen wordt een steekproef met teruglegging uit de pot genoemd. Het aantal erwten n in een steekproef getrokken, heet de omvang van de steekproef. Als er x rode erwten in een steekproef getrokken zijn, zeggen wij dat de frequentie van de rode erwten in de steekproef x is; men verstaat dan onder het frequentiequotient [1] van de rode erwten de breuk $\frac{x}{n}$. Wij kunnen dus ook zeggen, dat het fq van de rode erwten in een grote steekproef met terugleggingmeestal weinig van de fractie p der rode erwten in de pot zal afwijken. Men drukt dit ook wel uit door te zeggen, dat de kans of de waarschijnlijkheid [2] van het trekken van een rode erwt uit een pot met een fractie p aan rode erwten gelijk is aan p. Het verschijnsel, dat het fq van een rode erwt in lange reeksen waarnemingen ongeveer gelijk is aan p, wordt de wet der grote getallen genoemd.

Een consequentie van de wet der grote getallen is, dat de fqn van de rode erwten in grote steekproeven ongeveer gelijk zullen zijn. Indien men omgekeerd in lange reeksen waarnemingen van gelijksoortige verschijnselen (grootheden, voorwerpen etc.) steeds ongeveer gelijke fqn voor verschijnselen met een bepaald kenmerk K vindt, dan zal men aan het optreden van dit kenmerk ook een wh toekennen, welke wh benaderd wordt door de voor K gevonden fqn. Bij de erwten was de kans op het trekken van een

---

1) Afkorting: fq, meervoud:fqn.
2) Afkorting: wh, meervoud:whn.

rode erwt (onder bepaalde voorwaarden) gelijk aan de fractie rode erwten in de pot. Bij andere reeksen waarnemingen waarvoor de wet van de grote getallen blijkt te gelden, denkt men ook dikwijls aan een analogon van de pot met erwten, een pot, die alle mogelijke waarnemingen bevat betreffende hetzelfde verschijnsel die onder dezelfde omstandigheden verricht zouden kunnen worden. Deze uiteraard denkbeeldige pot wordt <u>populatie</u>, <u>universum</u> of <u>collectie</u> genoemd.

<u>Voorbeeld</u>:Lange reeksen worpen met een dobbelsteen voldoen in het algemeen aan de wet der grote getallen. Als men in dergelijke reeksen constateert dat de fqn van de 6 mogelijke aantallen ogen ongeveer gelijk zijn, zal men de kans op het werpen van ieder mogelijk aantal ogen gelijk stellen aan $\frac{1}{6}$. Een dobbelsteen waarvoor dit exact geldt heet zuiver. Indien de fqn van de aantallen ogen in lange reeksen worpen duidelijk verschillen, zullen wij de kansen op deze aantallen niet gelijk stellen. Een dobbelsteen, waarvoor deze kansen niet gelijk zijn, heet <u>onzuiver</u>. In beide gevallen kunnen wij de worpen beschouwen als trekkingen uit een populatie met ballen genummerd van 1 t/m 6. Bij een <u>zuivere</u> dobbelsteen zal deze populatie gelijke aantallen van deze ballen bevatten, bij een <u>onzuivere</u> dobbelsteen ongelijke aantallen.

1.2 <u>Eigenschappen van whn</u>

Wij hebben in het eerste deel van deze cursus gezien, dat men aan de whn bepaalde eigenschappen moet toekennen, welke corresponderen met die van fqn. Laten A, B, C etc. bepaalde gebeurtenissen[3] zijn, en P[A], P[B], P[C] de kansen op het optreden van die gebeurtenissen, dan kunnen die eigenschappen als volgt weergegeven worden:

P[A] = 0 als A onmogelijk is

P[A] = 1 als A zeker is

Als A en B elkaar uitsluiten:

(1.2.1) P[A of B] = P[A] + P[B].

Als A en B onderling <u>onafhankelijke</u> gebeurtenissen zijn (d.w.z. als het al of niet optreden van B niet afhangt van het al of niet optreden van A):

(1.2.2) P[A én B] = P[A].P[B].

---

3) Wij gebruiken hier het woord "gebeurtenis" tevens in de zin van "het optreden van een bepaald kenmerk" en "de uikomst van een experiment".

1.3 Binomiale Waarschijnlijkheidsverdeling

Indien wij uit de pot met erwten reeksen steekproeven van b.v. 25 exemplaren ieder nemen, kunnen daarbij alle mogelijkheden van 0 tot 25 rode erwten optreden. Als x het aantal rode erwten in een steekproef voorstelt, blijkt dat het fq van iedere mogelijke waarde van x in lange reeksen steekproeven weer redelijk constant is. Wij geven in figuur 1.1. de frequenties van de verschillende waarden van x in twee reeksen van 100 steekproeven van 25 waarnemingen uit de pot met erwten. Hier blijkt reeds een redelijke overeenstemming tussen beide reeksen te bestaan.

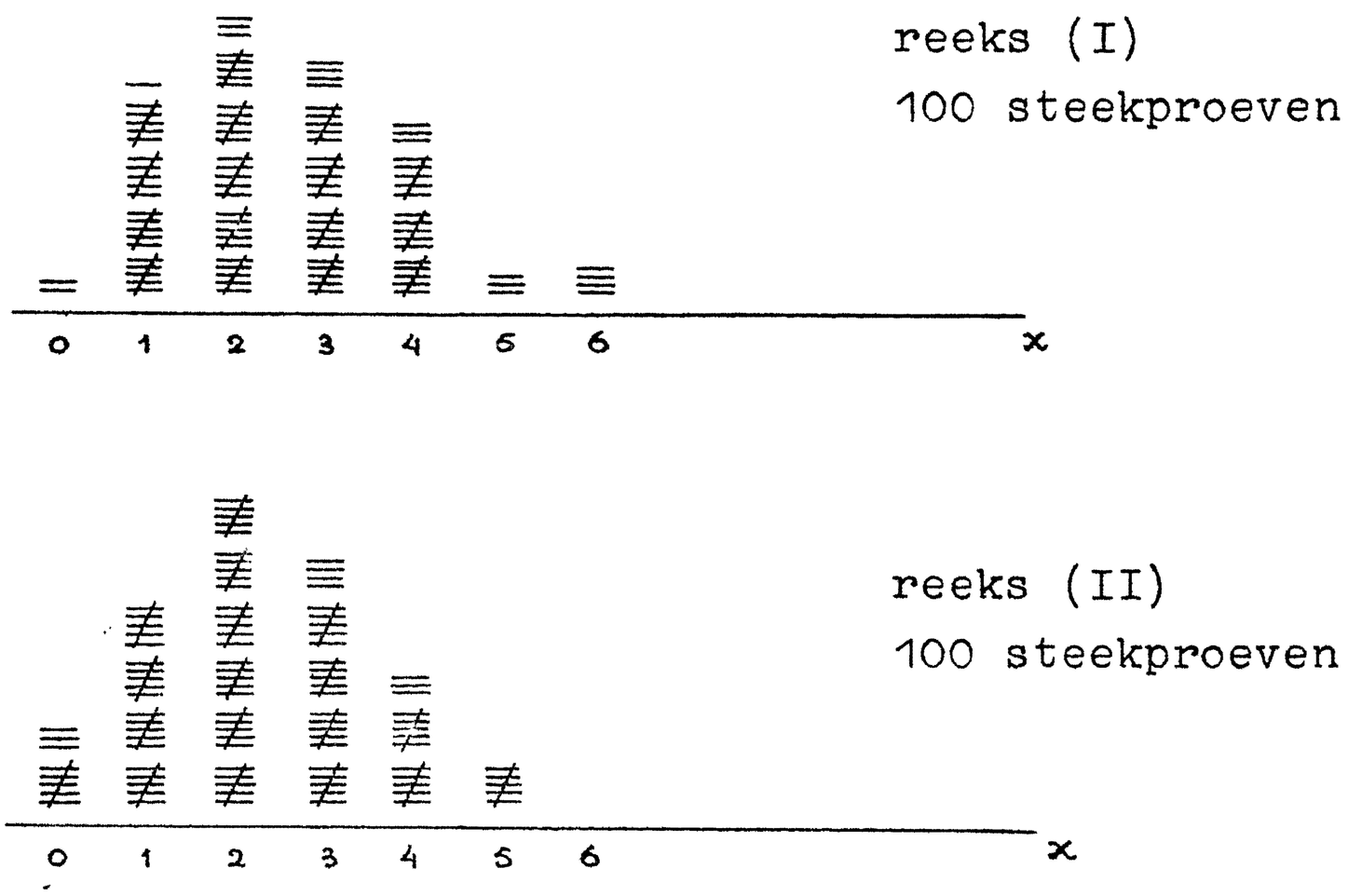


Fig. 1.1. Turfstaatjes van 2 reeksen van 100 waarnemingen van x

Steekproeven met 2 rode erwten komen in beide reeksen het meest voor, daarna achtereenvolgens die met 3, 1 en 4 rode erwten. Duidelijke verschillen treden alleen op in de "staarten" van de figuren.

Wij hebben hier te doen met trekkingen uit een populatie waarvan de elementen als kenmerken de nummers 0,1,2,.....,25 dragen. Als wij dit nummer voorstellen door $\underline{x}$ is er een zekere kans, dat $\underline{x}$ gelijk is aan 0, aan te duiden als $P[\underline{x}=0]$, een kans dat $\underline{x}$ gelijk is aan 1: $P[\underline{x}=1]$ etc. De kansen $P[\underline{x}=0]$, $P[\underline{x}=1]$, . . . , $P[\underline{x}=25]$ vormen tezamen een zogenaamde waarschijnlijkheidsverdeling.[4)]

Een grootheid, die een wh-verdeling bezit, noemen wij een stochastische variabele. Wij zullen stochastische variabel gewone variabelen onderscheiden door het symbool waarmee ze den aangeduid te onderstrepen.

4) Afkorting: wh-verdeling, ook wel kortweg: verdeling.

De wh-verdeling van de aantallen rode erwten in steekproeven van n waarnemingen uit een pot, die een fractie p aan rode erwten bevat, is een zogenaamde binomiale verdeling. Voor lage waarden van n kunnen wij deze verdeling gemakkelijk afleiden. Stelt b.v. RGGRG ..... een trekkingsreeks voor, waarin achtereenvolgens een rode, een groene, nog een groene, een rode, een groene enz. erwt getrokken is, dan kunnen wij, als de trekkingen onderling onafhankelijk zijn, uit par. 1.2 gemakkelijk de volgende kansen afleiden:

n = 1

$P[R]=p$, $P[G]=1-p$.
Geeft men $1-p$ aan met q, dan is dus:
$P[\underline{x}=1]=p$, $P[\underline{x}=0]=q$.

n = 2

$P[RR] = P[R].P[R] = p^2 \longrightarrow P[\underline{x}=2] = p^2$

$P[RG] = P[R].P[G] = pq$ } $P[\underline{x}=1] = P[RG] + P[GR] =$

$P[GR] = P[G].P[R] = qp$ } $= 2pq$

$P[GG] = P[G].P[G] = q^2 \longrightarrow P[\underline{x}=0] = q^2$.

n = 3

$P[RRR] = p^3$	$P[\underline{x}=3] = p^3$
$P[RRG] = P[RGR] = P[GRR] = p^2q$	$P[\underline{x}=2] = 3p^2q$
$P[RGG] = P[GRG] = P[GGR] = pq^2$	$P[\underline{x}=1] = 3pq^2$
$P[GGG] = q^3$	$P[\underline{x}=0] = q^3$.

Op deze wijze kan men afleiden, dat voor willekeurige waarden van n geldt:

(1.3.1) $$P[\underline{x}=x] = \binom{n}{x} p^x q^{n-x},$$

waarin $\binom{n}{x}$ het aantal verschillende manieren voorstelt om uit n voorwerpen x te kiezen. Men kan bewijzen, dat geldt:

$$\binom{n}{x} = \frac{n(n-1)\ldots(n-x+2)(n-x+1)}{x(x-1)\ldots 2 \cdot 1}.$$

De trekkingsreeksen I en II van figuur 1.1 waren afkomstig uit een pot met een fractie $p \approx 0{,}08$ aan rode erwten. Wij hebben in tabel 1.I de theoretische whn van de binomiale verdeling bij n = 25 vergeleken met de fqn der trekkingsreeksen.

Tabel 1.I.

Theoretische whn (P) voor een binomiale verdeling met p = 0,08 en n = 25 vergeleken met de fqn gevonden bij de reeksen I en II van fig. 1.1.

x	Theor. P (in %)	Reeks I fq(in %)	Reeks II fq(in %)
0	12,4	2	8
1	27,0	21	20
2	28,2	28	30
3	18,8	24	24
4	9,0	18	13
5	3,3	3	5
6	1,0	4	0
7	0,2	0	0
8 en hoger	0,1	0	0

De overeenkomst tussen de waargenomen fqn en de theorethische whn is niet bijzonder goed. De hoge waarden van x komen in de waarnemingsreeksen meer voor, dan overeenkomt met de wh-verdeling. Met statistische methoden, waarop hier niet nader zal worden ingegaan, kan men onderzoeken of de afwijkingen systematisch zijn, dan wel aan het toeval toegeschreven kunnen worden. Het is mogelijk, dat systematische afwijkingen ontstaan zijn, doordat tussen twee trekkingen niet voldoende geroerd is, waardoor de kans op het trekken van een rode erwt iets groter geworden is dan de fractie rode erwten in de pot.

Opgave:

1.3.a Bereken de Binomiale wh-verdelingen voor n = 4 en n = 5 uitgedrukt in p en q volgens de bovenbeschreven methode en vergelijk de resultaten met formule 1.3.1.

1.4 Discrete wh-verdelingen in het algemeen

De in par. 1.3 beschreven Binomiale wh-verdeling was een voorbeeld van een zogenaamde discrete wh-verdeling. De stochastische variabele ξ neemt hier met bepaalde whn discrete waarden aan. Een discrete wh-verdeling kan worden voorgesteld door een diagram, waarin horizontaal de grootheid x is uitgezet en door verticale strepen is aangegeven hoe groot de wh is van het optreden van bepaalde waarden van x. In fig. 1.2 hebben wij de wh-verdeling behorende bij een zuivere dobbelsteen (ξ = aantal

ogen) en de binomiale verdeling voor n = 6, p= 0,4 weergegeven.

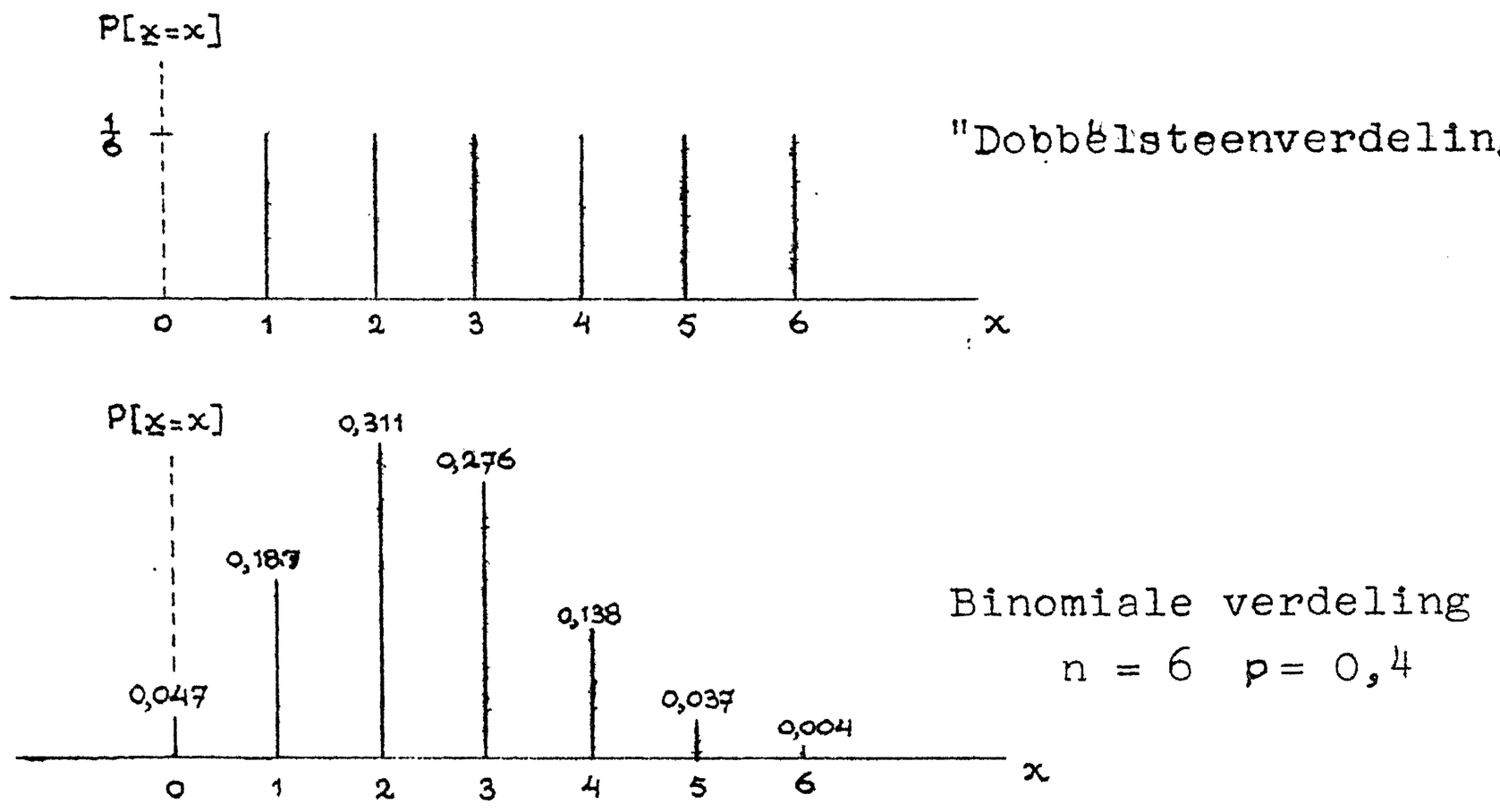


Fig. 1.2 Twee discrete verdelingen.

Een andere methode om een wh-verdeling te karakteriseren is het opgeven van de zogenaamde verdelingsfunctie. De verdelingsfunctie geeft voor iedere waarde van x de kans, dat de stochastische variabele $\underline{x}$ kleiner dan of gelijk aan x is. Zij wordt gewoonlijk voorgesteld door F(x). Er geldt dus per definitie:

(1.4.1) $$F(x) = P[\underline{x} \leqq x].$$

Deze functie is een zogenaamde <u>monotoon niet afnemende</u> functie: als x stijgt kan de waarde van F(x) stijgen of gelijk blijven, maar niet dalen. Bij een discrete wh-verdeling gaat deze stijging in sprongen. Bij ieder der discrete waarden die door $\underline{x}$ kan worden aangenomen, maakt F(x) een sprong ter grootte van de kans, dat $\underline{x}$ die waarde aanneemt. Een dergelijke functie krijgt dus de gedaante van een trap, die begint bij F(x)=0 en eindigt bij F(x)=1.(Zie figuur 1.3; aan het eind van iedere trede zijn pijlpunten geplaatst om aan te geven dat de treden lopen van $-\infty$ tot aan x=+1, vanaf x=+1 tot aan x=+2 etc. Als F(x) gedefinieerd zou zijn door $P[\underline{x}<x]$ zouden de treden tot en met de waarden x=+1, x=+2 etc. lopen, en zouden de pijlpunten aan de linker uiteinden der treden staan).

Het nut van de verdelingsfuncties is, dat men daarmee gemakkelijk de kans kan afleiden dat $\underline{x}$ in een bepaald interval ligt. Beschikt men b.v. bij de Binomiale verdeling (p=0,4 n=25) over een tabel

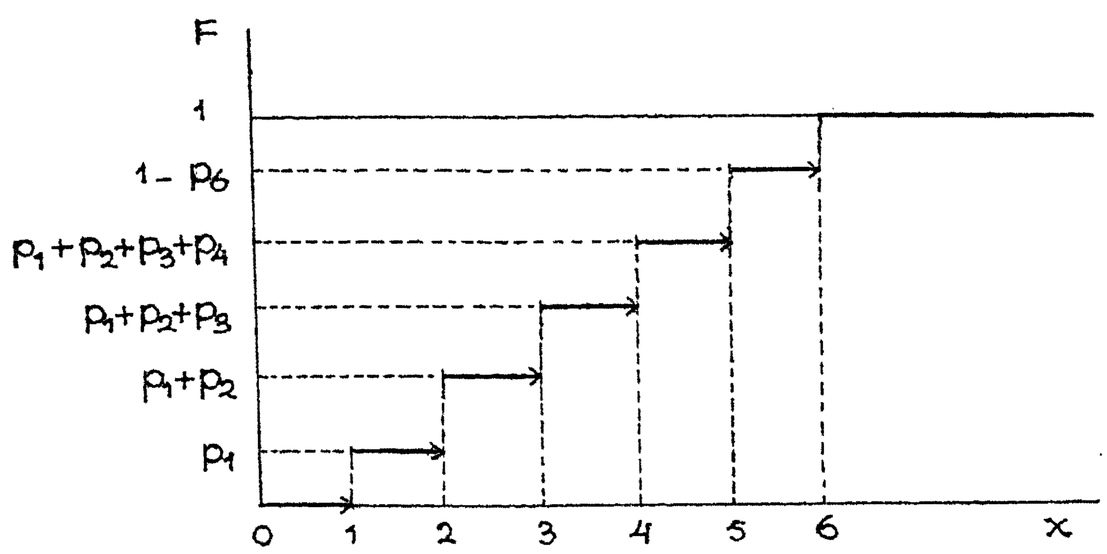


Fig. 1.3 F(x) voor een dobbelsteen.
(p_i is de kans, dat men met de dobbelsteen i ogen werpt)

van $P[\underline{x}=x]$ en wil men weten waaraan de kans $P[10 \leqq \underline{x} < 22]$ gelijk is, dan zal men de kansen $P[\underline{x}=10]$, $P[\underline{x}=11]$, ..., t/m $P[\underline{x}=21]$ moeten sommeren. Heeft men echter de beschikking over een tabel van $F(x)= P[\underline{x} \leqq x]$ dan vindt men veel gemakkelijker: $P[10 \leqq \underline{x} < 22] = F(21) - F(9)$.

Opgave:

1.4.a Teken het strependiagram (Fig.1.2) en de verdelingsfunctie $F(x)$ voor een binomiale verdeling met $n = 5$ en $p = 0,3$ (Vergelijk opgave 1.3.a zie Fig. 1.3)

1.5 Continue wh-verdelingen

Onder een continue wh-verdeling verstaan wij een verdeling waarbij F(x) niet een trapfunctie is, doch een vloeiende kromme. Dit zal uiteraard ook een kromme moeten zijn die begint op het niveau $F=0$ en geleidelijk stijgt naar het niveau $F=1$. In het bovenste gedeelte van figuur 1.4 hebben wij een willekeurige continue wh-verdeling geschetst. Bij continue wh-verdelingen is de kans dat de stochastische variabele een bepaalde waarde precies aanneemt, steeds gelijk aan 0. Men beschouwt hier alleen de kansen dat $\underline{x}$ in bepaalde intervallen ligt. Dergelijke kansen kan men, evenals bij de discrete wh-verdelingen, direct afleiden uit tabellen of grafieken van de verdelingsfunctie; de kans dat $\underline{x}$ valt in het interval (x_1, x_2) wordt dan gegeven door $F(x_2) - F(x_1)$; het doet er uiteraard niet meer toe of de eindpunten al dan niet zelf bij het interval behoren.

Een andere functie, dikwijls gebruikt om continue wh-verdelingen te karakteriseren, is de verdelingsdichtheid

(Symbool f(x)). Dit is een functie, die voor iedere x de helling van de kromme F(x) aangeeft. Wij kunnen deze functie dus vinden, door in ieder punt van de kromme F(x), de raaklijn aan deze kromme te tekenen, en daarna de tangens van de hoek te bepalen, die de raaklijn maakt met de x-as. In de terminologie van de differentiaalrekening heet f(x) de afgeleide van F(x).

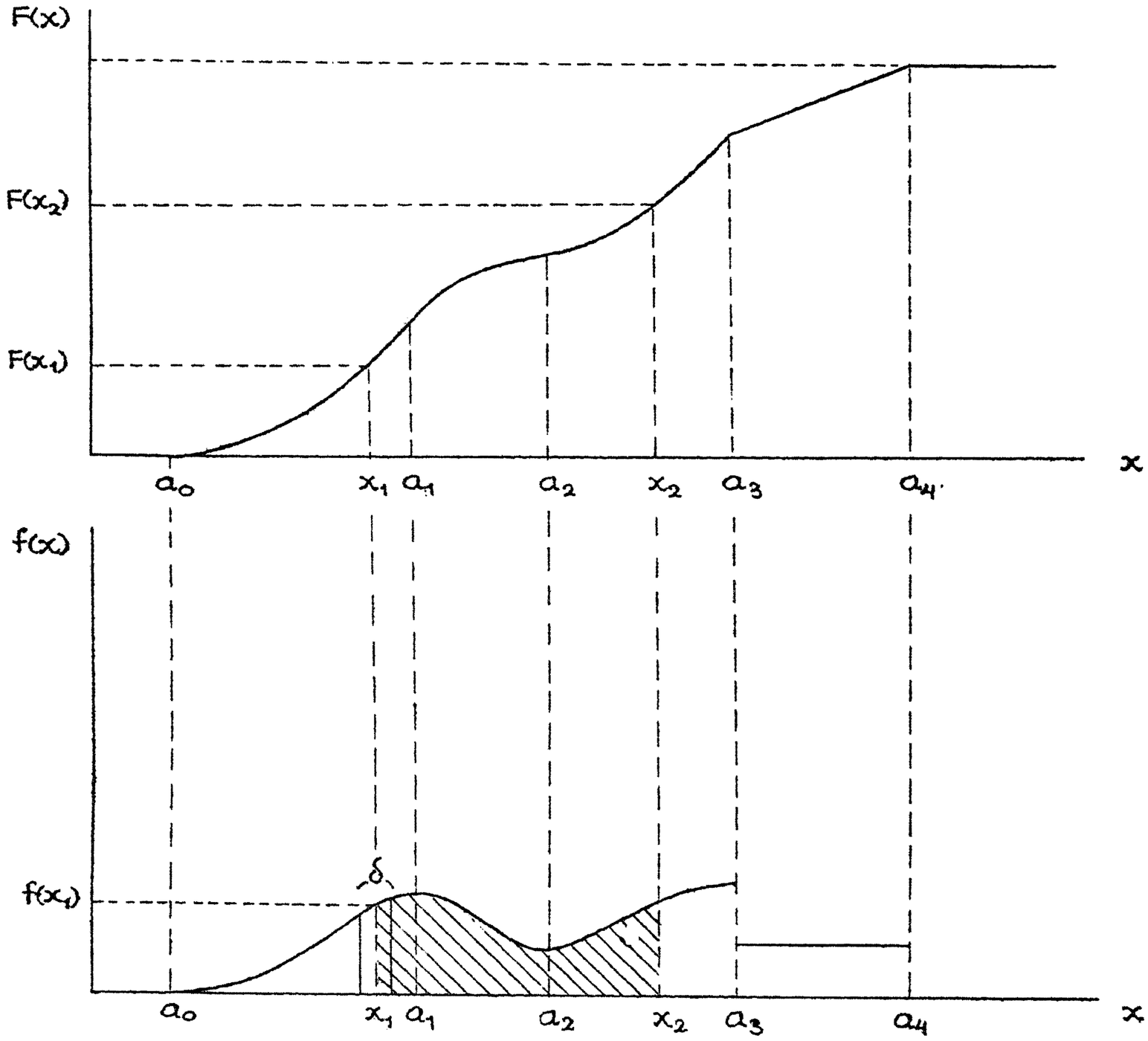


Fig. 1.4 Een continue wh-verdeling.

In het onderste gedeelte van Fig. 1.4 hebben wij de afgeleide f(x) geschetst van de verdelingsfunctie F(x), die in het bovenste gedeelte gegeven is. De kromme f(x) vertoont een stijgend verloop, als de kromme F(x) naar beneden gekromd is (in de figuur tussen a_0 en a_1 en tussen a_2 en a_3), een dalend verloop als F(x) naar boven gekromd is (in de figuur tussen a_1 en a_2) en zij blijft constant in de intervallen waar F(x) recht is (tussen a_3 en a_4). Karakteristieke punten in fig. 1.4 zijn verder a_1 en a_2, buigpunten van F(x) die corresponderen met een maximum of minimum bij f(x), en de punten a_3 en a_4 waar F(x) een knik maakt, waardoor f(x) een sprong vertoont.

De wh dat $\underline{x}$ een waarde tussen x_1 en x_2 aanneemt, kan ook uit het onderste gedeelte van fig. 1.4 worden afgeleid. Dit is de oppervlakte gelegen tussen de x-as en de lijn y=f(x), begrensd door de verticale lijnen door x_1 en x_2. Dit oppervlak wordt aan-

geduid met het symbool:

$$\int_{x_1}^{x_2} f(x)dx$$

te lezen als "de integraal van $f(x)$ van x_1 tot x_2". Hier geldt uiteraard:

$$\int_{x_1}^{x_2} f(x)\,dx = F(x_2) - F(x_1)$$

daar links en rechts dezelfde kans wordt voorgesteld.

De wh.dat $\underline{x}$ een waarde aanneemt, die ligt in een klein intervalletje ter lengte δ om x_1 is blijkens de figuur bij benadering gelijk aan $f(x_1).\delta$. Hierin schuilt de betekenis van de term wh- dichtheid. Wanneer $f(x)$ groot is, is"meer wh. in de omgeving van het punt x geconcentreerd" dan wanneer $f(x)$ klein is.Een grafiek van $f(x)$ geeft ons daardoor een beter beeld van het verloop der wh- verdeling dan een grafiek van $F(x)$. Indien men echter berekeningen moet uitvoeren gebaseerd op een bepaalde wh- verdeling, heeft men meer aan tabellen van $F(x)$ dan aan die van $f(x)$.

Om de functie $f(x)$ af te leiden uit de functie $F(x)$ moet men een wiskundige bewerking uitvoeren, welke als <u>differentiëren</u> bekend staat. Deze bewerking is niet altijd uitvoerbaar, ook niet bij alle krommen die in de wiskunde continu genoemd worden. In de wh-rekening gebruikt men echter alleen verdelingen, die hetzij discreet zijn, hetzij zodanig continu, dat de afgeleide functie $f(x)$ bestaat; men noemt de verdelingen waarvoor $f(x)$ bestaat dan kortweg continu.

Om de functie $F(x)$ af te leiden uit $f(x)$ moet men een andere bewerking uit voeren, die <u>integreren</u> genoemd wordt. Dit integreren kan in sommige gevallen volgens eenvoudige rekenregels gebeuren. In vele gevallen moet men echter zijn toevlucht nemen tot een of andere benaderingsmethode om het oppervlak onder de kromme $f(x)$ te bepalen; benaderingsmethoden, waarmee overigens, als $f(x)$ exact bekend is, iedere gewenste nauwkeurigheid bereikt kan worden.

<u>Opgaven:</u>

1.5a. Teken de verdelingsfunctie $F(x)$ en de verdelingsdichtheid $f(x)$ voor een wh-verdeling, die als volgt gedefiniëerd is:

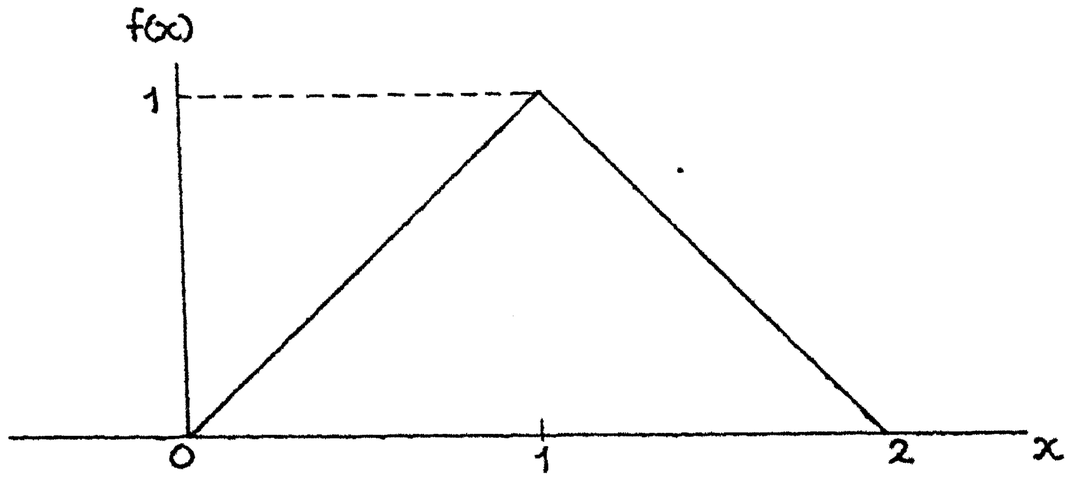
$F(x) = 0$ voor $x < 0$

$F(x) = x$ voor $0 \leqq x \leqq 1$

$F(x) = 1$ voor $x > 1$.

(Homogene of rechthoekige verdeling).

1.5b. De verdelingsdichtheid van een wh-verdeling heeft de volgende vorm:



Geef een formule voor $f(x)$ en een voor $F(x)$:

1) als $0 \leqq x \leqq 1$,

2) als $1 \leqq x \leqq 2$.

(Driehoeksverdeling)

Bereken hieruit die waarden van x waarvoor geldt:

3) $P[\underline{x} > x] = 0{,}05$,

4) $P[-x \leqq \underline{x} \leqq +x] = 0{,}98$.

1.6 Toepassing van Continue wh-verdelingen

Continue wh-verdelingen gebruikt men voor de beschrijving van de variatie van grootheden, die in een bepaald interval (eventueel van $-\infty$ tot $+\infty$) alle mogelijke waarden aan kunnen nemen. Een dergelijke grootheid is b.v. de lengte van mensen (beschouwd over de populatie van àlle mensen of van een bepaalde groep mensen). Vele physische en chemische grootheden vertonen een stochastisch karakter, tengevolge van schommelingen in de omstandigheden welke men niet in zijn macht heeft. Men pleegt de hierdoor optredende verschillen meetfouten te noemen. Een waarneming van een dergelijke grootheid kan men dan beschouwen als een trekking uit de populatie van alle mogelijke waarnemingen, die men van dezelfde grootheid zou kunnen verrichten als alle controleerbare en relevante omstandigheden gelijk gehouden worden.

De steekproeven, die men aldus verkrijgt, vertonen toch een discreet karakter, omdat ieder meetresultaat uit een beperkt aantal cijfers bestaat. Indien wij opgeven, dat de lengte van een persoon 1,79 m is, bedoelen wij in feite, de lengte van deze persoon ligt tussen 1,785 en 1,795 m (afgezien van kleine lengte variaties in de tijd). Men geeft dus één waarde aan, maar bedoelt een interval van mogelijke waarden, waarvan de opgegeven waarde het midden is.

Indien wij een steekproef willen vergelijken met een bepaalde continue wh-verdeling, dient men daarom ook na te gaan op welke intervallen de steekproefwaarden in feite betrekking hebben. De fqn van de waarnemingen in die intervallen kunnen dan vergeleken worden met de overeenkomstige whn in de populatie.

Een goed beeld van een steekproef uit een continue verdeling krijgen wij door het tekenen van een"histogram". Men verdeelt daarbij de x-as in intervallen, en tekent boven ieder interval een rechthoek die evenredig is met het fq van de steekproefwaarden die binnen dat interval liggen. Indien men nu deze steekproef wil vergelijken met een bepaalde wh-verdeling kan men op twee manieren te werk gaan. Men kan boven ieder interval een rechthoek tekenen, waarvan de oppervlakte evenredig is met de wh. dat de beschouwde stochastische variabele een waarde aanneemt, die in dat interval ligt; men verkrijgt dan het histogram van de verdeling. Men kan ook de verdelingsdichtheidscurve f(x) in de figuur tekenen. Men zorge er voor, dat de oppervlakte van het histogram der verdeling resp.de oppervlakte onder de lijn f(x) , gelijk is aan de oppervlakte van het histogram der steekproef.

In figuur 1.5 hebben wij een dergelijk histogram gegeven voor de schedelbreedten in Engelse duimen van 1000 studenten uit Cambridge[5].

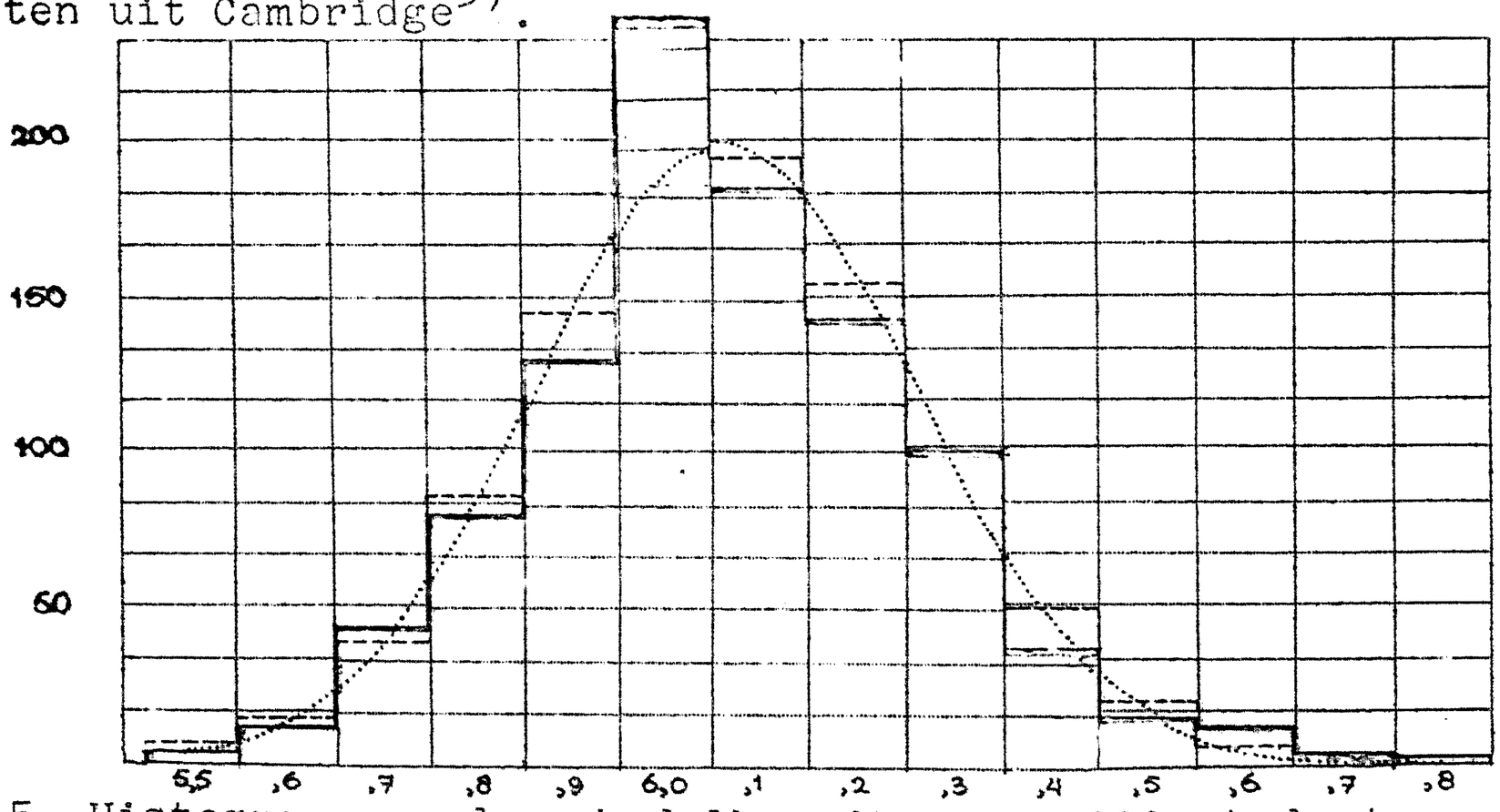


Fig 1.5 Histogram van de schedelbreedte van 1000 studenten uit Cambridge.

—— = Histogram der gegevens.

---- = Histogram van de aangepaste normale verdeling.

—— = Verdelingsdichtheidscurve der aangepaste normale verdeling.

5) G.U.Yule and M.G.Kendall, An Introduction to the Theory of statistics, London 1946, 13th edition, fig.6.2 pg 91.

In dezelfde figuur zijn ter vergelijking de interval-whn èn de verdelingsdichtheidscurve getekend van een wh-verdeling van een veel voorkomend type, de zogenaamde normale verdeling, die zo goed mogelijk bij de experimentele resultaten is aangepast.Op deze normale verdeling en de methoden van aanpassing zullen wij nog nader terugkomen. (Zie Hoofdstuk 3)

Het is uiteraard ook mogelijk om histogrammen te vervaardigen van steekproeven uit discrete verdelingen. Men verdeelt de x-as dan in intervallen, die één of meer van de mogelijke steekproefwaarden bevatten en plaatst op deze intervallen rechthoeken waarvan de oppervlakte evenredig is met fqn der steekproefwaarden in de intervallen.Evenzo kan men een histogram maken van een discrete wh-verdeling zelf. Het is zodoende mogelijk om een histogram van een discrete steekproef te vergelijken met dat van een discrete wh-verdeling en ook met het histogram van een continue wh-verdeling, mits steeds dezelfde indeling in intervallen wordt gebruikt.

Met behulp van histogrammen kan men eveneens wh-verdelingen onderling vergelijken.In de praktijk is het vaak van belang om een discrete wh-verdeling te vergelijken met een aangepaste continue verdeling. Dit is bijvoorbeeld gedaan in figuur 1.6voor een binomiale en een normale verdeling. Wij zien hier in, dat de kans $P[\underline{x} = x]$ voor de binomiale verdeling vergeleken wordt met de kans $P[x-\frac{1}{2} \leqq \underline{x} \leqq x+\frac{1}{2}]$ van de normale verdeling.

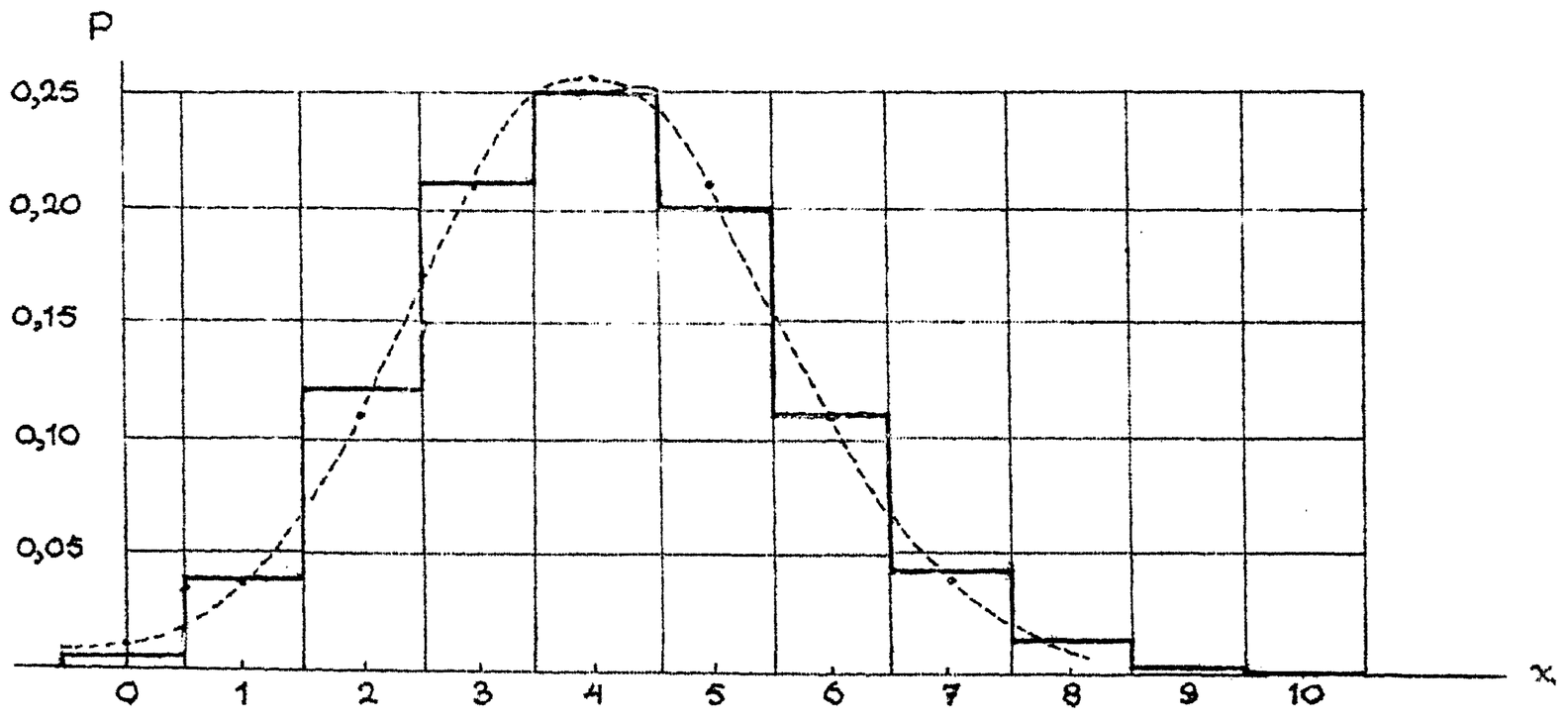


Fig.1.6

———————— Histogram binomiale verdeling

---------- f(x) van aangepaste normale verdeling.

Er kan bewezen worden dat, als n toeneemt, die beide kansen elkaar steeds dichter naderen. Men drukt dit uit door te zeggen, dat de Binomiale verdeling voor $n \to \infty$ (bij vaste p) asymptotisch overgaat in een normale verdeling. Dit gaat sneller naarmate p dichter in de buurt van $\frac{1}{2}$ ligt. Hieruit volgt, dat de Binomiale verdeling voor niet te kleine waarden van n, als p niet te dicht bij 0 of 1 ligt, goed benaderd kan worden door een normale verdeling.

Het blijkt in vele gevallen mogelijk te zijn discrete verdelingen door continue te benaderen. Continue verdelingen treden dus niet alleen op bij de studie van continue of als continu beschouwde variabele grootheden, doch ook als benaderingen van discrete verdelingen. Het vervangen van discrete verdelingen door continue biedt uit praktisch oogpunt meestal grote voordelen.

Indien men bijvoorbeeld de whn van alle binomiale verdelingen zou willen tabelleren volgens formule (1.3.1.) zou men voor iedere n en iedere p een afzonderlijke tabel moeten vervaardigen. De normale verdelingen, die de binomiale benaderen, zijn echter alle gemakkelijk tot één standaard type te herleiden, zodat men bij toepassing van deze benadering met één tabel kan volstaan. Een tweede voordeel van continue verdelingen is, dat zij, wiskundig gezien, gemakkelijker hanteerbaar zijn dan de meeste discrete verdelingen. Zo kan men dikwijls, de verdeling afleiden van functies van grootheden, die een bepaalde continue verdeling hebben, terwijl dit bij de meeste discrete verdelingen practisch onuitvoerbaar is. Bijzonder prettige eigenschappen bezit in dit verband de normale verdeling. (Zie Hoofdstuk 3)

1.7. Massatheoretische interpretatie van whn-verdelingen.

Men kan zich de tot nu toe beschouwde wh-verdelingen ook aanschouwelijk voorstellen als massaverdelingen over een rechte lijn. Men denkt zich daartoe een rechte lijn, die een totale massa van 1 eenheid (b.v. 1 kg) heeft en verdeeld op de lijn een nulpunt en schaalverdeling, zodat men bij iedere waarde x, die de beschouwde stochastische variabele kan aannemen, een daarmee overeenkomend punt met coördinaat x op de lijn kan aanwijzen. Indien men te doen heeft met een discrete verdeling, waarbij de stochastische variabele de discrete waarden $x_1, x_2, x_3,$ etc. aanneemt, is de massa in de daarmee overeenkomende punten samengeperst en wel in punt x_i

punt x_i een massa gelijk aan $P[\underline{x} = x_i]$. Bij een continue verdeling is de massa uitgesmeerd over de gehele lijn of over gedeelten daarvan, zodanig dat de soortelijke massa (de "dichtheid") in ieder punt x gelijk is aan de wh-dichtheid $f(x)$. De term wh-dichtheid wordt door deze voorstellingswijze verhelderd. Bovendien blijkt hieruit duidelijk het verschil tussen wh- en wh-dichtheid; dit is een overeenkomstig verschil als bestaat tussen een massa en een soortelijke massa, tussen een afgelegde weg er een snelheid etc. Wij zullen in het vervolg herhaaldelijk van deze voorstellingswijze gebruik maken.

1.8 Tweedimensionale wh-verdelingen

De tot nu toe beschouwde wh-verdelingen hebben betrekking op één variabele en kunnen vergeleken worden met massaverdelingen over een rechte lijn. Zij worden daarom één-dimensionale wh-verdelingen genoemd. Wij beschouwen nu een plat vlak, waarin een zeker coördinatenstelsel is aangebracht, zodat ieder punt van het vlak door een coördinatenpaar (x,y) bepaald is. Laat er over dit vlak een massa 1 verdeeld zijn, hetzij geconcentreerd op bepaalde lijnen of in bepaalde punten, hetzij uitgesmeerd over het gehele vlak of gedeelten ervan. Men kan nu de massa van ieder gedeelte van het vlak bepalen. Een tweedimensionale wh-verdeling kan met een dergelijke massaverdeling vergeleken worden; aan ieder gedeelte van het vlak wordt een wh toegekend gelijk aan de massa van dat gedeelte bij de massaverdeling.

Een voorbeeld van een tweedimensionale wh-verdeling heeft men in de worpresultaten van twee dobbelstenen A en B. Deze kunnen gekarakteriseerd worden door 2 stochastische variabelen $\underline{x}$ en $\underline{y}$; $\underline{x}$ is het aantal ogen geworpen met A, $\underline{y}$ het aantal ogen geworpen met B. Wij onderstellen hierbij dat de dobbelstenen A en B te onderscheiden zijn, b.v. doordat ze verschillende kleuren hebben. De wh-massa is hier geconcentreerd in 36 punten; $\underline{x}$ en $\underline{y}$ kunnen beide namelijk 6 waarden aannemen. Met ieder van deze punten correspondeert een kans, dat het betreffende worpresultaat bereikt wordt. De kans, dat men b.v. met de eerste dobbelsteen 2 ogen gooit en met de tweede 4 zal worden aangeduid met $P[\underline{x} = 2, \underline{y} = 4]$. De kans, dat men met de eerste dobbelsteen hoogstens 3 en met de tweede hoogsten 5 gooit wordt aangeduid met $P[\underline{x} \leqq 3, \underline{y} \leqq 5]$.

In het algemeen wordt bij een tweedimensionale wh-verdeling de kans $P[\underline{x} \leqq x, \underline{y} \leqq y]$ de <u>verdelingsfunctie</u> $F(x,y)$ in het punt x,y genoemd. In fig. 1.7 is aangegeven op welk gebied deze kans betrekking heeft. Men trekt door het punt (x,y) twee rechte lijnen, één evenwijdig aan de x-as en één evenwijdig aan de y-as. Van de vier gedeelten waarin het platte vlak door deze lijnen verdeeld wordt, neemt men het gebied links onder het punt (x,y). De verdelingsfunctie F(x,y) is dan gelijk aan de wh van dat gebied met inbegrip van zijn grenzen.

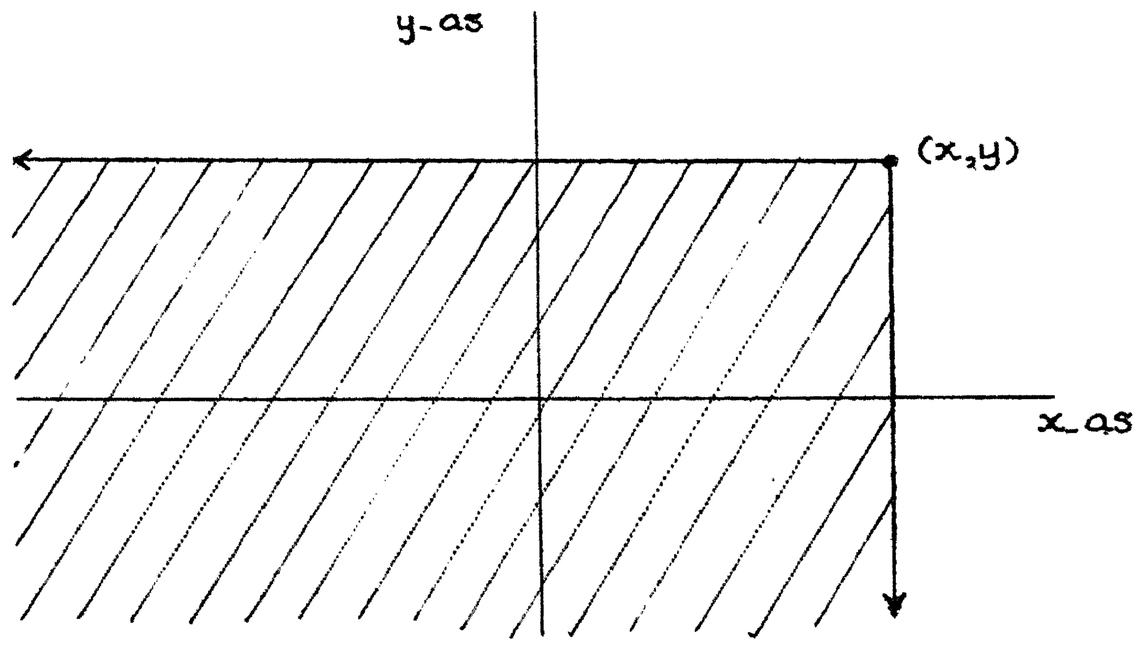


Fig. 1.7: F(x,y) is de wh dat ($\underline{x},\underline{y}$) valt in het gearceerde gebied of op de grenzen daarvan.

De functie F(x,y) is, omdat zij een kans voorstelt, steeds gelegen tussen 0 en 1 Zij kan niet afnemen, als wij ons in het vlak naar rechts, naar boven, of in een van de daar tussen gelegen richtingen bewegen. Als er op bepaalde lijnen, of in bepaalde punten een zekere wh-massa geconcentreerd is, zoals b.v. het geval is bij de twee dobbelstenen, dan zal F(x,y) op die lijnen of in die punten sprongen vertonen. Bij zogenaamde <u>continue</u> verdelingen, waarbij dergelijke punten of lijnen niet voorkomen, verloopt F(x,y) geleideijk.

Indien men in F(x,y) één van beide variabelen b.v. y naar oneindig laat gaan, gaat F(x,y) over in $F(x,\infty) = P[\underline{x} \leqq x, \underline{y} \leqq \infty] = P[\underline{x} \leqq x]$. Men verkrijgt dan de waarschijnlijkheid dat het punt ($\underline{x},\underline{y}$) links van of op een verticale lijn ligt. Indien men de gehele wh-massa in dat gebied evenwijdig aan de y-as verschuift naar de x-as, verkrijgt men op die as een ééndimensionale wh-verdeling. Deze wordt de <u>marginale verdeling</u> van $\underline{x}$ genoemd; $F(x,\infty)$ is de verdelingsfunctie van deze verdeling. Evenzo kan men de marginale verdeling van $\underline{y}$ definiëren. De verdelingsfunctie daarvan wordt gegeven door:

$F(\infty,y) = P[\underline{x} \leqq \infty, \underline{y} \leqq y] = P[\underline{y} \leqq y]$.

Na het voorgaande is het niet moeilijk om vast te stellen, wat wij onder $P[\underline{x} \leqq x,\ y_1 \leqq \underline{y} \leqq y_2]$ zullen verstaan. Dit is de wh van de strook begrensd door de horizontale lijnen $y=y_1$ en $y = y_2$, links van de verticale lijn die de punten met abscis x verbindt, met inbegrip van de grenzen van dat gebied (zie figuur 1.8).

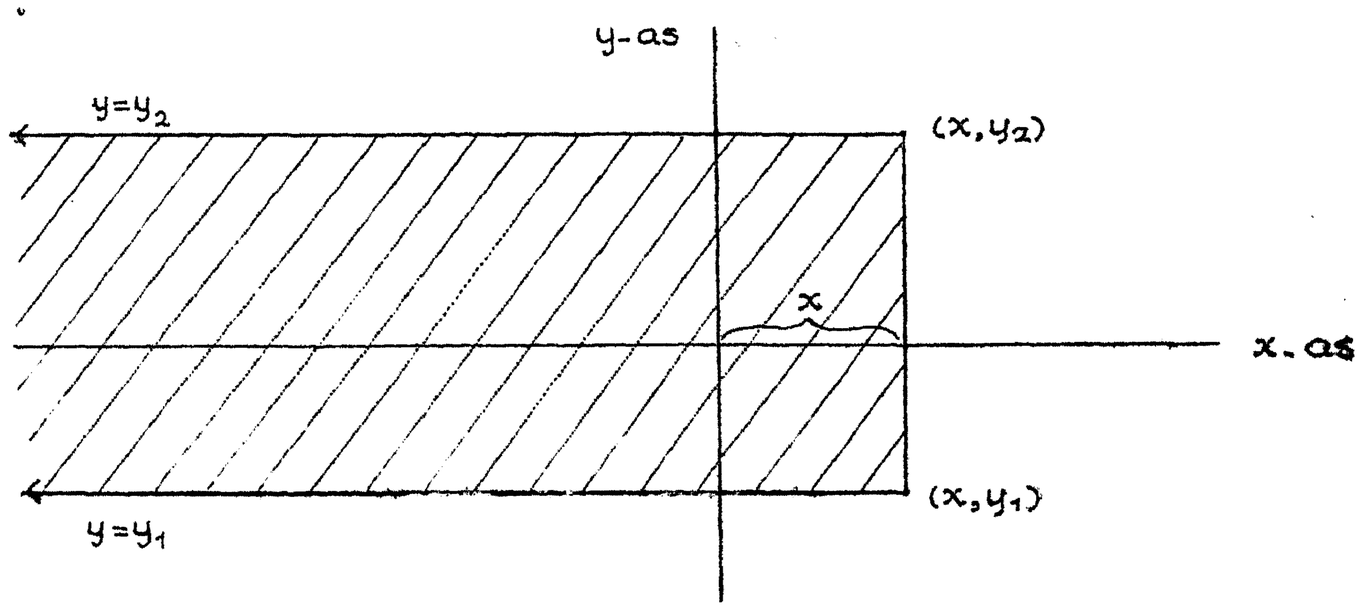


Fig.1.8. Gebied ($\underline{x} \leqq x,\ y_1 \leqq \underline{y} \leqq y_2$).

Als we de kans $P[\underline{x} \leqq x,\ y_1 \leqq \underline{y} \leqq y_2]$ beschouwen als functie van x, is dit een functie die veel lijkt op de verdelingsfunctie van een één-dimensionale wh-verdeling. Zij gaat in het algemeen echter niet naar 1, als x naar oneindig gaat, doch wordt dan gelijk aan $P[y_1 \leqq \underline{y} \leqq y_2]$. Als deze laatste kans niet nul is, zal het quotiënt

$$\frac{P[\underline{x} \leqq x,\ y_1 \leqq \underline{y} \leqq y_2]}{P[y_1 \leqq \underline{y} \leqq y_2]}$$

wel als een verdelingsfunctie te beschouwen zijn. Dit quotiënt wordt aangeduid met

$P[\underline{x} \leqq x \mid y_1 \leqq \underline{y} \leqq y_2]$ en wordt de kans, dat $\underline{x} \leqq x$ onder de voorwaarde $y_1 \leqq \underline{y} \leqq y_2$ genoemd. De corresponderende verdeling wordt de voorwaardelijke verdeling van $\underline{x}$ onder de voorwaarde $y_1 \leqq \underline{y} \leqq y_2$ genoemd.

Als de voorwaardelijke verdelingen van $\underline{x}$ in alle mogelijke horizontale stroken hetzelfde zijn, dus als de verdeling van $\underline{x}$ niet afhangt van de door $\underline{y}$ aangenomen waarden, zeggen wij dat $\underline{x}$ onafhankelijk is van $\underline{y}$.

In dat geval geldt:

$$P[\underline{x} \leqq x \mid \underline{y} \leqq y] = P[\underline{x} \leqq x \mid \underline{y} \leqq \infty] = P[\underline{x} \leqq x]$$

volgens de definitie van de voorwaardelijke kans in het linkerlid geldt:

$$P[\underline{x} \leqq x \mid \underline{y} \leqq y] = \frac{P[\underline{x} \leqq x,\ \underline{y} \leqq y]}{P[\underline{y} \leqq y]}.$$

Indien we nu de rechterleden van beide regels aan elkaar gelijk stellen vinden we, dat, als $\underline{x}$ onafhankelijk is van $\underline{y}$, geldt:

$P[\underline{x} \leqq x, \underline{y} \leqq y] = P[\underline{x} \leqq x] \cdot P[\underline{y} \leqq y]$ (Vergelijk par.1.2)

Hieruit is gemakkelijk af te leiden, dat als $\underline{x}$ onafhankelijk is van $\underline{y}$, ook $\underline{y}$ onafhankelijk is van $\underline{x}$. Men kan dan dus zeggen dat $\underline{x}$ en $\underline{y}$ (onderling) onafhankelijk zijn.

In het geval dat $\underline{x}$ en $\underline{y}$ onafhankelijk zijn, kan men dus de tweedimensionale verdeling van $\underline{x}$ en $\underline{y}$ afleiden als men de marginale verdelingen kent. Als $\underline{x}$ en $\underline{y}$ afhankelijk zijn, is dit niet mogelijk.

1.9 Twee dobbelstenen

Boven hebben we reeds 2 dobbelstenen A en B beschouwd; $\underline{x}$ was het aantal ogen geworpen met A, $\underline{y}$ het aantal ogen geworpen met B. Als men de stenen vóór iedere worp goed schudt, zal het worpresultaat van A in het algemeen niet afhangen van dat van B.Men zal $\underline{x}$ en $\underline{y}$ dan dus als onderling onafhankelijke variabelen kunnen beschouwen. We hebben in dat geval te doen met een wh-verdeling over de punten in het in fig.1.9 getekende rooster. Als we de whn $P[\underline{x} = i]$ aangeven door p_i en de whn $P[\underline{y} = j]$ door p'_j, geldt voor de wh van een worpresultaat $\underline{x} = i$, $\underline{y} = j$:

$$P[\underline{x} = i, \underline{y} = j] = p_i p'_j.$$

Indien beide dobbelstenen zuiver zijn ($p_1 = \ldots = p_6 = p'_1 = \ldots = p'_6 = \frac{1}{6}$), is dus de wh. van ieder worpresultaat $\frac{1}{36}$.

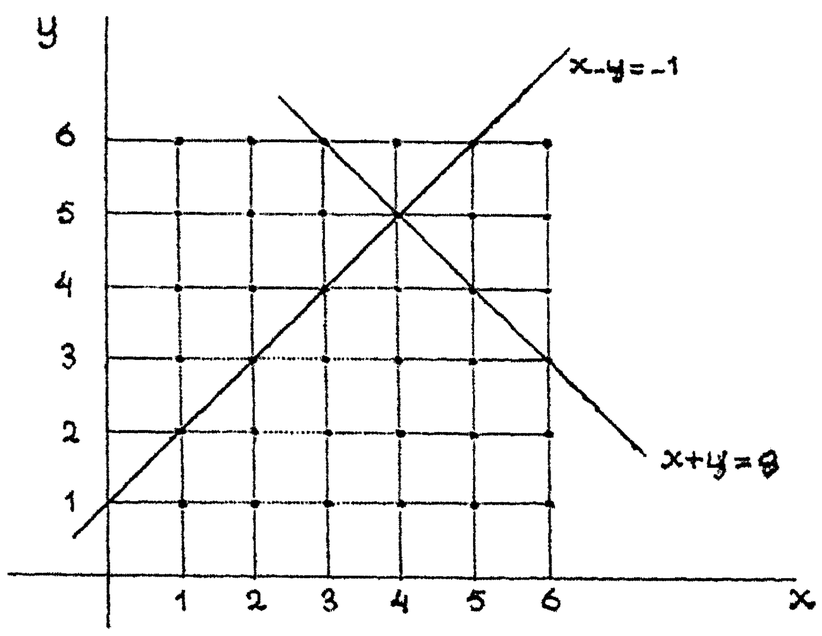


Fig.1.9 Mogelijke worpresultaten bij twee dobbelstenen.

Wij kunnen hierbij het volgende opmerken, als $P[\underline{x}=i, \underline{y}=j] = 0$ is, zal p_i en/of $p_j = 0$ moeten zijn. Tevens zal als b.v. $p_k = 0$ dan zal $P[\underline{x} = k, \underline{y} = j] = 0$ moeten zijn voor iedere waarde van j. Als dus (bij onafhankelijkheid van $\underline{x}$ en $\underline{y}$) één punt in het rooster een wh. nul heeft, moet dit het geval zijn met een hele verticale of horizontale rij van punten.Dit betekent dus, dat bij een discrete twee-dimensionale wh-verdeling (waar dus de wh.in punten geconcentreerd is) de variabelen $\underline{x}$ en $\underline{y}$ alleen onafhankelijk kunnen

zijn, als de punten een volledig rechthoekig rooster vormen, waarvan de zijden evenwijdig lopen met de coördinaatassen. Dit wil, zoals we straks zullen zien. nog niet zeggen, dat $\underline{x}$ en $\underline{y}$ onafhankelijk moeten zijn, als de punten, waarin de wh.geconcentreerd is, een rechthoekig rooster vormen.

We kunnen, gebruik makend van figuur 1.9 ook gemakkelijk de verdeling van $\underline{x} + \underline{y}$ afleiden. Willen we b.v. de kans $P[\underline{x}+\underline{y}=9]$ weten, dan sommeren we de whn van alle punten liggend op de lijn $x + y = 9$ in het rooster. Bij onafhankelijkheid geldt dus:

$$P[\underline{x} + \underline{y} = 9] = p_6p'_3 + p_5p'_4 + p_4p'_5 + p_3p'_6.$$

Om de kans $P[\underline{x} + \underline{y} \leqq 9]$ te vinden sommeren we de whn van alle roosterpunten op en onder de lijn $x + y = 9$. Evenzo kunnen we de verdeling van $\underline{x} - \underline{y}$ afleiden. Er geldt b.v. (bij onafhankelijkheid):

$$P[\underline{x} - \underline{y} = -1] = p_1p'_2 + p_2p'_3 + p_3p'_4 + p_4p'_5 + p_5p'_6.$$

Opgave:

1.9.a Bepaal de wh-verdeling van $\underline{x} + \underline{y}$ als beide dobbelstenen zuiver zijn en als $\underline{x}$ onafhankelijk is van $\underline{y}$. Voor dezelfde gevallen ook de verdeling van $\underline{x} - \underline{y}$. Teken een strepen-diagram van beide verdelingen (Verg.Fig.1.9)

We willen nu de beide dobbelstenen niet meer onderscheiden, en dus het resultaat: i ogen geworpen met A en j met B, identificeren met j ogen geworpen met A en i met B. Wij veronderstellen, dat de resultaten steeds genoteerd worden met (x,y), waarbij $x \geqq y$ genomen wordt. De mogelijke worpresultaten worden nu gegeven door fig.1.10:

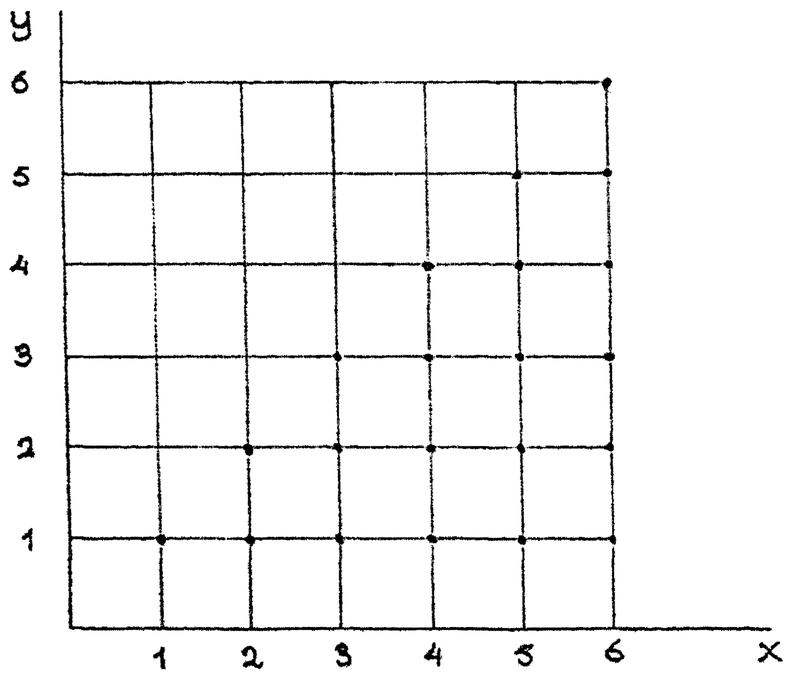


Fig.1.10.
Mogelijke worpresultaten met twee niet onderscheiden dobbelstenen.

We zien, dat de punten hier een driehoekig rooster vormen. Hieruit blijkt, dat de variabelen $\underline{x}$ en $\underline{y}$ niet meer onafhankelijk kunnen zijn. Als $\underline{y}$ b.v. gelijk is aan 2 kan $\underline{x}$ de waarden 2, 3, 4,5, en 6 aannemen, doch als $\underline{y}$ gelijk is aan 5 blijven er alleen de waarden 5 en 6 voor $\underline{x}$ over.

Indien de dobbelstenen onafhankelijk geworpen worden en we de hierboven ingevoerde notatie p_i en p'_j gebruiken, geldt:

Als $i \neq j$: $P[\underline{x} = i, \underline{y} = j] = p_i p'_j + p'_i p_j$

en als $i = j$: $P[\underline{x} = i, \underline{y} = i] = p_i p'_i$.

Opgave

1.9.b Bereken voor het geval van fig.1.10 de verdeling van $\underline{x}+\underline{y}$ en die van $\underline{x} - \underline{y}$, als beide dobbelstenen zuiver zijn en onafhankelijk van elkaar geworpen worden. Wat merkt ge op bij vergelijking van dit resultaat met dat van vraagstuk 1.9.a ?

1.10 Grafische voorstelling van 2-dimensionale wh-verdelingen. Continue 2-dimensionale verdelingen

Om een goed beeld te verkrijgen van een 2-dimensionale wh-verdeling dient men in feite 3-dimensionale diagrammen (zogenaamde blokdiagrammen) te vervaardigen. Men heeft immers reeds 2 dimensies nodig om de variabelen uit te zetten, een derde dimensie is vereist om de whn uit te zetten. Men zou een diagram kunnen vervaardigen van de dobbelsteenverdeling van fig.1.9.door in ieder roosterpunt een staafje op te richten waarvan de lengte evenredig is met de wh van dat punt. Als $\underline{x}$ en $\underline{y}$ onafhankelijk zijn zullen de lengten van de staafjes op iedere horizontale lijn van het rooster hetzelfde patroon vertonen, omdat voor iedere j geldt:

$$p_{1j} : p_{2j} : \ldots\ldots : p_{6j} = p_1 : p_2 : \ldots\ldots : p_6$$

Hetzelfde geldt voor de staafjes op de verticale roosterlijnen.

Histogrammen van 2-dimensionale verdelingen kunnen vervaardigd worden, door het platte vlak met behulp van lijnen evenwijdig aan de assen in rechthoeken te verdelen, en op iedere rechthoek een blokje te plaatsen, waarvan de inhoud evenredig is met de wh binnen de rechthoek. De hoogte van een dergelijk blokje gedeeld door de oppervlakte van het grondvlak geeft ons de relatieve wh (de wh per oppervlakteëenheid) voor de rechthoek, waarop het blokje staat. Als de wh-verdeling niet geconcentreerd is in bepaalde punten, of op bepaalde lijnen, kan men de indeling van het platte vlak steeds fijner maken en zodoende voor iedere willekeurige kleine omgeving van ieder punt een dergelijke relatieve wh vinden. In het limietgeval, gaat deze relatieve wh over in een grootheid, die de wh-dichtheid in het beschouwde punt genoemd worden;[6)]

---

6) De hier bedoelde limietovergang is niet altijd mogelijk als de wh-verdeling continu is. Bij de in de practijk voorkomende typen continue verdelingen is dit echter wel steeds het geval.

deze zal alleen een functie zijn van de coördinaten x en y van het punt en wordt aangeduid met f(x,y). Als men nu boven ieder punt van het platte vlak, de waarde van f(x,y) uitzet, verkrijgt men een "berglandschap" dat de verdeling karakteriseert. In de buurt van hooggelegen punten zal meer wh geconcentreerd zijn dan in de buurt van laaggelegen punten. Om in een vlakke tekening een indruk te krijgen van het diagram van f(x,y) tekent men wel lijnen die alle punten van gelijke dichtheid verbinden, een methode die ook dikwijls bij landkaarten wordt toegepast. In gebieden waar deze lijnen dicht bij elkaar liggen, zal het "berglandschap" een sterke helling vertonen en vindt men dus grote dichtheidsverschillen; waar de lijnen ver van elkaar liggen heeft de verdeling een vlakker verloop.

Ook in het geval van de tweedimensionale verdeling, kan de wh-dichtheid afgeleid worden uit de verdelingsfunctie en omgekeerd. Om f(x,y) te vinden uit F(x,y) moet de bewerking van het differentiëren tweemaal uitgevoerd worden; eerst voor de variabele x en dan voor de variabele y of omgekeerd. Wij schrijven

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

De bewerking die we dan uitvoeren komt in feite neer op het proces dat we boven gevolgd hebben om f(x,y) te definiëren. De wiskunde biedt echter voor speciale functies F(x,y) formules om het resultaat van dit proces direct op te schrijven. Om F(x,y) af te leiden uit f(x,y) zullen wij tweemaal moeten integreren; ten aanzien van de variabele x en ten aanzien van de variabele y. Men schrijft:

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

F(x,y) is de inhoud van het gedeelte van het "berglandschap" gelegen binnen de rechte hoek met hoekpunt (x,y) en benen evenwijdig aan de coördinaatassen. Hieruit volgt dat de totale inhoud van het "berglandschap" steeds = 1 moet zijn.

Als de wh-verdeling van (x,y) continu is, zullen ook de marginale verdelingen continu zijn. Als de verdelingsdichtheden van deze verdelingen worden voorgesteld door g(x) en h(y) en als x en y onafhankelijk zijn, zal gelden voor ieder punt (x,y): f(x,y) = g(x) h(y). In dat geval zullen dus alle verticale doorsneden door het "berglandschap" evenwijdig aan de x-as gelijkvormig moeten zijn, en evenzo alle verticale doorsneden evenwijdig aan de y-as. Tevens

blijkt hieruit, dat bij onafhankelijkheid van $\underline{x}$ en $\underline{y}$, $f(x,y)$ alleen maar nul kan zijn, in punten waar of $g(x)$ en of $h(y)$ nul is en omgekeerd als $g(x)$ b.v. nul is voor $x > x_1$. Het gebied waar de dichtheid $f(x,y) > 0$ is, zal dus als $\underline{x}$ en $\underline{y}$ onafhankelijk zijn, steeds begrensd moeten worden door lijnen evenwijdig aan de assen van het coördinatenstelsel, of het zal het gehele platte vlak moeten beslaan.

Opgaven

1.10.a Ga na of $\underline{x}$ en $\underline{y}$ onafhankelijk zijn in de volgende gevallen:

1) De functie $f(x,y)$ wordt voorgesteld door het bovenvlak van een kubus, waarvan de zijvlakken evenwijdig zijn aan de coördinaatassen.

2) De functie $f(x,y)$ wordt voorgesteld door het bovenvlak van een kubus, waarvan nu echter de verticale diagonaalvlakken evenwijdig zijn aan de coördinaatassen. (Deze kubus is dus 45° gedraaid ten opzichte van de vorige).

3) De functie $f(x,y)$ wordt voorgesteld door een der zijvlakken van een regelmatige vierzijdigen pyramide (Vierkant grondvlak met top boven het midden van het vierkant). De zijden van het grondvlak lopen evenwijdig aan de coördinaatassen.

Opmerking: In het bovenstaande hebben wij vooral de aandacht geschonken aan het geval dat $\underline{x}$ en $\underline{y}$ onafhankelijk zijn. In de hoofdstukken over correlatie zullen wij de twee dimensionale verdelingen opnieuw tegenkomen. Wij zullen daarin methoden leren kennen om bepaalde structurele afhankelijkheden tussen $\underline{x}$ en $\underline{y}$ te onderkennen.

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 155

Cursus

Toegepaste Statistiek II

door

Ph. van Elteren en J. Kriens

II. Verwachting en spreiding

1955

2. Verwachting en spreiding.

2.1. De verwachting van een alternatieve verdeling.

Wij beschouwen een zeer eenvoudige stochastische variabele $\underline{x}$, die met kansen $\frac{1}{2}$ de waarden 0 of 1 aanneemt en nemen een grote steekproef b.v. van N waarnemingen uit de corresponderende verdeling. In dat geval zullen in het algemeen ongeveer $\frac{1}{2}N$ waarnemingen de waarde 0 en $\frac{1}{2}N$ waarnemingen de waarde 1 hebben. De gemiddelde waarde van deze waarnemingen zal ongeveer worden:

$$(2.1.1) \qquad \frac{\frac{1}{2}N.1 + \frac{1}{2}N.0}{N} = \frac{1}{2}$$

De gemiddelde waarde van lange reeksen waarnemingen uit een dergelijke verdeling zal dus in het algemeen weinig van $\frac{1}{2}$ afwijken. Wij kennen daarom aan onze wh-verdeling ook een "gemiddelde" toe. Dit gemiddelde stellen wij gelijk aan $\frac{1}{2}$. Men dient het gemiddelde van de wh-verdeling goed te onderscheiden van het gemiddelde van steekproeven. Voor de bovenbeschreven wh-verdeling is het gemiddelde steeds precies $\frac{1}{2}$. Indien wij meerdere grote steekproeven nemen, zullen de gemiddelden daarvan weliswaar zelden veel van $\frac{1}{2}$ afwijken, maar toch in het algemeen onderling wel enigszins verschillen.

In plaats van de uitdrukking het gemiddelde van een verdeling, spreekt men ook vaak van de verwachting van die verdeling, om verwarring te voorkomen. Wij zullen dat hier ook doen.

De verwachting van een stochastische variabele $\underline{x}$, dit is dus het gemiddelde van de wh-verdeling van $\underline{x}$, wordt hier aangeduid met $\mathcal{E}\underline{x}$. Zij wordt in het algemeen zo gedefinieerd, dat het gemiddelde van grote steekproeven uit de verdeling van $\underline{x}$ er ongeveer mee overeenkomt. In het hier beschouwde geval was dus $\mathcal{E}\underline{x} = \frac{1}{2}$.

Als nu $\underline{x}$ niet met kansen $\frac{1}{2}$ en $\frac{1}{2}$, maar bijvoorbeeld met kans $\frac{1}{3}$ resp. $\frac{2}{3}$ de waarden 0 resp. 1 aanneemt, vinden wij door een redenering als boven gegeven, dat de gemiddelde waarde van een reeks van N waarnemingen ongeveer gelijk zal worden aan

$$(2.1.2) \qquad \frac{\frac{2}{3}N.1 + \frac{1}{3}N.0}{N} = \frac{2}{3}$$

In dat geval zullen wij de verwachting van $\underline{x}$ dus gelijk stellen aan $\frac{2}{3}$ (dus $\mathcal{E}\underline{x} = \frac{2}{3}$). De twee beschreven gevallen zijn voorbeelden van een zogenaamde alternatieve

verdeling. In het algemeen is dat een verdeling, waarbij $\underline{X}$ de waarden x_1 en x_2 kan aannemen, met kansen p_1 en p_2 ($p_1 + p_2 = 1$). We vinden nu voor het gemiddelde van grote steekproeven geheel analoog: ongeveer $p_1 x_1 + p_2 x_2$.

We zullen dus voor een alternatieve verdeling de verwachting van $\underline{X}$ definieren als:

$$(2.1.3) \qquad \mathcal{E}\,\underline{X} = p_1 x_1 + p_2 x_2 \ .$$

2.2. Voorbeeld: Kruis en Munt.

Een van de oudste toepassingen van het begrip verwachting vindt men in de theorie van de kansspelen. Stel dat twee personen A en B kruis of munt gaan spelen en afspreken dat A a gulden aan B zal betalen telkens als er kruis geworpen wordt en B b gulden aan A telkens als er munt geworpen wordt. Speler A zal in de regel niet geneigd zijn aan het spel deel te nemen, als hij mag verwachten dat hij in lange reeksen spelen gemiddeld meer zal verliezen dan winnen. De door A aan het spel te stellen voorwaarde kunnen we nu wiskundig formuleren. Als de kans om kruis te werpen p is, en de kans om munt te werpen q, dan is de winst die A met een spel maakt een stochastische variabele $\underline{X}$, die met de kans p de waarde $-a$ en met een kans q de waarde $+b$ aanneemt; immers bij kruis verliest A het bedrag a aan B, hij wint dus een bedrag $-a$. Wij weten nu dat de gemiddelde winst van A in een lange reeks spelen, ongeveer gelijk zal worden aan het gemiddelde van de wh-verdeling van $\underline{X}$, dus aan de verwachting $\mathcal{E}\,\underline{X}$ van $\underline{X}$. Volgens formule (2.1.3), waarin we invullen $x_1 = -a$, $x_2 = b$, $p_1 = p$, $p_2 = q$ wordt deze verwachting

$$\mathcal{E}\,\underline{X} = -pa + qb \ .$$

A zal dus alleen willen spelen als $\mathcal{E}\,\underline{X}$ niet negatief is, dus als geldt:

$$(2.2.1) \qquad -pa + qb \geqq 0 \ .$$

Maar de winst van B is eveneens een stochastische variabele $\underline{Y}$. Deze neemt met een kans p de waarde $+a$ en met een kans q de waarde $-b$ aan; immers als er kruis geworpen wordt wint B a, doch bij munt verliest hij b. B zal dus alleen willen spelen als de verwachting van zijn winst niet negatief is, dus als geldt

$$(2.2.2) \qquad pa - qb \geqq 0 \ .$$

Uiteindelijk vinden we dus, dat beide spelers alleen bereid zijn om te spelen, als tegelijk aan (2.2.1) en (2.2.2) voldaan is, dus als geldt:

(2.2.3) $pa - qb = 0$.

In dat geval zegt men dat het spel eerlijk is. Het blijkt dat het hier beschreven kruis-munt spel eerlijk is als geldt:

$pa = qb$ of volgens de hoofdeigenschap van evenredigheden:

$a : b = q : p$.

Aangezien p de kans op kruis is, of de kans dat B wint, en q de kans op munt of de kans dat A wint, geldt dat het bedrag dat een speler bij verlies moet betalen evenredig moet zijn aan de kans dat die speler wint, of dat het bedrag dat een speler bij winst ontvangt omgekeerd evenredig moet zijn aan de kans dat hij wint. Hieruit volgt dat bij een spel met een zuivere munt ($p = q = \frac{1}{2}$) de bedragen a en b aan elkaar gelijk moeten zijn.

Indien het spel niet eerlijk is, bijvoorbeeld $p = \frac{2}{3}$ $a = 3$ $b = 3$ met $pa - qb = 1$, wat betekent dat B per spel gemiddeld 1 gulden van A zal winnen, dan zal A toch bereid zijn op het spel in te gaan, als B hem vóór het begin van ieder spel f 1.- betaalt. In het algemeen zal als $pa - qb > 0$ is, B bij het begin van ieder spel $pa - qb$ aan A moeten betalen, en als $pa - qb < 0$ is, zal A $-(pa - qb)$ aan B moeten betalen. Een "oneerlijk" spel kan dus eerlijk worden, als de speler waarvan de verwachte winst positief is, vóór ieder spel deze verwachte winst aan de tegenspeler wil betalen.

Het is duidelijk dat men, om een kruis-munt-spel eerlijk te maken, de kansen p en q moet kennen. Dikwijls zal dit niet het geval zijn. We kunnen dan echter gebruik maken van de wet van de grote getallen. Voor het kruis-munt-spel betekent dit, dat we een groot aantal, b.v. 10000 keren met de munt werpen en dan het aantal keren noteren dat er kruis optreedt. Laat dit 3014 maal zijn. We weten dan dat de kans p op kruis ongeveer $\frac{3014}{10000} \approx 0{,}3$ moet zijn. Het spel zal dus ongeveer eerlijk zijn als a : b = 7 : 3 is.

Opm. Mocht één van beide spelers zich niet veilig voelen met deze schatting van p , dan kan hij zich een indruk vormen van de nauwkeurigheid van de schatting door een betrouwbaarheidsinterval voor p te bepalen. We hebben het

bepalen van een betrouwbaarheidsinterval voor een onbekende kans uitvoerig besproken in de Cursus Toegepaste Statistiek I (Hoofdstuk 3). We willen hier slechts vermelden, dat een betrouwbaarheidsinterval met tweezijdige onbetrouwbaarheid 0,05 hier voor p de grenzen 0,292 en 0,310 oplevert; dit betekent dat onze schatting behoudens een kans 0,05 niet meer dan 0,01 fout kan zijn.

Uit het voorafgaande blijkt, dat het begrip verwachting van belang is als kriterium voor de "eerlijkheid" van een kansspel. Men mag hieruit niet concluderen, dat de kennis van de verwachte winst voldoende is om uit te kunnen maken of het verantwoord is aan een kansspel deel te nemen. Er zijn nl. gevallen, waarbij men een zeer groot aantal keren moet spelen, om een gemiddelde winst te bereiken, die de verwachte winst van het spel goed benadert. Stel men heeft b.v. een loterij waaraan 100.000 personen deelnemen, terwijl er één prijs is ter waarde van ƒ 1.000.000.--. Een willekeurige deelnemer heeft nu een kans 0,00001 om ƒ 1.000.000.- te winnen, en een kans 0,99999 om ƒ 0.- te winnen. De <u>verwachting</u> van zijn winst is dus 0,00001 x ƒ 1.000.000.- = ƒ 10.- . Ook al neemt men honderd keren aan een dergelijke loterij deel, dan is het nog bijna zeker, dat men geen prijs wint, (de kans daarop is nl. ongeveer 0,999). De gemiddelde winst van 100 spelen zal dus bijna zeker 0 zijn. Als de prijs van een lot in deze loterij ƒ 10.- bedraagt, is zij volkomen eerlijk, men zal echter zeer vaak (duizenden keren) in dergelijke loterijen moeten spelen, om gemiddeld ongeveer even veel te winnen als men voor de loten betaalt.

In het algemeen is de verwachte winst van een spel moeilijk te realiseren, als het gaat om <u>kleine</u> kansen op het winnen van <u>grote</u> bedragen. Iemand die aan een dergelijke loterij deelneemt is er bijna zeker van de prijs van het lot te zullen verliezen. Meestal is ook de verwachte winst negatief, omdat een gedeelte van de opbrengst der loten bestemd is voor administratiekosten en winst voor de organisatoren (of voor een liefdadig doel). Hieruit volgt niet, dat het onverstandig zou zijn aan een dergelijke loterij deel te nemen; als de prijs per lot niet te hoog is, zal men deze veelal overhebben voor de <u>mogelijkheid</u> om een groot bedrag te winnen (en eventueel uiteraard voor het liefdadige doel).

2.3. 2e Voorbeeld: Een Verzekeringscontract.

Een zeer belangrijke toepassing vindt het begrip verwachting op het terrein der verzekeringen. De contracten van verzekeringsmaatschappijen hebben het karakter van kansspelen; of de maatschappij winst of verlies maakt op een bepaald contract hangt af van het al dan niet optreden van een bepaalde onzekere gebeurtenis. Wij willen hier een zeer eenvoudig voorbeeld van een dergelijk contract behandelen. Voor zijn vertrek per vliegtuig van Amsterdam naar New York sluit iemand een contract met een verzekeringsmaatschappij, inhoudende, dat ƒ 10.000.- aan zijn nabestaanden zal worden uitgekeerd als hij verongelukt. De prijs van de polis bedraagt c gulden. De relevante kans ρ is hier de kans, dat een reiziger die per vliegtuig van Amsterdam naar New York reist, omkomt. De winst van de reiziger (of beter van diens nabestaanden) is 10000-c als hij omkomt, en -c als dat niet gebeurt, de verwachting van de winst is dus: $+(10000-c).\rho + -c(1-\rho) = 10000\rho - c$. De reiziger zal zich in het algemeen niet bekommeren om de grootte van deze verwachte winst (die in de regel negatief zal zijn). Hij zal vermoedelijk zelfs liever de verliezer dan de winnaar van het spel zijn. De verzekeringsmaatschappij heeft er echter belang bij om te zorgen, dat haar verwachte winst, dat is dus het tegengestelde van de verwachte winst der verzekerde, positief is, omdat zij er een bedrijf van maakt dit "spel" zeer vaak te spelen. Zij zal er dus voor moeten zorgen dat geldt:

$$-(10000\rho - c) \geqq 0, \text{ of } c \geqq 10000\rho .$$

Hiertoe zal de maatschappij de kans ρ althans bij benadering moeten kennen en dan het bedrag c zo moeten kiezen, dat aan bovenstaande ongelijkheid voldaan is. Ter bepaling van ρ kan men gebruik maken van statistieken betreffende het aantal dodelijke ongevallen van luchtreizigers op het traject Amsterdam-New York, betrokken op het totaal aantal luchtreizigers op dit traject in een bepaalde periode. Zou men b.v. vinden, dat ongeveer 1 op de 5000 luchtreizigers op dit traject omkomt, dan zal de premie c minstens $ƒ \frac{10000}{5000} = ƒ 2.-$ moeten bedragen. Uiteraard zal dit bedrag verhoogd worden om de onkosten te dekken en een winstmarge voor de maatschappij te verkrijgen.

In principe moet een verzekeringsmaatschappij er voor zorgen, alléén contracten te sluiten, waarvoor de verwachting der winst niet negatief is. Het voorbeeld, dat

we hier behandeld hebben, is een voorbeeld van een <u>levens-verzekering</u>; de uitkering van de verzekerde som hangt af van het al dan niet in leven blijven van een bepaalde persoon in een bepaalde periode (hier de duur van de luchtreis). Het betreft hier echter een zeer speciaal voorbeeld, omdat de duur van de verzekering maar zeer kort is en alleen verzekerd wordt tegen overlijden tengevolge van een bepaalde doodsoorzaak (nl.: een vliegtuigongeval). Bij de "gewone" levensverzekeringscontracten is de duur gewoonlijk veel langer. Men moet daardoor tevens rekening houden met een rentefactor, waardoor de vergelijking waaruit de premies bepaald worden ingewikkelder zijn. Zij berusten echter op hetzelfde beginsel als hierboven beschreven.

2.4. <u>De verwachting van een discrete verdeling</u>.

Een stochastische variabele $\underline{x}$ is discreet verdeeld als zij met whn $p_1, p_2, \ldots$ discrete waarden $x_1, x_2, \ldots$ kan aannemen. In de voorafgaande paragrafen hebben wij het geval beschouwd, waarbij $\underline{x}$ slechts twee waarden kon aannemen: x_1 en x_2. De verwachting van $\underline{x}$ werd daarbij gedefinieerd als:

$$\mathcal{E}\,\underline{x} = p_1 x_1 + p_2 x_2 .$$

Wij hebben verder gezien dat het gemiddelde van grote steekproeven uit een dergelijke verdeling in het algemeen weinig van de verwachting zal afwijken.

Indien de variabele meer dan 2, b.v. n waarden, kan aannemen, definieren wij de verwachting als;

$$(2.4.1) \qquad \mathcal{E}\,\underline{x} = p_1 x_1 + p_2 x_2 + \ldots + p_n x_n = \sum_{i=1}^{n} p_i x_i ,$$

waarin p_i de kans $P\{\underline{x} = x_i\}$ voorstelt (Er moet uiteraard gelden: $\sum_i p_i = 1$).

Ook deze verwachting kan benaderd worden door het gemiddelde van een grote steekproef uit de verdeling. Immers de fractie van de waarnemingen met de waarde x_i zal in grote steekproeven niet veel afwijken van de kans p_i. Zoals wij reeds bij een eenvoudig voorbeeld in par. 2.3 gezien hebben hangt het aantal waarnemingen in de steekproeven, nodig om een goede benadering van $\mathcal{E}\,\underline{x}$ te verkrijgen, van de vorm van de verdeling af.

Men kan met behulp van formule (2.4.1) de verwachte winst becijferen, van meer ingewikkelde kansspelen, dan het in par. 2.2 beschreven kruis-munt spelletje.

Laten A en B b.v. spelen met een dobbelsteen; telkens als er i ogen geworpen worden (i = 1,2,...,6) betaalt A i gulden aan B. De verwachting van de winst van B is dan per spel:

$$\mathcal{E}\underline{x} = 1.p_1 + 2.p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6 ,$$

als p_i de kans voorstelt dat er i ogen geworpen worden. Bij een zuivere dobbelsteen zal dus gelden:

$$\mathcal{E}\underline{x} = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3\tfrac{1}{2} .$$

Om dit spel met A te mogen spelen, zou dus B voor ieder spel ƒ 3½,- aan A moeten betalen.

De verwachting van een binomiale verdeling met $n = 3$ kan als volgt berekend worden: (zie par. 1.3)

$$\mathcal{E}\underline{x} = 0.P\{\underline{x} = 0\} + 1.P\{\underline{x} = 1\} + 2.P\{\underline{x} = 2\} + 3P\{\underline{x} = 3\}$$

$$= 0.q^3 + 1.3pq^2 + 2.3pq^2 + 3.p^3$$

$$= 3p(q^2 + 2pq + p^2) = 3(p+q)^2 p = 3p,$$

want $p + q = 1$.

<u>Opgave 2.4.a</u>. Ga na wat de verwachtingen der binomiale verdelingen voor n = 1,2,4 en 5 zijn. Wat merkt ge hierbij op? (Vergelijk opgave 1.3.a).

Men kan ook nagaan, wat een lot waard is in een loterij met meerdere prijzen. Laten er bijvoorbeeld 100.000 personen deelnemen aan een loterij, met een hoofdprijs van ƒ 10.000.-, 2 prijzen van ƒ 5.000.-, 4 prijzen van ƒ 2.500.- en 5 prijzen van ƒ 1.000.-. Een deelnemer aan een dergelijke loterij heeft kansen:

0,00001 op ƒ 10.000.-
0,00002 op ƒ 5.000.-
0,00004 op ƒ 2.500.-
0,00005 op ƒ 1.000.-
en 0,99988 op ƒ 0.- .

Om de verwachting van deze verdeling te vinden moet men de kansen met de overeenkomstige bedragen vermenigvuldigen. Men vindt dan $\mathcal{E}\underline{x} = 0,35$. Dit betekent dus, dat de verwachte winst ƒ 0,35 bedraagt, zodat men voor het lot ƒ 0,35 kan betalen. Andere voorbeelden kan men ontlenen aan premies voor verzekeringscontracten, waarbij tegen meerdere elkaar uitsluitende risico's verzekerd wordt voor verschillende bedragen. We willen hier echter niet nader op ingaan.

Opgemerkt moet worden, dat n niet altijd eindig behoeft te zijn. Wij kunnen b.v. het volgende spel beschouwen: A en B spelen kruis of munt met een zuivere munt; telkens als er kruis geworpen wordt betaalt B ƒ 1.- aan A; het spel is beëindigd zodra er een keer munt geworpen wordt. Bij een dergelijk spel ontvangt A dus evenveel gulden als er achter elkaar kruis geworpen wordt. Als er de eerste keer munt geworpen wordt ontvangt A niets, de kans daarop is $\frac{1}{2}$; als een keer kruis en daarna munt geworpen wordt, ontvangt A ƒ 1.--; de kans daarop is $P[K].P[M] = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$.
In het algemeen is de kans op k maal achtereen kruis en daarna munt, dus de kans dat A ƒ k.- ontvangt

$$\{P[K]\}^k P[M] = \frac{1}{2^k}\cdot\frac{1}{2} = \frac{1}{2^{k+1}} .$$

De mogelijkheden zijn hier dus (als $\underline{x}$ de winst voorstelt, die A maakt):

$\underline{x} = 0,1,2,3,4,$ etc.

met kansen:

$$P\{\underline{x}=0\} = \frac{1}{2}\ ,\ P\{\underline{x}=1\} = \frac{1}{2^2}\ ,\ \ldots\ ,\ P\{\underline{x}=k\} = \frac{1}{2^{k+1}}\quad etc.$$

Dus geldt:

$$\mathcal{E}\,\underline{x} = \frac{1}{2}\cdot 0 + \frac{1}{2^2}\cdot 1 + \frac{1}{2^3}\cdot 2 + \ldots = \sum_{k=0}^{\infty}\frac{k}{2^{k+1}} .$$

Er kan bewezen worden dat deze som van oneindig veel termen gelijk is aan 1. De verwachte winst van A wordt dus ƒ 1.-, zodat A ƒ 1.- zal moeten inzetten om dit spel te mogen spelen.

Bij verdelingen, waarbij $\underline{x}$ oneindig veel waarden kan aannemen, kan zich de moeilijkheid voordoen, dat de verwachting zoals we die hier gedefinieerd hebben, oneindig groot wordt. Indien b.v. A en B zouden afspreken, dat B aan A 2^k gulden betaalt, als er k maal achtereenvolgens kruis geworpen wordt met een zuivere munt, dan wordt

$$\mathcal{E}\,\underline{x} = \frac{1}{2}\cdot 2^0 + \frac{1}{2^2}\cdot 2^1 + \frac{1}{2^3}\cdot 2^2 + \ldots .$$

Dit is een reeks van oneindig veel termen, waarbij iedere term gelijk wordt aan $\frac{1}{2}$. Als we onze formule (2.4.1) hier dus consequent toepassen, vinden we dat de verwachte winst van A oneindig wordt. Toch zal A bij dit spelletje, weliswaar soms een groot, maar toch steeds een eindig bedrag

winnen. Ook over een groot aantal spelen zal de gemiddelde winst van A de verwachting van de verdeling niet benaderen. Men heeft hier te doen met een extreem geval van de kwestie, die wij aangeroerd hebben aan het eind van par. 2.2: de verwachting van de verdeling is helemaal niet meer te benaderen door het nemen van steekproeven uit die verdeling. De gemiddelden van steekproeven zullen dan onderling ook veel sterker uiteenlopen dan bij "gewone" wh-verdelingen het geval is, waar de gemiddelden meestal dicht bij de verwachting liggen. Indien de verwachting van een verdeling oneindig groot is, kan men dan ook beter zeggen dat de verdeling geen verwachting bezit.

Opgave:

2.4.b. Bereken de verwachting van de wh-verdeling van het aantal ogen geworpen met twee zuivere dobbelstenen (Zie opgave 1.9.a.).

2 5 De verwachting van de continue verdeling.

De verwachting van een willekeurige continue verdeling kan voorgesteld worden door:

$$(2.5.1) \qquad \mathcal{E}\,\underline{x} = \int_{-\infty}^{+\infty} x\, f(x)\, dx \quad ,$$

waarin $f(x)$ de verdelingsdichtheid van $\underline{x}$ voorstelt. De integraal in het rechterlid kan wederom als een oppervlakte beschouwd worden. Men tekent daartoe de functie $g(x) = x f(x)$. Omdat $f(x)$ nooit negatief is, zal deze functie positief of nul zijn voor $x>0$ en negatief of 0 voor $x<0$; zij verloopt dus voor $x>0$ boven of langs de x-as; de oppervlakte tussen de kromme en de x-as vanaf $x=0$ stellen wij nu voor door O_+ ; voor $x<0$ verloopt $g(x)$ beneden of langs de x-as; de oppervlakte tussen de kromme en de x-as voor $x<0$ noemen wij O_-. Er geldt dan:

$$\int_{-\infty}^{+\infty} x\, f(x)\, dx = O_+ - O_- \quad .$$

Wij kunnen hieruit bijvoorbeeld direct zien, dat het gemiddelde van een continue verdeling 0 is, als zij symmetrisch is ten opzichte van 0, hetgeen wil zeggen: als $f(x) = f(-x)$ (Zie fig. 2.1 blz. 33).

Immers nu geldt $g(x) = x f(x) = x f(-x) = -g(-x)$, dus het gedeelte van de kromme $g(x)$ links van 0 is vanuit het gedeelte rechts van 0 te verkrijgen door dit ten opzichte van het punt 0 over 180° te draaien. Aangezien bij deze draaiing het oppervlak tussen de kromme en de x-as ongewijzigd blijft geldt $O_+ = O_-$, dus is $\int_{-\infty}^{+\infty} x\, f(x)\, dx = 0$ ..

Opgave:
2.5.a: Geldt de hier bewezen stelling ook voor discrete verdelingen?

In bepaalde gevallen is het mogelijk de oppervlakten O_+ en O_- exact te bepalen; in vele andere gevallen zullen benaderingen moeten worden gebruikt. Deze benaderingen bestaan in principe daarin, dat de met O_+ en O_- corresponderende vlakdelen door verticale lijnen in strookjes worden verdeeld en de oppervlakten van deze strookjes geschat worden.

Wij zullen op de techniek van de berekening van het gemiddelde van een continue verdeling niet verder ingaan, en er hier slechts op wijzen, dat het gemiddelde van een continue verdeling evenals dan van een discrete verdeling in het algemeen geschat kan worden door het gemiddelde van een lange reeks waarnemingen. Enige eigenschappen van het gemiddelde zullen in een volgende paragraaf volgen.

Nu wij het gemiddelde zowel van continue als discrete verdelingen gedefiniëerd hebben, kunnen wij ook de massatheoretische interpretatie geven van dit begrip (verg. par. 1.7) Deze is de volgende: als de massa 1 over een rechte lijn verdeeld is als aangegeven wordt door de wh-verdeling van $\underline{x}$, dan is $\mathcal{E}\underline{x}$ het moment van deze massa t.o.v. het punt $x=0$. Men noemt $\mathcal{E}\underline{x}$ daarom ook wel het eerste moment van $\underline{x}$. Tevens blijkt dat $\mathcal{E}\underline{x}$ de coordinaat voorstelt van het aangrijpingspunt van de resultante der zwaartekrachten, die op een lijn met massaverdeling $f(x)$ werken.

Opgave:
2.5.b. $\underline{x}$ is homogeen verdeeld tussen a en b (d.w.z $f(x)=0$ als $x \leqq a$, $f(x)=\frac{1}{b-a}$ als $a<x\leqq b$, $f(x)=0$ als $x>b$). Toon aan dat $\mathcal{E}\underline{x}=\frac{a+b}{2}$ is voor de gevallen $a<b<0$, $a<0<b$ en $0<a<b$

2.6 Eigenschappen van de verwachting van een verdeling.

Wij zullen hieronder de verwachting van een willekeurige stochastische variabele steeds voorstellen door $\mathcal{E}\underline{x}$. Als $\underline{x}$ een discrete verdeling heeft zullen wij aannemen, dat $\underline{x}$ met de whn p_i de waarden x_i aanneemt ($i=1,2,\ldots$ etc.) $\sum_i p_i=1$ dan is dus: $\mathcal{E}\underline{x}=\sum_i p_i x_i$. Als $\underline{x}$ een continue verdeling heeft duiden wij de verdelingsdichtheid aan met $f(x)$, zodat dan $\mathcal{E}\underline{x}=\int_{-\infty}^{+\infty} x f(x)\,dx$
Wij geven hieronder nu een tweetal eigenschappen van het gemiddelde, die zowel voor discrete als voor continue verdelingen gelden. Wij zullen deze eigenschappen alleen voor discrete verdelingen bewijzen.

Eigenschap 1: Als c een constante is, geldt:

(2 6.1) $\mathcal{E}(\underline{x} + c) = \mathcal{E}\underline{x} + c$

Voorbeeld: Laat $\underline{x}$ het aantal ogen zijn, geworpen met een zuivere dobbelsteen; dus $\mathcal{E}\underline{x} = 3\frac{1}{2}$ (zie 2.4) Laat nu A en B een spel spelen, waarbij A $i - 2$ gulden van B ontvangt, als er i ogen geworpen worden; met dien verstande dat ontvangen van een negatief bedrag overeenkomt met betalen van het tegengestelde Als dus $i = 1$ betaalt A f1,-- aan B, als $i = 2$ gebeurt er niets, als $i = 3$ betaalt B f1,-- aan A etc De winst van A is dus steeds 2 minder dan het aantal geworpen ogen en kan dus voorgesteld worden als $\underline{x} - 2$. Volgens bovenstaande stelling geldt dus voor de verwachting van de winst van A:

$$\mathcal{E}(\underline{x} - 2) = \mathcal{E}\underline{x} - 2 = 3\tfrac{1}{2} - 2 = 1\tfrac{1}{2}.$$

Bewijs van eigenschap 1:

Stel $\underline{y} = \underline{x} + c$, en $y_i = x_i + c$ dan is:

$$\mathcal{E}\underline{x} = \mathcal{E}\underline{y} = \sum_i p_i y_i = \sum_i p_i (x_i + c) = \sum_i p_i x_i + c \sum_i p_i .$$

Er geldt echter steeds $\sum_i p_i = 1$, dus:

$$\sum_i p_i x_i + c \sum_i p_i = \mathcal{E}\underline{x} + c ,$$

waarmee de eigenschap bewezen is.

Deze eigenschap volgt ook gemakkelijk uit de massatheoretische interpretatie. De massa-verdeling corresponderend met de wh-verdeling van $\underline{x} + c$ is t.o.v. de verdeling van $\underline{x}$ over een afstand c verschoven; het zwaartepunt $\mathcal{E}(\underline{x} + c)$ van de verdeling van $\underline{x} + c$ moet dus ook door een verschuiving over dezelfde afstand uit het zwaartepunt $\mathcal{E}\underline{x}$ van $\underline{x}$ verkregen worden.

Opgave:

2.6.a Als men in de loterij, vermeld in par 2.4 de waarde van de 12 prijzen met f 1000,-- verhoogt, moet dan volgens deze stelling ook de prijs van een lot met f 1000,-- verhoogd worden?

Eigenschap 2: Als a een constante is, geldt:

(2.6.2) $\mathcal{E}(a\underline{x}) = a\,\mathcal{E}\underline{x}$

Voorbeeld: Bij de loterij beschreven in par 2.4, (met gemiddelde winst voor een speler: $\mathcal{E}\underline{x} = 0{,}35$) wordt het bedrag van alle prijzen met 10 vermenigvuldigd. Hierbij wordt ook de winst van de

"nieten" met 10 vermenigvuldigd ($10 \times 0 = 0$). Alle waarden die de stochastische variabele kan aannemen worden dus met 10 vermenigvuldigd. Dus geldt volgens bovenstaande stelling:

$$\mathcal{E}(10\,\underline{x}) = 10\,\mathcal{E}\,\underline{x} = 3{,}50 .$$

De prijs van een lot moet dus minstens f 3,50 bedragen.

Bewijs van eigenschap 2:

Stel $\underline{y} = a\,\underline{x}$ en $y_i = a x_i$ dan geldt:

$$\mathcal{E}\,a\underline{x} = \mathcal{E}\,\underline{y} = \sum_i p_i y_i = \sum_i p_i\, a\, x_i = a \sum_i p_i x_i = a\,\mathcal{E}\,\underline{x}.$$

De overgang van $\underline{x}$ op $a\,\underline{x}$ komt overeen met een gelijkmatige uitrekking (als $a > 1$) of samenpersing (als $a < 1$) van de massaverdeling van $\underline{x}$ over een rechte lijn, zó dat de afstand van de massapunten tot het punt O met het bedrag a wordt vermenigvuldigd. Het is duidelijk dat dit dan ook met het zwaartepunt moet gebeuren.

Opgaven:

2.6.b Als de verwachting van $\underline{x}$ gelijk is aan μ , waaraan is dan de verwachting van $a\underline{x} + c$ gelijk, als a ën c gegeven constanten zijn?

2.6.c Als $\underline{x}$ symmetrisch verdeeld is t.o.v. O, wat is dan de verwachting van $a\underline{x}$, van $\underline{x} + c$ en van $a\underline{x} + c$? (vergelijk par. 2.5). Wat volgt hieruit voor de verwachting van een variabele die symmetrisch verdeeld is t.o.v. $x = c$?

Ad blz. 30.

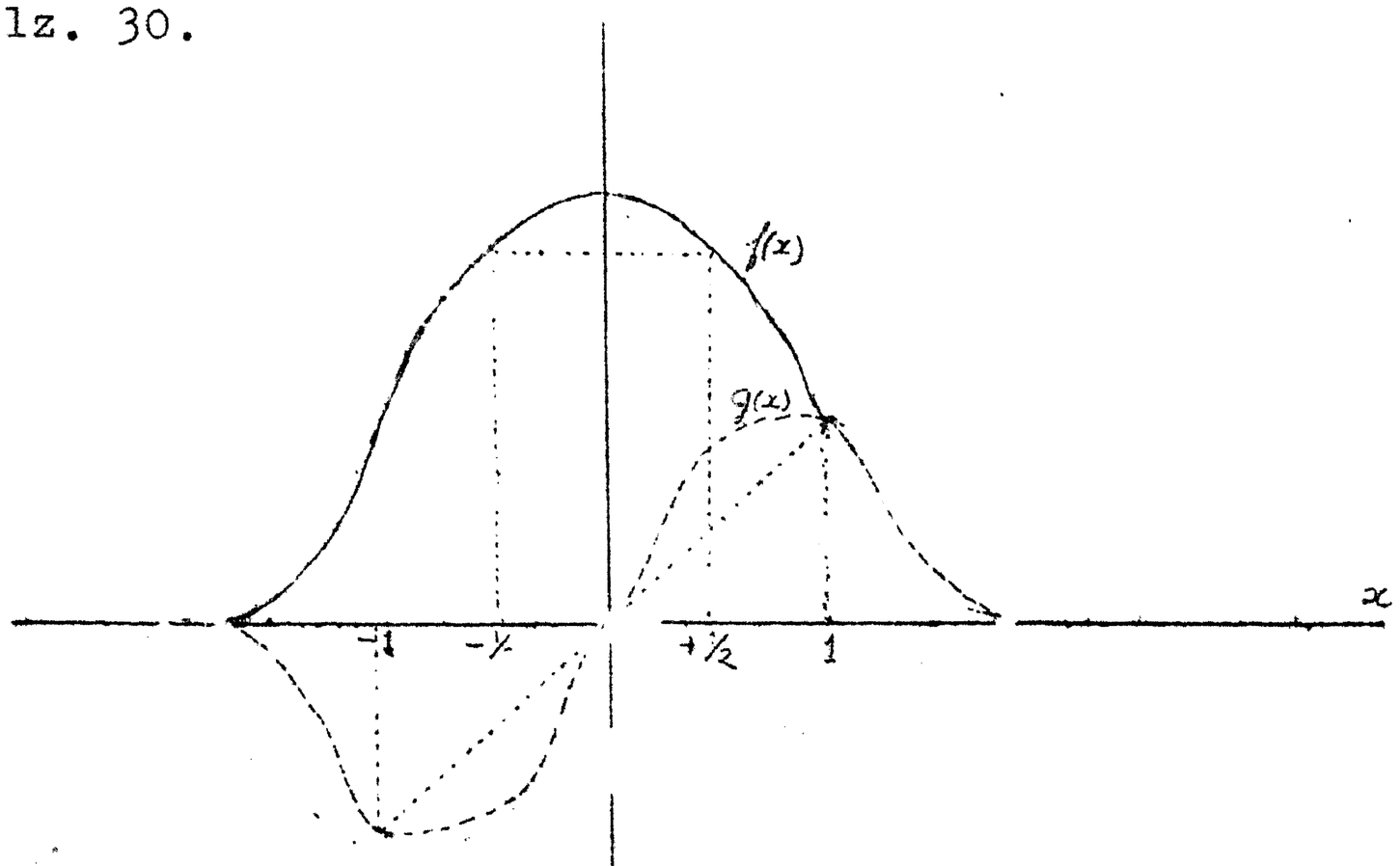


Fig 2.1

$f(x)$ en $g(x) = x f(x)$ bij een t.o.v. O symmetrische verdeling (vgl. par 2.1)

2.7. Het begrip verwachting bij tweedimensionale verdelingen.

Wij onderstellen, dat de stochastische variabelen $\underline{x}$ en $\underline{y}$ simultaan een discrete tweedimensionale verdeling hebben b.v. zoals gegeven in figuur 2.2. Hierin is

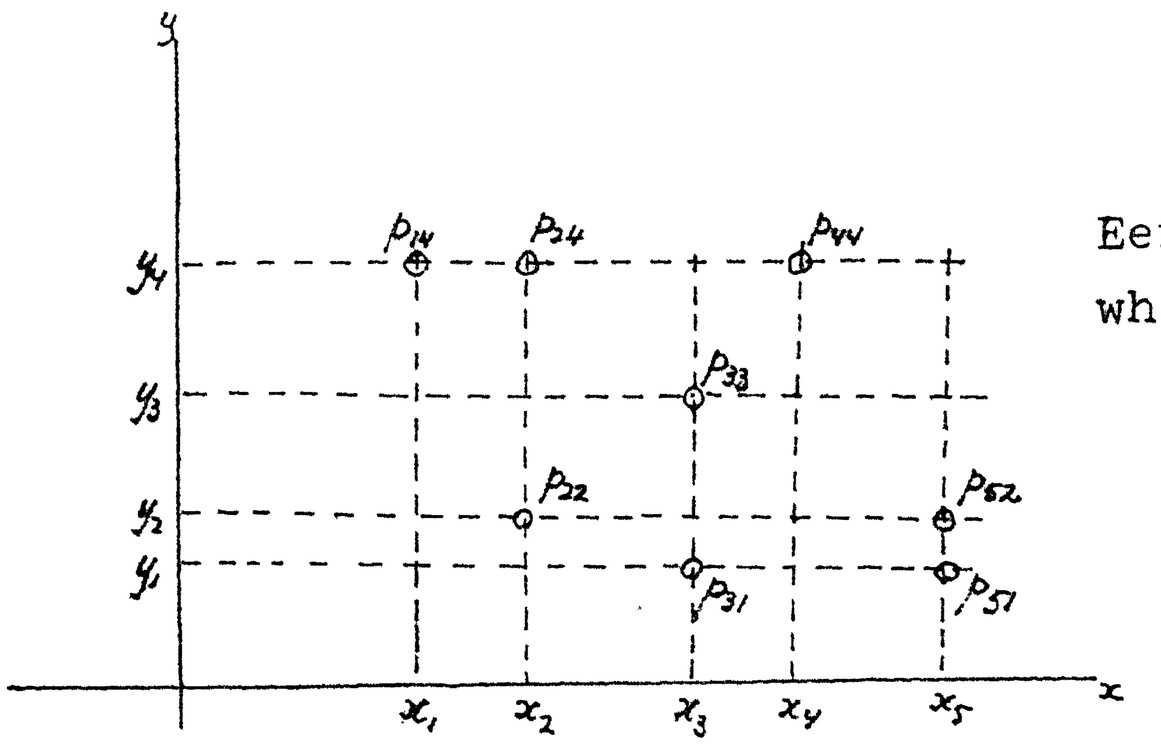


Fig. 2.2
Een tweedimensionale wh-verdeling.

de kans $P[\underline{x} = x_i \text{ en } \underline{y} = y_j]$ aangegeven door p_{ij} . Uiteraard moet gelden: $\sum_i \sum_j p_{ij} = 1$.

De verwachting van $\underline{x}$ zal in dit geval worden:

$$\mathcal{E}\,\underline{x} = \sum_{i=1}^{5} x_i\, P[\underline{x} = x_i].$$

Dit is dus het gemiddelde van de marginale verdeling van $\underline{x}$. In het geval van figuur 2.2 vinden wij:

$$\mathcal{E}\,\underline{x} = x_1 p_{14} + x_2 (p_{22} + p_{24}) + x_3 (p_{31} + p_{33}) + x_4 p_{44} + x_5 (p_{51} + p_{52}).$$

Als wij aan de combinaties (x_i, y_j), die in de verdeling niet voorkomen, whn $p_{ij} = 0$ toekennen, kunnen wij $\mathcal{E}\underline{x}$ ook in de volgende vorm schrijven:

$$\mathcal{E}\,\underline{x} = \sum_{i=1}^{5} x_i \sum_{j=1}^{4} p_{ij} = \sum_{i=1}^{5} \sum_{j=1}^{4} p_{ij}\, x_i .$$

Evenzo vinden wij:

$$\mathcal{E}\,\underline{y} = y_1 (p_{31} + p_{51}) + y_2 (p_{22} + p_{52}) + y_3 p_{33} + y_4 (p_{14} + p_{24} + p_{44}) =$$
$$= \sum_{j=1}^{4} y_j \sum_{i=1}^{5} p_{ij} = \sum_{i=1}^{5} \sum_{j=1}^{4} p_{ij}\, y_j .$$

Algemeen zullen wij, als $\underline{x}$ de waarden $x_1, \ldots, x_m$ aanneemt en $\underline{y}$ de waarden $y_1, \ldots, y_n$ en als p_{ij} gedefinieerd wordt als boven, vinden:

$$(2.7.1) \qquad \mathcal{E}\underline{x} = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij}\, x_i \qquad en$$

$$(2.7.2) \qquad \mathcal{E}\underline{y} = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij}\, y_j .$$

Als bij een continue tweedimensionale verdeling de marginale verdelingsdichtheid van $\underline{x}$ wordt voorgesteld door $f_1(x)$ geldt:

$$(2.7.3) \qquad \mathcal{E}\underline{x} = \int_{-\infty}^{+\infty} x\, f_1(x)\, dx .$$

Deze kan dus als $f_1(x)$ bekend is, berekend worden zoals geschetst is in par. 2.6. Als $f_2(y)$ de marginale verdelingsdichtheid van $\underline{y}$ is geldt analoog:

$$(2.7.4) \qquad \mathcal{E}\underline{y} = \int_{-\infty}^{+\infty} y\, f_2(y)\, dy .$$

$\mathcal{E}\underline{x}$ en $\mathcal{E}\underline{y}$ kunnen ook in de verdelingsdichtheid $f(x,y)$ uitgedrukt worden; er geldt voor $\mathcal{E}\underline{x}$:

$$(2.7.5) \qquad \mathcal{E}\underline{x} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x\, f(x,y)\, dx\, dy .$$

Deze dubbele integraal stelt het verschil van twee inhouden voor nl.: de inhoud tussen het oppervlak voorgesteld door $x f(x,y)$ en het (x,y)-vlak voor $x > 0$ verminderd met de overeenkomstige inhoud voor $x < 0$.

Omdat $\mathcal{E}\underline{x}$ en $\mathcal{E}\underline{y}$ beschouwd kunnen worden als de gemiddelden van de marginale verdelingen van $\underline{x}$ resp. $\underline{y}$, gelden hiervoor de eigenschappen die in par. 2.6 beschreven zijn. Ook kunnen wij, evenals in par. 2.5. gedaan is, aantonen dat $\mathcal{E}\underline{x} = 0$ is, als de verdeling symmetrisch is t.o.v. $x = 0$, dus t.o.v. de y-as; dit betekent, dat voor iedere y geldt:

$$P[\underline{x} = x \text{ én } \underline{y} = y] = P[\underline{x} = -x \text{ én } \underline{y} = y] \quad \text{resp.:}$$

$$f(x,y) = f(-x,y).$$

Wij kunnen bij een tweedimensionale verdeling de verwachting van $\underline{x}+y$ uitdrukken in de verwachtingen van $\underline{x}$ en y. Om $\mathcal{E}(\underline{x}+y)$ te verkrijgen moeten wij voor ieder punt (x_i, y_j) van de verdeling, de kans $p_{ij} = P[\underline{x} = x_i \text{ en } y = y_j]$ vermenigvuldigen met $x_i + y_j$ en de producten sommeren. Er geldt dus:

$$\mathcal{E}(\underline{x}+y) = \sum_i \sum_j (x_i + y_j) p_{ij} = \sum_i \sum_j x_i p_{ij} + \sum_i \sum_j y_j p_{ij} = \mathcal{E}\underline{x} + \mathcal{E}y.$$

We hebben hiermee de volgende belangrijke eigenschap aangetoond:

Eigenschap 3: De verwachting van de som van twee stochastische variabelen is gelijk aan de som van hun verwachtingen, of in formulevorm:

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

Deze eigenschap geldt zowel als $\underline{x}$ en y onafhankelijk als wanneer deze grootheden afhankelijk zijn.

Voorbeelden:

$\underline{x}$ en y onafhankelijk: De verwachting van de som van het aantal ogen geworpen met twee zuivere dobbelstenen is $3\frac{1}{2} + 3\frac{1}{2} = 7$. Dit is ook gemakkelijk in te zien als men de verdeling van $\underline{x}+y$ berekend heeft (zie opgave 1.9.b).

$\underline{x}$ en y afhankelijk: In een vaas bevinden zich 5 loten, genummerd 1,..,5. Een persoon trekt uit de vaas 2 loten zònder terug-legging. Gevraagd wordt de verwachting van de som der op de getrokken loten voorkomende nummers.

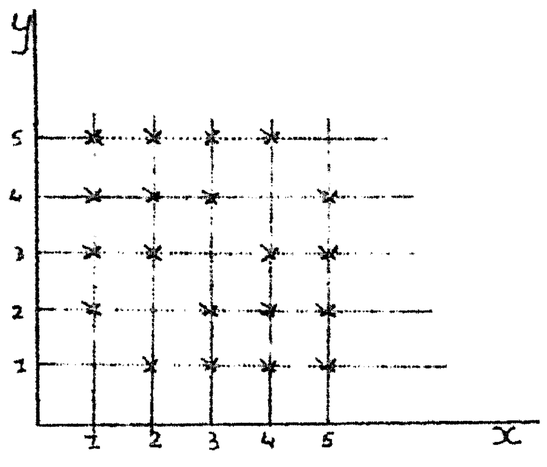


Fig.2.3.
Trekking van twee loten uit 5 zonder terugleggíng.

In fig.2.3 hebben wij de hier voorkomende mogelijkheden door kruisjes aangegeven; $\underline{x}$ stelt voor het nummer van het eerst getrokken lot, y het nummer van het tweede lot. Vanwege het feit, dat getrokken wordt zonder terugleggíng vervallen de mogelijkheden met $x = y$. Er blijven 20 mogelijkheden over, die alle even waarschijnlijk zijn. Men ziet verder dat de marginale verdelingen van $\underline{x}$ en y hetzelfde zijn, en wel dat met kansen $\frac{1}{5}$ de waarden $1,2,\ldots,5$ worden aangenomen. Er geldt dus: $\mathcal{E}\underline{x} = \mathcal{E}y = 3$ dus volgens

eigenschap 3: $\mathcal{E}(\underline{x}+y)=6$.

Indien wij nu trekkingen van 2 loten beschouwen met teruglegging, worden de punten $(x=1, y=1)$ t/m $(x=5, y=5)$ in figuur 2.3 eveneens mogelijke trekkingsresultaten: $\underline{x}$ en y zijn dan onafhankelijk, er zijn nu 25 even waarschijnlijke mogelijkheden. De marginale verdelingen van $\underline{x}$ en y blijven echter ongewijzigd, zodat ook hier geldt: $\mathcal{E}(\underline{x}+y)=6$. Wij zien hier dus dat het voor de verwachtingen van $\underline{x}$ en y voor de verwachting van $\underline{x}+y$ geen verschil maakt, of de twee trekkingen zonder of met teruglegging geschieden. Men dient hierbij wel te bedenken, dat het gaat om de verwachtingen van $\underline{x}$ en y vóórdat de eerste trekking verricht is. Als men de verwachting van y gaat bepalen wanneer de trekking van $\underline{x}$ reeds geschiedt is, maakt het wel verschil. Indien b.v. getrokken is $x=4$, en het getrokken lot wordt niet teruggelegd, blijven er voor de tweede trekking de mogelijkheden 1,2,3 en 5 over, zodat de verwachting is: $\frac{1}{4}(1+2+3+5)=2\frac{3}{4}$. Indien er echter met teruglegging getrokken wordt, is de tweede trekking volkomen onafhankelijk van de eerste en blijft de verwachting daarvan dus 3.

Eigenschap 1 in par.2.6 kan als een bijzonder geval van eigenschap 3 beschouwd worden. Er geldt immers volgens eigenschap 3:

$$\mathcal{E}(\underline{x}+c)=\mathcal{E}\,\underline{x}+\mathcal{E}\,c\ .$$

c is een constante, en kan dus beschouwd worden als een stochastische variabele, die met wh 1 de waarde c aanneemt. De verwachting van een dergelijke variabele is c. Dus $\mathcal{E}\,c=c$, waaruit eigenschap 1 volgt.

2.8 Consequenties van eigenschap 3.

Eigenschap 3 kan gemakkelijk uitgebreid worden tot meer dan twee stochastische variabelen. Laten wij bijvoorbeeld 3 grootheden $\underline{x}$, y en $\underline{z}$ beschouwen. Zij nu $\underline{u}=\underline{x}+y$. Dan geldt volgens eigenschap 3:

$$\mathcal{E}(\underline{x}+y+\underline{z})=\mathcal{E}(\underline{u}+\underline{z})=\mathcal{E}\,\underline{u}+\mathcal{E}\,\underline{z} \quad \text{en}$$

$$\mathcal{E}\,\underline{u}=\mathcal{E}(\underline{x}+y)=\mathcal{E}\,\underline{x}+\mathcal{E}\,y\ .\ \text{Dus}$$

$$\mathcal{E}(\underline{x}+y+\underline{z})=\mathcal{E}\,\underline{x}+\mathcal{E}\,y+\mathcal{E}\,\underline{z}\ ,$$

zo voortgaande vinden wij:

Eigenschap 4: De verwachting van de som van een aantal stochastische variabelen is gelijk aan de som van hun ver-

wachtingen. In formule luidt deze stelling:

$$(2.8.1) \qquad \mathcal{E}\left(\sum_{i=1}^{n} \underline{x}_i\right) = \sum_{i=1}^{n} \mathcal{E}\,\underline{x}_i \ .$$

Ook hier is het niet nodig dat de variabelen $\underline{x}_i$ onderling onafhankelijk zijn.

Met behulp van eigenschap 4 kunnen wij bijvoorbeeld gemakkelijk het gemiddelde van een binomiale verdeling uitrekenen. Zoals wij reeds gezien hebben in par.1.3 heeft het aantal successen, in een reeks van n onafhankelijke experimenten, ieder met kans p op succes een binomiale verdeling. Als wij nu voor het i^{de} experiment afzonderlijk een stochastische variabele $\underline{x}_i$ definiëren, die 1 is als er een succes optreedt en 0 zo dit niet gebeurt, dan zal

$$\underline{y} = \sum_{i=1}^{n} \underline{x}_i$$

juist de stochastische variabele voorstellen, die het aantal successen bij de n experimenten weergeeft. Omdat de kans op een succes p is, geldt dus voor $\underline{x}_i$:

$$P\left[\underline{x}_i = 1\right] = p \quad , \quad P\left[\underline{x}_i = 0\right] = 1 - p \ .$$

De verwachting van $\underline{x}_i$ wordt dus:

$$\mathcal{E}\,\underline{x}_i = 1.p + 0.(1-p) = p \ .$$

Dit geldt voor iedere waarde van i $(i = 1, \ldots, n)$.

Volgens eigenschap 4 geldt nu:

$$\mathcal{E}\,\underline{y} = \sum_{i=1}^{n} \mathcal{E}\,\underline{x}_i = \sum_{i=1}^{n} p = np \ .$$

Het gemiddelde van een binomiale verdeling voor n experimenten met een kans p op succes wordt dus gelijk aan np. (Verg. opgave 2.4 a).

Als we eigenschap 4 combineren met eigenschap 2 vinden we, als $a_1, \ldots, a_n$ constanten zijn:

$$(2.8.2) \qquad \mathcal{E}\left(\sum_{i=1}^{n} a_i\,\underline{x}_i\right) = \sum_{i=1}^{n} \mathcal{E}\,a_i\,\underline{x}_i = \sum_{i=1}^{n} a_i\,\mathcal{E}\,\underline{x}_i \ .$$

We kunnen dus de verwachting van een zogenaamde <u>lineaire combinatie</u> $\sum_{i=1}^{n} a_i\,\underline{x}_i$ van de stochastische variabelen $\underline{x}_i$ in de verwachtingen van deze variabelen uitdrukken.

Een bijzonder geval hiervan is b.v. de verwachting van het <u>verschil</u> van twee variabelen:

$$(2.8.3) \qquad \mathcal{E}(\underline{x} - \underline{y}) = \mathcal{E}(\underline{x} + (-)\,\underline{y}) = \mathcal{E}\,\underline{x} - \mathcal{E}\,\underline{y} \ .$$

Een ander voorbeeld hebben wij in het rekenkundig gemiddelde van een aantal stochastische grootheden:

$$(2.8.4) \qquad \mathcal{E}\,\frac{1}{n}\sum_{i=1}^{n} \underline{x}_i = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}\,\underline{x}_i \ .$$

Als nu alle $\underline{x}_i$ dezelfde wh-verdeling hebben, met $\mathcal{E}\,\underline{x}_i = \mu$ geldt dus:

(2.8.5) $$\mathcal{E}\,\frac{1}{n}\sum_{i=1}^{n}\underline{x}_i = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n\mu}{n} = \mu .$$

De verwachting van het rekenkundig gemiddelde der $\underline{x}_i$ is dan dus gelijk aan de verwachting van die grootheden zelf. Dit geval doet zich in de praktijk voor, als wij een steekproef nemen van meerdere waarnemingen van een zelfde stochastische variabele, b.v. verschillende metingen van een zelfde physische grootheid. Bij herhaling van de steekproef zullen in het algemeen andere steekproefwaarden optreden. De waarnemingen van een steekproef kunnen dus ook als stochastische variabelen beschouwd worden. Als wij de i^{de} waarneming van een steekproef van n aanduiden met $\underline{x}_i$ zullen de stochastische grootheden $\underline{x}_1,\ldots,\underline{x}_n$ dezelfde wh-verdeling hebben, namelijk de verdeling der waargenomen variabele $\underline{x}$. Uit het bovenstaande blijkt nu, dat de verwachting van het rekenkundig gemiddelde van een aantal waarnemingen van een stochastische variabele $\underline{x}$ gelijk is aan de verwachting van $\underline{x}$ zelf. Het is hiervoor niet nodig, dat de steekproef uit onderling onafhankelijke waarnemingen bestaat; zo mag het b.v. een steekproef zonder teruglegging zijn.

Voorbeeld: In een kelder bevinden zich 1000 appels met een totaal gewicht van 150 kilo. Het gemiddelde gewicht van een appel is dan dus 150 gram. Wij kunnen dit gemiddelde gewicht op de volgende wijzen door het nemen van steekproeven benaderen:

a. We trekken een aantal keren een appel, waarvan we het gewicht bepalen en die wij vervolgens terugleggen. Als de opeenvolgende trekkingen aselect zijn, d.w.z. voldoen aan de eisen van gelijkwaardigheid en onafhankelijkheid (zie: Cursus, Toegepaste Statistiek I, par.4.1) zal het gemiddelde van de bepaalde gewichten volgens de wet der grote getallen naderen tot het gemiddelde gewicht der appels van de partij, dus 150 gram.

b. Wij trekken een aantal keren een greep van b.v. 5 appels en nemen dus steekproeven van 5 zonder teruglegging. Wij bepalen dan het gemiddelde gewicht van ieder van die steekproeven, d.w.z. wij wegen de 5 appels samen en delen de uitkomst door 5. De grepen van 5 appels worden na de bepaling van het gewicht steeds teruggelegd, vóór wij de vol-

gende trekking verrichten. Als de opeenvolgende grepen weer aselect genomen worden zal het gemiddelde van de gemiddelde appelgewichten der afzonderlijke steekproeven eveneens naderen tot het gemiddelde gewicht der appels van de partij. Beide beschreven methoden voeren dus, volgens de tot nu toe behandelde theorie, tot het doel: het bepalen van het gemiddelde gewicht van de appels in de partij. Wij kunnen nog niet uitmaken, welke methode het kleinste aantal experimenten vergt. Dit is pas mogelijk als wij de spreiding van de verdeling beschouwen (Zie par.2.14).

Opgave:

2.8.a. Aan ieder der getallen $1,2,\ldots,n$ wordt met kans $\frac{1}{2}$ een positief of een negatief teken toegekend. De dan verkregen getallen worden gesommeerd. Gevraagd wordt de verwachting van deze som. (Aanwijzing: deze verwachting kan op een dergelijke wijze bepaald worden als boven gegeven voor de binomiale verdeling).

2.9. De gereduceerde variabele.

Het blijkt in de practijk zeer gemakkelijk te zijn een afzonderlijke naam en een afzonderlijk symbool te hebben voor een stochastische variabele, verminderd met haar verwachting. Wij noemen dit de gereduceerde variabele; de gereduceerde variabele behorende bij de stochastische variabele $\underline{x}$ duiden we aan met $\tilde{\underline{x}}$. Er geldt dus:

$$\tilde{\underline{x}} = \underline{x} - \mathcal{E}\,\underline{x} . \qquad (2.9.1)$$

$\tilde{\underline{x}}$ is uiteraard ook stochastisch; haar verdeling wordt uit die van $\underline{x}$ verkregen door verschuiving over de constante afstand $\mathcal{E}\,\underline{x}$. (Zie par.2.6) Volgens eigenschap 1 geldt nu:

$$\mathcal{E}\,\tilde{\underline{x}} = 0 . \qquad (2.9.2)$$

Soms wordt bij een steekproef $x_1, \ldots, x_n$ ook van gereduceerde waarden gesproken; dit zijn dan de steekproefwaarden verminderd met het steekproefgemiddelde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, zij worden wel aangeduid met $\breve{x}_i$; er geldt dus:

$$\breve{x}_i = x_i - \bar{x} . \qquad (2.9.3)$$

Als het een steekproef betreft van dezelfde stochastische variabele $\underline{x}$, en als we de steekproefwaarden opvatten als stochastische variabelen, zoals gedaan is in par. 2.8, geldt:

$$\mathcal{E}\,\underline{x}_i = \mathcal{E}\,\bar{\underline{x}} = \mathcal{E}\,\underline{x} .$$

Dus: $\mathcal{E}\ \check{\underline{x}}_i = 0$.

De gereduceerde variabele $\check{\underline{x}}$ en de gereduceerde steekproefwaarde $\check{\underline{x}}_i$ hebben dus beide dezelfde verwachting. Maar zij hebben niet dezelfde verdeling; $\check{\underline{x}}$ heeft een verdeling die uit die van $\underline{x}$ kan worden afgeleid door verschuiving over een constante $\mathcal{E}\,\underline{x}$, doch $\check{\underline{x}}_i$ is ontstaan door samenstelling van twee stochastische variabelen $\underline{x}_i$ en $\bar{\underline{x}}$.

2.10. Het 2° moment.

We zijn nu enigszins vertrouwd geraakt met verwachtingen van stochastische variabelen, van sommen en verschillen van twee stochastische variabelen, alsmede van lineaire combinaties van dergelijke grootheden. Wij beschouwen nu een uitdrukking van de gedaante $\mathcal{E}\,\underline{x}^2$, de verwachting van $\underline{x}^2$, meestal genoemd het 2° moment van $\underline{x}$. De definitie hiervan is geheel analoog aan die van $\mathcal{E}\,\underline{x}$.

Daarvoor gold: $\mathcal{E}\,\underline{x} = \sum_{i=1}^{n} p_i\, x_i^1$ of $\int_{-\infty}^{+\infty} x^1 f(x)\,dx$.

We definieren nu het 2° moment als:

(2.10.1) $\mathcal{E}\,\underline{x}^2 = \sum_{i=1}^{n} p_i\, x_i^2$ bij een discrete verdeling en

(2.10.2) $\mathcal{E}\,\underline{x}^2 = \int_{-\infty}^{+\infty} x^2 f(x)\,dx$, bij een continue verdeling.

Opmerking:

Een andere methode om $\mathcal{E}\,\underline{x}^2$ te definieren is de volgende: bepaal eerst de verdeling van $\underline{x}^2$ en neem daarvan het gemiddelde. Deze definitie leidt tot hetzelfde resultaat als de bovenstaande. Wij kunnen dit aan een eenvoudig voorbeeld laten zien:

Laat $\underline{x}$ de waarden -3, -2, -1, 0, 2, 3 en 5 met kansen: $P[\underline{x} = -3]$, $P[\underline{x} = -2]$ etc. aannemen. Volgens definitie (2.10.1) wordt nu

$$\mathcal{E}\,\underline{x}^2 = 9\,P[\underline{x} = -3] + 4\,P[\underline{x} = -2] + 1\,P[\underline{x} = -1] + 0\,P[\underline{x} = 0] + 4\,P[\underline{x} = 2] + 9\,P[\underline{x} = 3] + 25\,P[\underline{x} = 5].$$

De verdeling van $\underline{x}^2$ wordt:

$$\begin{aligned} P[\underline{x}^2 = 25] &= P[\underline{x} = 5] \\ P[\underline{x}^2 = 9] &= P[\underline{x} = -3] + P[\underline{x} = 3] \\ P[\underline{x}^2 = 4] &= P[\underline{x} = -2] + P[\underline{x} = 2] \\ P[\underline{x}^2 = 1] &= P[\underline{x} = -1] \\ P[\underline{x}^2 = 0] &= P[\underline{x} = 0]. \end{aligned}$$

De verwachting van deze verdeling wordt dezelfde uitdrukking als boven, alleen worden nu de termen met dezelfde coëfficiënten bij elkaar genomen.

Dus: $\mathcal{E}\ \breve{x}_i = 0$.

De gereduceerde variabele $\breve{x}$ en de gereduceerde steekproefwaarde $\breve{x}_i$ hebben dus beide dezelfde verwachting. Maar zij hebben niet dezelfde verdeling; $\breve{x}$ heeft een verdeling die uit die van x kan worden afgeleid door verschuiving over een constante $\mathcal{E}\,x$, doch $\breve{x}_i$ is ontstaan door samenstelling van twee stochastische variabelen x_i en $\bar{x}$.

2.10. Het 2° moment.

We zijn nu enigszins vertrouwd geraakt met verwachtingen van stochastische variabelen, van sommen en verschillen van twee stochastische variabelen, alsmede van lineaire combinaties van dergelijke grootheden. Wij beschouwen nu een uitdrukking van de gedaante $\mathcal{E}\,x^2$, de verwachting van x^2, meestal genoemd het 2° moment van x. De definitie hiervan is geheel analoog aan die van $\mathcal{E}\,x$.

Daarvoor gold: $\mathcal{E}\,x = \sum_{i=1}^{n} p_i x_i^1$ of $\int_{-\infty}^{+\infty} x^1 f(x)\,dx$.

We definieren nu het 2° moment als:

(2.10.1) $\mathcal{E}\,x^2 = \sum_{i=1}^{n} p_i x_i^2$ bij een discrete verdeling en

(2.10.2) $\mathcal{E}\,x^2 = \int_{-\infty}^{+\infty} x^2 f(x)dx$, bij een continue verdeling.

Opmerking:

Een andere methode om $\mathcal{E}\,x^2$ te definieren is de volgende: bepaal eerst de verdeling van x^2 en neem daarvan het gemiddelde. Deze definitie leidt tot hetzelfde resultaat als de bovenstaande. Wij kunnen dit aan een eenvoudig voorbeeld laten zien:

Laat x de waarden -3, -2, -1, 0, 2, 3 en 5 met kansen: $P[x=-3]$, $P[x=-2]$ etc. aannemen. Volgens definitie (2.10.1) wordt nu

$$\mathcal{E}\,x^2 = 9\,P[x=-3] + 4\,P[x=-2] + 1\,P[x=-1] + 0\,P[x=0] + 4\,P[x=2] + 9\,P[x=3] + 25\,P[x=5].$$

De verdeling van x^2 wordt:

$$\begin{aligned} P[x^2=25] &= P[x=5] \\ P[x^2=9] &= P[x=-3] + P[x=3] \\ P[x^2=4] &= P[x=-2] + P[x=2] \\ P[x^2=1] &= P[x=-1] \\ P[x^2=0] &= P[x=0]. \end{aligned}$$

De verwachting van deze verdeling wordt dezelfde uitdrukking als boven, alleen worden nu de termen met dezelfde coëfficiënten bij elkaar genomen.

Hierin is het draaimoment het moment van het koppel, d.w.z. $K \times a$ als K de grootte en a de afstand tussen de aangrijpingspunten van de koppelkrachten is. De hoekversnelling is de vermeerdering van de hoeksnelheid per seconde, die onder invloed van het koppel optreedt. De hoeksnelheid wordt op een analoge wijze gedefinieerd als de snelheid bij een rechtlijnige beweging en wordt uitgedrukt in hoekeenheden (graden of radialen) per seconde. Het traagheidsmoment is tenslotte de grootheid, die overeenkomt met de massa bij een rechtlijnige beweging.

Het blijkt uit de wet, dat bij eenzelfde koppel, de hoekversnelling groter wordt en dus de lijn gemakkelijker in beweging te krijgen is, naarmate het traagheidsmoment kleiner is. Dit traagheidsmoment wordt nu gedefinieerd door dezelfde formule als het 2^o moment van de met de massaverdeling corresponderende wh-verdeling.

Zoals men aan de formules (2.10.1) en (2.10.2) kan zien neemt het 2^o moment en dus het traagheidsmoment toe, als de massa geheel of gedeeltelijk verder van $x=o$ wordt verwijderd; het doet er daarom niet toe, of zij links of rechts van $x=o$ komt te liggen. Bij een onzuivere dobbelsteen waarbij de kans op $\underline{x}=2$ iets toegenomen is ten koste van de kans op $\underline{x}=1$; b.v.: $P[\underline{x}=2]=\frac{1}{5}$ en $P[\underline{x}=1]=\frac{2}{6}-\frac{1}{5}=\frac{2}{15}$, wordt het tweede moment

$$\mathcal{E}\,\underline{x}^2=\tfrac{2}{15}\cdot 1^2+\tfrac{1}{5}\cdot 2^2+\tfrac{1}{6}\left(3^2+4^2+5^2+6^2\right)=15\tfrac{4}{15}>15\tfrac{1}{6}.$$

Men kan de invloed van veranderingen van het traagheidsmoment door verschuivingen van de massa van het draaiende lichaam aardig illustreren door gebruik te maken van een andere bekende wet uit de mechanica: de wet van het behoud van arbeidsvermogen. Een draaiend lichaam heeft een arbeidsvermogen van beweging A dat gegeven wordt door:

$A = \frac{1}{2}$ Traagheidsmoment $\times$ (hoeksnelheid)^2.

(analoog aan de bekende wet $A=\frac{1}{2}mv^2$ bij de rechtlijnige beweging). Als er geen krachten op het lichaam werken is A constant; vergroting van het traagheidsmoment betekent dan dus verkleining van de snelheid. Een persoon, die draait op een kantoorkruk, kan dus zijn snelheid verminderen door zijn benen uit te strekken (verplaatsing van de massa van het steunpunt af, dus vergroting van het traagheidsmoment) en vermeerderen door intrekken van de benen. Andere toepassingen vindt men bij de techniek van kunstschaatsenrijders, etc.

Opgave

2.10.a: Bereken het tweede moment van de som van het aantal ogen geworpen met twee zuivere dobbelstenen (Verg.opgave 1.9.a).

Als men het tweede moment van $\underline{x}$ uit een steekproef moet schatten, gaat men te werk op een wijze die volkomen analoog is aan hetgeen gebeurt bij het schatten van de verwachting. Men bepaalt dan het rekenkundig gemiddelde van de kwadraten der steekproefwaarden. De schatting wordt dus:

$\frac{1}{n}\sum_{i=1}^{n}\underline{x}_i^2$, als $x_1, \ldots\ldots, x_n$ de steekproefwaarden voorstellen. Deze schatting zal, als de omvang van de steekproef toeneemt, volgens de wet van de grote getallen in het algemeen steeds dichter tot het tweede moment van de verdeling naderen.

2.11. Variantie en spreiding.

In de statistiek interesseert ons speciaal het tweede moment van de gereduceerde variabele. Als $\widetilde{\underline{x}}$ de gereduceerde variable van $\underline{x}$ voorstelt, wordt $\mathcal{E}\,\widetilde{\underline{x}}^2$ de variantie van $\underline{x}$ genoemd. Deze grootheid wordt ook wel met $Var\,\underline{x}$ of $\sigma^2\{\underline{x}\}$ aangeduid. Er geldt dus:

(2.11.1) $$Var\,\underline{x} = \sigma^2\{\underline{x}\} = \mathcal{E}\,\widetilde{\underline{x}}^2 = \mathcal{E}(\underline{x} - \mathcal{E}\,\underline{x})^2 =$$

$$= \sum_{i=1}^{n} p_i (x_i - \mathcal{E}\,\underline{x})^2 \qquad \text{bij een discrete en}$$

$$= \int_{-\infty}^{+\infty} (x - \mathcal{E}\,\underline{x})^2 f(x)\,dx \qquad \text{bij een continue verdeling.}$$

Voorbeelden: Bij een zuivere dobbelsteen neemt $\widetilde{\underline{x}}$ de waarden $(1-3\frac{1}{2}), (2-3\frac{1}{2}), \ldots\ldots, (6-3\frac{1}{2})$ aan met kansen $\frac{1}{6}$; dus

$$\mathcal{E}\,\widetilde{\underline{x}}^2 = \frac{1}{6}\left\{(-2\tfrac{1}{2})^2 + (-1\tfrac{1}{2})^2 + (-\tfrac{1}{2})^2 + (\tfrac{1}{2})^2 + (1\tfrac{1}{2})^2 + (2\tfrac{1}{2})^2\right\} =$$

$$= \frac{1}{6}\left\{2.6\tfrac{1}{4} + 2.2\tfrac{1}{4} + 2.\tfrac{1}{4}\right\} = \frac{1}{6}.17\tfrac{1}{2} = 2\tfrac{11}{12}.$$

De verwachting van een variabele $\underline{x}$ die homogeen verdeeld is tussen 0 en 1 is gelijk aan $\frac{1}{2}$, de verdelingsdichtheid van $\widetilde{\underline{x}}$ wordt daarvoor dus 1 tussen $-\frac{1}{2}$ en $+\frac{1}{2}$, elders nul. Dus geldt:

$$\mathcal{E}\,\widetilde{\underline{x}}^2 = \int_{-\frac{1}{2}}^{+\frac{1}{2}} x^2\,dx = \frac{1}{3}.\left(\tfrac{1}{2}\right)^3 - \frac{1}{3}\left(-\tfrac{1}{2}\right)^3 = \frac{1}{12}.$$

Men kan de variantie van $\underline{x}$ ook direct uit haar 2^o moment afleiden, door gebruik te maken van de volgende eigenschap:

Eigenschap 5: De variantie van een stochastische variabele is gelijk aan haar tweede moment, verminderd met het kwadraat van haar verwachting. Dus in formule:

(2.11.2) $$\mathcal{E}\,\widetilde{\underline{x}}^2 = \mathcal{E}(\underline{x} - \mathcal{E}\,\underline{x})^2 = \mathcal{E}\,\underline{x}^2 - (\mathcal{E}\,\underline{x})^2.$$

Voorbeelden:

Zuivere dobbelsteen: $\mathcal{E}\,\underline{x}^2 = 15\frac{1}{6}$

$\mathcal{E}\,\underline{x} = 3\frac{1}{2}$, $\longrightarrow$ $(\mathcal{E}\,\underline{x})^2 = 12\frac{1}{4}$

$\mathcal{E}\,\widetilde{\underline{x}}^2 = 2\frac{11}{12}$ (Zie boven).

Homogene verdeling over (0,1): $\mathcal{E}\,\underline{x}^2 = \frac{1}{3}$

$\mathcal{E}\,\underline{x} = \frac{1}{2} \longrightarrow \underline{(\mathcal{E}\,\underline{x})^2 = \frac{1}{4}}$

$\mathcal{E}\,\underline{\tilde{x}}^2 = \frac{1}{12}$.

Eigenschap 5 kan bewezen worden met behulp van eigenschap 4: (Vergelijk de opmerking bij de definitie van $\mathcal{E}\,\underline{x}^2$ in par.2.10). Wij stellen daartoe: $\mathcal{E}\,\underline{x} = \mu$. We vinden dan:

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

De massatheoretische interpretatie van het tweede moment van de gereduceerde variabele is het traagheidsmoment ten opzichte van een bepaald steunpunt, n.l.: het zwaartepunt van de massaverdeling. Het nulpunt van $\underline{\tilde{x}}$ is immers het zwaartepunt (de verwachting) van $\underline{x}$.

Op dezelfde wijze als we hierboven hebben afgeleid, kunnen we aantonen, dat geldt (zowel voor $c>0$ als voor $c<0$)

$$\mathcal{E}\,(\underline{\tilde{x}} + c)^2 = \mathcal{E}\,\underline{\tilde{x}}^2 + c^2 .$$

Dit betekent, dat het tweede moment van een verdeling c^2 groter is dan haar variantie, als het nulpunt van het coördinatenstelsel op een afstand c van haar gemiddelde gekozen wordt. Massatheoretisch betekent dit: het traagheidsmoment van een lijn, waarover een massa 1 verdeeld is, t.o.v. van een steunpunt dat op een afstand c van het zwaartepunt ligt, is c^2 groter dan het traagheidsmoment ten opzichte van het zwaartepunt. We verkrijgen dus het kleinste traagheidsmoment, als de lijn ondersteund wordt in haar zwaartepunt.

Voor de statistiek heeft de variantie vooral betekenis als maat voor de spreiding, die de stochastische variabele om haar gemiddelde vertoont. De term spreiding is daarbij verbonden aan de positieve vierkantswortel uit de variantie; deze grootheid wordt gewoonlijk aangeduid met $\sigma\{\underline{x}\}$. Er geldt dus:

(2.11.3) $$\sigma\{\underline{x}\} = \sqrt{\mathcal{E}\,\underline{\tilde{x}}^2} = \sqrt{\mathcal{E}(\underline{x} - \mathcal{E}\,\underline{x})^2} .$$

Naarmate de spreiding van een verdeling kleiner is, liggen de waarden die de stochastische variabele aanneemt, meer om haar verwachting geconcentreerd. Bij een variabele met een kleine spreiding zal men de verwachting nauwkeuriger kunnen schatten uit het gemiddelde van een steekproef van een bepaalde omvang, dan bij een variable met een grote spreiding; in het eerste geval liggen de waarnemingen immers over het algemeen dichter bij de verwachting dan in het tweede geval. De spreiding kan dus ge-

bruikt worden als een maat voor de nauwkeurigheid waarmee een stochastische grootheid kan worden bepaald.

Opgave

2.11.a: Bereken de variantie van de som van de aantallen ogen geworpen met twee zuivere dobbelstenen (Verg.opgave 2.10.a).

2.12. Eigenschappen van variantie en spreiding.

Eén eigenschap van de variantie is reeds behandeld in de vorige paragraaf (Zie eigenschap 5).

Eigenschap 6: De variantie van een verdeling verandert niet als we het nulpunt van het coördinatenstelsel veranderen. In formule:

(2.12.1) $$\sigma^2(\underline{x}+c) = \sigma^2(\underline{x}).$$

Deze eigenschap volgt direct uit het feit, dat de gereduceerde variabele behorende bij $\underline{x}+c$, hetzelfde is als de gereduceerde variabele behorende bij $\underline{x}$; immers:

$$\underline{x}+c-\mathcal{E}(\underline{x}+c) = \underline{x}+c-(\mathcal{E}\underline{x}+c) = \underline{x}-\mathcal{E}\underline{x}.$$

Voorbeeld: $\underline{x}$ neemt met kansen $\frac{1}{6}$ de waarden: 0,1,2,3,4 en 5 aan. De variantie van $\underline{x}$ is dan dezelfde als die van de zuivere dobbelsteen $(c=-1)$, dus $2\frac{11}{12}$. (Zie par.2.11).

Eigenschap 7: Indien een stochastische variabele met een constante a wordt vermenigvuldigd, wordt haar variantie met a^2 en dus haar spreiding met a vermenigvuldigd. In formule vorm:

(2.12.2) $$\sigma^2\{a\underline{x}\} = a^2\sigma^2\{\underline{x}\} \quad \text{of} \quad \sigma(a\underline{x}) = a\,\sigma\{\underline{x}\}.$$

Voorbeeld: Indien $\underline{x}$ homogeen verdeeld is tussen 0 en 1, is $2\underline{x}$ homogeen verdeeld tussen 0 en 2. De variantie van $\underline{x}$ is, zoals we in par.2.11 hebben aangetoond $\frac{1}{12}$: de variantie van $2\underline{x}$ en dus van een variabele die homogeen verdeeld is tussen 0 en 2 is dus $\frac{1}{3}$.

Algemeen geldt, dat de variantie van een variabele, die homogeen verdeeld is over een interval ter breedte a, gelijk is aan $\frac{1}{12}a^2$. Dit volgt gemakkelijk uit de eigenschappen 6 en 7.

Bewijs van eig.7:

$$\sigma^2\{a\underline{x}\} = \mathcal{E}(a\underline{x}-\mathcal{E}a\underline{x})^2 = \mathcal{E}a^2(\underline{x}-\mathcal{E}\underline{x})^2 = a^2\sigma^2\{\underline{x}\}.$$

Uit eigenschap 7 blijkt de invloed van een schaalverandering op de variantie van een stochastische variabele. Zij b.v. $\underline{x}$ een stochastische variabele, uitgedrukt in de meter als eenheid, die de lengten van de personen van een bepaalde groep representeert, dan zal men bij gebruik van de voet als lengte-eenheid een variabele verkrijgen die gelijk is aan $3{,}28\,\underline{x}$ (1 meter = 3,28 voet). De variantie van de lengten in voeten zal dan $(3{,}28)^2$ x

de variantie van de lengten in meters zijn. De spreiding zal bij deze overgang met de schaalfactor zelf worden vermenigvuldigd. Dit is de reden waarom men als nauwkeurigheidsmaat de voorkeur geeft aan de spreiding; bij schaalveranderingen gedraagt zij zich **het** zelfde als de gemeten grootheid.

Wij kunnen ons nu afvragen, wat er geldt voor de variantie van de som van twee stochastische variabelen $\underline{x}$ en y. De gereduceerde variabele van $\underline{x}+y$ is eenvoudig te bepalen, immers:

$$\underline{x} + y - \mathcal{E}(\underline{x}+y) = \underline{x} - \mathcal{E}\underline{x} + y - \mathcal{E}y = \tilde{\underline{x}} + \tilde{y}.$$

De variantie van $\underline{x}+y$ wordt dus:

(2.12.3) $$\mathcal{E}(\tilde{\underline{x}}+\tilde{y})^2 = \mathcal{E}(\tilde{\underline{x}}^2 + 2\tilde{\underline{x}}\tilde{y} + \tilde{y}^2) = \mathcal{E}\tilde{\underline{x}}^2 + \mathcal{E}\tilde{y}^2 + 2\,\mathcal{E}\tilde{\underline{x}}\tilde{y}.$$

We zien in het rechterlid naast de varianties van $\underline{x}$ en y nog een derde term optreden, waarin $\mathcal{E}\tilde{\underline{x}}\tilde{y}$ voorkomt. Deze verwachting wordt de <u>covariantie</u> van $\underline{x}$ en y genoemd. Bij een discrete tweedimensionale verdeling wordt zij gedefinieerd door:

$$\mathcal{E}\tilde{\underline{x}}\tilde{y} = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i - \mathcal{E}\underline{x})(y_j - \mathcal{E}y)\,P\left[\underline{x} = x_i \text{ en } y = y_j\right].$$

Als nu $\underline{x}$ en y <u>onafhankelijk</u> zijn geldt:

$$P\left[\underline{x} = x_i \text{ en } y = y_j\right] = P\left[\underline{x} = x_i\right] \cdot P\left[y = y_j\right].$$

Dan is dus:

$$\mathcal{E}\tilde{\underline{x}}\tilde{y} = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i - \mathcal{E}\underline{x})\,P\left[\underline{x} = x_i\right]\cdot(y_j - \mathcal{E}y)\,P\left[y = y_j\right] =$$

$$= \sum_{i=1}^{m}(x_i - \mathcal{E}\underline{x})\,P\left[\underline{x} = x_i\right]\cdot\sum_{j=1}^{n}(y_j - \mathcal{E}y)\,P\left[y = y_j\right] = 0.$$

(Zie formule 2.9.2)

De relatie $\mathcal{E}\tilde{\underline{x}}\tilde{y} = 0$ geldt, zoals door een analoge redenering met integralen kan worden bewezen, ook voor continue verdelingen. Uit (2.12.3) volgt dus voor <u>onafhankelijke</u> stochastische variabelen $\underline{x}$ en y:

$$\mathcal{E}(\tilde{\underline{x}}+\tilde{y})^2 = \mathcal{E}\tilde{\underline{x}}^2 + \mathcal{E}\tilde{y}^2 \quad \text{of}$$

(2.12.4) $$\sigma^2\{\underline{x}+y\} = \sigma^2\{\underline{x}\} + \sigma^2\{y\}.$$

Hiermee is de volgende eigenschap bewezen:

<u>Eigenschap 8</u>: De variantie van de som van twee <u>onafhankelijke</u> stochastische variabelen is gelijk aan de som van hun varianties.

<u>Voorbeeld</u>: We hebben gezien, dat de variantie van het aantal ogen, geworpen met een zuivere dobbelsteen $2\frac{11}{12}$ is (Zie par.2.11). De variantie van de som van de aantallen ogen van twee (onafhankelijk) geworpen, zuivere dobbelstenen is dus $2 \times 2\frac{11}{12} = 5\frac{5}{6}$ (Verg. opgave 2.11a).

Opmerking: Het is voor de geldigheid van eigenschap 8 niet strikt nodig, dat $\underline{x}$ en $\underline{y}$ onafhankelijk zijn. Zoals uit het bewijs duidelijk blijkt, is het voldoende, dat geldt $\mathcal{E}\,\underline{\tilde{x}}\,\underline{\tilde{y}} = 0$. Dit kan zoals we later zullen zien, ook het geval zijn als $\underline{x}$ en $\underline{y}$ niet onafhankelijk zijn. Voorlopig zullen we echter alleen gebruik maken van toepassingen van eigenschap 8.

Opgave:

2.12.a: $\underline{x}$ en $\underline{y}$ zijn beide homogeen verdeeld tussen 0 en 1. Hoe groot zijn de verwachting, de variantie en de spreiding van $\underline{x}+\underline{y}$:

1) als $\underline{x}$ en $\underline{y}$ onafhankelijk zijn
2) als $\underline{x}$ en $\underline{y}$ volkomen afhankelijk zijn: d.w.z. als $\underline{y}$ altijd dezelfde waarden aanneemt als $\underline{x}$.

2.13 Consequenties van eigenschap 8.

Eigenschap 8 kan evenals eigenschap 3 tot meer dan 2 variabelen worden uitgebreid.

Eigenschap 9: De variantie van de som van een aantal onafhankelijke stochastische variabelen, is gelijk aan de som van hun varianties. In formulevorm:

$$\sigma^2\left\{\sum_{i=1}^{n} \underline{x}_i\right\} = \sum_{i=1}^{n} \sigma^2\{\underline{x}_i\} \quad . \tag{2.13.1}$$

Hiertoe is het feitelijk al voldoende dat voor ieder paar variabelen $\underline{x}_i$ en $\underline{x}_j$ geldt: $\mathcal{E}\,\underline{\tilde{x}}_i\,\underline{\tilde{x}}_j = 0$. Onafhankelijkheid van de variabelen is een veel zwaardere eis: deze houdt in, dat de verdeling van geen enkele der beschouwde variabelen afhangt van de waarden die andere variabelen aannemen. De stelling wordt echter dikwijls toegepast in de hierboven gegeven vorm.

Met behulp van Eigenschap 9 kan men de variantie van een willekeurige binomiale verdeling berekenen. In par. 2.8 hebben wij gezien, dat, als de variabelen $\underline{x}_1, \ldots\ldots, \underline{x}_n$ onafhankelijk zijn en alle met kansen p resp. $1-p$ de waarden 1 resp. 0 aannemen, hun som

$$y = \sum_{i=1}^{n} \underline{x}_i$$

een binomiale verdeling heeft met parameters n en p .

Nu geldt: $\mathcal{E}\,\underline{x}_i = p$,

$$\mathcal{E}\,\underline{x}_i^2 = p \cdot 1^2 + (1-p)\cdot 0^2 = p \quad .$$

Dus: $\sigma^2\{\underline{x}_i\} = \mathcal{E}\,\underline{\tilde{x}}_i^2 = \mathcal{E}\,\underline{x}_i^2 - (\mathcal{E}\,\underline{x}_i)^2 = p - p^2 = p(1-p)$.

Dus: $\sigma^2\{\underline{y}\} = \sum_{i=1}^{n} \sigma^2\{\underline{x}_i\} = np(1-p)$.

De variantie van een binomiale verdeling met parameters n en p is dus $np(1-p)$ haar spreiding is dus $\sigma\{\underline{y}\} = \sqrt{np(1-p)}$.

Door combinatie van Eigenschap 7 en Eigenschap 9 kunnen wij nu ook de variantie van lineaire combinaties van onafhankelijke stochastische variabelen in de varianties van die variabelen uitdrukken. Men vindt:

$$\sigma^2\left\{\sum_{i=1}^{n} a_i \underline{x}_i\right\} = \sum_{i=1}^{n} a_i^2\, \sigma^2\{\underline{x}_i\}. \qquad (2.13.2)$$

Als bijzonder geval vinden wij voor twee variabelen met $a_1 = 1$, $a_2 = -1$:

$$\sigma^2\{\underline{x}_1 - \underline{x}_2\} = \sigma^2\{\underline{x}_1\} + \sigma^2\{\underline{x}_2\}. \qquad (2.13.3)$$

De variantie van het <u>verschil</u> van twee <u>onafhankelijke</u> variabelen is dus dezelfde als de variantie van hun som. Dit kan men ook direct als volgt inzien. Er geldt:

$$\mathcal{E}(\underline{\tilde{x}} - \underline{\tilde{y}})^2 = \mathcal{E}\,\underline{\tilde{x}}^2 + \mathcal{E}\,\underline{\tilde{y}}^2 - 2\,\mathcal{E}\,\underline{\tilde{x}}\,\underline{\tilde{y}}$$

en de laatste term in deze uitdrukking is nul als $\underline{x}$ en $\underline{y}$ onafhankelijk zijn, zoals wij in par. 2.12 gezien hebben.

Een andere toepassing van formule (2.13.2) verkrijgen wij, als wij de variantie willen bepalen van het rekenkundig gemiddelde van n stochastische variabelen $\underline{x}_1, \ldots, \underline{x}_n$ met dezelfde variantie σ^2. Als dan verder voor alle $a_i = \frac{1}{n}$ genomen wordt, vinden wij:

$$\sigma^2\left\{\frac{1}{n}\sum_{i=1}^{n} \underline{x}_i\right\} = \sum_{i=1}^{n} \frac{1}{n^2}\, \sigma^2\{\underline{x}_i\} = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}. \qquad (2.13.4)$$

Deze formule geldt b v als alle $\underline{x}_i$ dezelfde verdeling hebben.

Wij hebben dan de variantie van het gemiddelde van n onafhankelijke waarnemingen van een stochastische variabele $\underline{x}$ uitgedrukt in de variantie van $\underline{x}$ zelf. Het gemiddelde van de waarnemingen wordt daarbij weer opgevat als een stochastische variabele, zoals wij ook gedaan hebben in par. 2.8 ; als wij dit gemiddelde voorstellen door $\underline{\bar{x}}$, vinden wij dus:

$$\mathcal{E}\,\underline{\bar{x}} = \mathcal{E}\,\underline{x} \quad \text{(Vgl. formule 2.8.5)} \qquad (2.13.5)$$

en:

(2.13 6) $$\sigma\{\bar{x}\} = \frac{\sigma\{x\}}{\sqrt{n}} .$$

Het gemiddelde van n onafhankelijke waarnemingen van een stochastische variabele x heeft dus een verwachting die gelijk is aan de verwachting van x , doch een spreiding die gelijk is aan $1/\sqrt{n}$ maal de spreiding van x . Hieruit volgt, dat als n aangroeit, de spreiding van $\bar{x}$ steeds dichter tot nul nadert, zodat de verdeling van $\bar{x}$ steeds meer gaat lijken op de verdeling van een variabele die met kans 1 de waarde $\mathcal{E}_x$ aanneemt. Hierop is de theoretische afleiding van de wet der grote getallen gebaseerd, die, zoals bekend is, inhoudt, dat het gemiddelde van grote steekproeven bijna zeker convergeert naar $\mathcal{E}_x$. Tevens blijkt, dat het moeilijker wordt om $\mathcal{E}_x$ te schatten, naarmate $\sigma\{x\}$ groter is, immers $\sigma\{\bar{x}\}$ is evenredig met $\sigma\{x\}$.

Formule (2.13.6) geldt niet voor het gemiddelde van steekproeven zonder teruglegging uit een eindige populatie. Dit geval hebben wij beschouwd in het voorbeeld betreffende de appels van par. 2.8 (methode b). Bij een partij van N appels, met gewichten $g_1, \ldots, g_N$, zal het gewicht van een aselect getrokken appel een stochastische variabele x zijn, die met kansen $\frac{1}{N}$ de waarden $g_1, \ldots, g_N$ aanneemt. Daaruit volgt:

(2.13.7) $$\mathcal{E}_x = \frac{1}{N} \sum_{i=1}^{N} g_i \qquad \text{en}$$

(2.13.8) $$\sigma\{x\} = \frac{1}{N} \sum_{i=1}^{N} (g_i - \mathcal{E}_x)^2 .$$

Als $\bar{x}$ het gemiddelde gewicht is van een aselecte steekproef met teruglegging van n waarnemingen, gelden voor $\bar{x}$ de formules (2.13.5) en (2.13.6).

Is daarentegen $\bar{x}'$ het gemiddelde gewicht van een steekproef zonder teruglegging van n waarnemingen, dan geldt:

(2.13.9) $\mathcal{E}_{\bar{x}'} = \mathcal{E}_x$ en:

(2.13.10) $$\sigma\{\bar{x}'\} = \sqrt{\frac{N-n}{N-1}} \cdot \sigma\{\bar{x}\} = \sqrt{\frac{N-n}{n(N-1)}} \cdot \sigma\{x\} .$$

Deze laatste twee formules zullen wij hier niet bewijzen.

Wij zien aan de formules (2.13.5) t/m (2.13.10) dat x , $\bar{x}$ en $\bar{x}'$ alle dezelfde verwachtingswaarde hebben; hun spreidingen zijn echter verschillend; als $n > 1$ is, heeft x een grotere spreiding dan $\bar{x}$, $\bar{x}$ weer een grotere spreiding dan $\bar{x}'$; $\bar{x}'$ zal dus in het algemeen een betere benadering zijn voor het partijgemiddelde

dan $\bar{x}$; als n toeneemt, wordt zowel de spreiding van $\bar{x}$ als die van $\bar{x}'$ kleiner; vanwege de extra factor $\sqrt{\frac{N-n}{n(N-1)}}$ gaat dit bij $\bar{x}'$ sneller dan bij $\bar{x}$

In par 2 8 was bij methode b: $N = 1000$ en $n = 5$. In dat geval maakt het nagenoeg niets uit, of wij mét of zonder teruglegging trekken In de spreiding scheelt het slechts een factor $\sqrt{\frac{995}{999}} = 0{,}998$ .Doch de spreiding van het steekproefgemiddelde is wel veel kleiner dan die van een trekking, immers:

$$\sigma(\bar{x}') = 0{,}998 \frac{\sigma(x)}{\sqrt{5}} = 0{,}446\ \sigma(x). \qquad \left(\sigma(\bar{x}) = 0{,}447\ \sigma(x)\right)$$

Hieruit volgt, dat het gemiddelde van een aantal waarnemingen van $\bar{x}$ of $\bar{x}'$ in het algemeen veel dichter bij het partijgemiddelde zal liggen dan het gemiddelde van hetzelfde aantal waarnemingen van x zelf. De methode b genoemd in par. 2.8 is dus veel effectiever dan methode a, als wij mogen aannemen, dat het trekken en wegen van 5 appels niet noemenswaard meer tijd kost dan het trekken en wegen van één appel.

Opgaven:

2 13 a. Wat wordt volgens formule (2.13 10) de spreiding van $\bar{x}'$ als $m \to N$? Kunt U dit verklaren?

2.13.b. Uit een partij van 1000 appels worden 100 appels getrokken zónder teruglegging. Het gemiddelde gewicht hiervan zij $\bar{x}'$. Hoeveel appels moet men mét teruglegging uit de partij nemen, om een gemiddeld gewicht te verkrijgen met dezelfde spreiding als $\bar{x}'$ Hetzelfde wordt gevraagd als $\bar{x}'$ correspondeert met een steekproef zonder teruglegging van 500 resp. 800 exemplaren.

2 13.c Als men 20 appels uit een partij van 1000 mag trekken om het gemiddelde gewicht van de 1000 appels te schatten, met welke van de volgende methoden zou men dan het nauwkeurigste resultaat kunnen verkrijgen:

1) Een steekproef van 20 exemplaren met teruglegging?
2) 4 steekproeven van 5 met teruglegging?

3) 4 steekproeven van 5 zonder terugleggíng?
4) 1 steekproef van 20 zonder terugleggíng?

Er behoeft hierbij geen rekening te worden gehouden met weegfouten. In geval 3) worden de appels na iedere steekproef weer aan de partij toegevoegd alvorens men de volgende steekproef neemt.

§ 14. Over het schatten van variantie en spreiding van een verdeling

Zoals wij in par. § 11 gezien hebben, kan de spreiding van een verdeling beschouwd worden als een maat voor de nauwkeurigheid waarmee men het gemiddelde van de verdeling kan schatten uit het gemiddelde van een steekproef. In de gevallen, waarin men het gemiddelde van een verdeling aldus wenst te schatten, zal men in de regel de spreiding van de verdeling evenmin kennen. Het is dus gewenst om methoden aan te geven, volgens welke men de spreiding of de variantie van een verdeling kan schatten uit de gegevens van een steekproef.

Laat nu $\underline{x}$ een stochastische variabele zijn met onbekende verwachting $\mathcal{E}_{\underline{x}=\mu}$ en onbekende spreiding $\sigma\{\underline{x}\} = \sigma$. Wij beschikken over een steekproef $x_1, \ldots, x_n$ van onderling onafhankelijke waarnemingen van $\underline{x}$, waaruit wij σ trachten te schatten.

Wij hebben in par. 2.10 gezien hoe men het tweede moment van $\underline{x}$ kan schatten. Aangezien σ^2 het tweede moment is van de gereduceerde variabele $\underline{\tilde{x}}$, zou men op deze wijze σ^2 kunnen schatten, als men de beschikking had over waarnemingen van $\underline{\tilde{x}}$. Wij hebben deze echter niet; men zou ze kunnen afleiden uit de steekproefwaarden $x_1, \ldots, x_n$ door deze alle met de onbekende verwachting μ te verminderen. In plaats daarvan zullen wij gebruik moeten maken van het gemiddelde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ van de steekproef, hetgeen een schatting van μ is. Als wij dan verder volgens par. 2.10 te werk gaan, zouden wij als schatting voor σ^2 vinden:

$$(2.14.1) \qquad s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Deze schatting heeft echter een bezwaar. Als wij de steekproefwaarden als stochastische variabelen opvatten, blijkt de verwachting van $\underline{s}^2$ niet gelijk te zijn aan de variantie σ^2. Een dergelijke schatting noemt men onzuiver (Eng.: biased). Een consequentie hiervan is, dat het gemiddelde van de schattingen s^2, berekend uit een groot aantal steekproeven van n waarnemingen, in de regel σ^2 niet goed zal benaderen. Wij kunnen aantonen, dat geldt:

$$(2.14.2) \qquad \mathcal{E}\, \underline{s}^2 = \frac{n-1}{n}\, \sigma^2 .$$

Deze formule kan worden afgeleid, met behulp van de in dit hoofdstuk behandelde eigenschappen. Wij zullen dat hier niet doen en volstaan met enkele opmerkingen.

Als $n = 2$, blijkt uit formule (2.14.2) dat de verwachting van $\underline{s}^2$ slechts $\frac{1}{2}\sigma^2$ is. Als wij σ^2 dus schatten door de s^2 van steekproeven van 2 waarnemingen, zullen wij in het algemeen veel te kleine waarden vinden. Naarmate n toeneemt wordt de fout minder ernstig, immers als n aangroeit, convergeert $\frac{n-1}{n} = 1 - \frac{1}{n}$ naar 1. Door een kleine wijziging kunnen wij echter een schatting verkrijgen die ook voor kleine n bevredigend is. Als wij namelijk de verwachting van

$$(2.14.3) \qquad \underline{s}'^2 = \frac{n}{n-1}\, \underline{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})^2$$

bepalen, vinden wij uit Eigenschap 2 en formule (2.14.2) gemakkelijk:

$$\mathcal{E}\, \underline{s}'^2 = \frac{n}{n-1}\, \mathcal{E}\, \underline{s}^2 = \sigma^2 .$$

De verwachting van $\underline{s}'^2$ is dus voor iedere n gelijk aan de geschatte grootheid σ^2. Een dergelijke schatting noemt men zuiver (Eng.: unbiased).

Men moet zich door de term zuivere schatting niet laten misleiden. Het feit, dat een schatting zuiver is, houdt slechts in, dat men door veelvuldige herhaling van het schattingsprocédé een reeks waarden kan verkrijgen, waarvan het gemiddelde in de regel niet veel van de geschatte grootheid zal verschillen. Een enkele waarde verkrijgen met een zuivere schattingsmethode behoeft dus helemaal niet gelijk te zijn aan de geschatte grootheid

en niet eens een vrij nauwkeurige schatting daarvan te zijn. Als $\underline{x}$ een stochastische variabele is, is één willekeurige waarneming daarvan reeds een zuivere schatting van $\mathcal{E}_{\underline{x}}$, het gemiddelde van 10000 waarnemingen echter ook en is, zoals wij gezien hebben, veel nauwkeuriger; s'^2 berekend uit 2 waarnemingen is een zuivere schatting van $\sigma^2\{\underline{x}\}$; s^2 berekend uit 10.000 waarnemingen is géén zuivere, maar in het algemeen wel een veel nauwkeuriger schatting. Dit neemt nog niet weg, dat men, zo men te kiezen heeft tussen een zuivere en een onzuivere schatting, gebaseerd op evenveel waarnemingen, in het algemeen de voorkeur geeft aan de zuivere schatting.

Om s^2 of s'^2 te kunnen berekenen, moeten wij de som $\sum_{i=1}^{n}(x_i-\bar{x})^2$ bepalen, dit is de som van de kwadraten van de gereduceerde steekproefwaarden (zie par. 2.9). Door gebruik te maken van een eigenschap die analoog is aan de eigenschap van de variantie van een verdeling, kunnen wij het aftrekken van het gemiddelde vermijden. Er geldt namelijk:

$$(2.14.4)\qquad \sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2=\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}.$$

Wij bewijzen deze formule als volgt:

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

$$=\sum_{i=1}^{n}x_i^2-2\bar{x}.n\bar{x}+n\bar{x}^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2.$$

Voorbeeld: Gegeven zij de waarnemingen:
-44, -35, -23, -13, -8, +12, +27, +40.
Wij berekenen s'^2 als volgt:

Waarnemingen x_i	Kwadraten x_i^2
-44	1936
-35	1225
-23	529
-13	169
-8	64
+12	144
+27	729
+40	1600
$\sum_{i=1}^{8}x_i=-44$	$\sum_{i=1}^{8}x_i^2=6396$

$$\sum_{i=1}^{8}(x_i-\bar{x})^2=\sum_{i=1}^{8}x_i^2-\frac{\left(\sum_{i=1}^{8}x_i\right)^2}{8}=6396-\frac{44^2}{8}=6396-\frac{1936}{8}=6154.$$

$$s'^2=\frac{1}{7}\sum_{i=1}^{8}(x_i-\bar{x})^2=879\tfrac{1}{7}.$$

Als schatting van de spreiding σ gebruikt men gewoonlijk de wortel uit de schatting van de variantie, dus $s'=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$. Uit het feit, dat $\underline{s}'^2$ een zuivere schatting is van σ^2 volgt niet dat $\underline{s}'$ een zuivere schatting van σ is. Daarvoor zou moeten gelden:

$\mathcal{E}_{\underline{s}'} = \sigma$ en dus:

$$\left(\mathcal{E}_{\underline{s}'}\right)^2 = \sigma^2 = \mathcal{E}_{\underline{s}'^2}$$

of : $\mathcal{E}_{\underline{s}'^2} - \left(\mathcal{E}_{\underline{s}'}\right)^2 = \mathcal{E}\left(\underline{s}' - \mathcal{E}_{\underline{s}'}\right)^2 = 0$ (Vgl. Eig. 5).

Dit zou dus betekenen, dat de variantie van $\underline{s}'$ gelijk was aan nul, en dat dus de schatting $\underline{s}'$ steeds gelijk was aan de geschatte grootheid σ Dit is in het algemeen natuurlijk niet het geval, er zal gelden:

$\mathcal{E}\left(\underline{s}' - \mathcal{E}_{\underline{s}'}\right)^2 > 0$ waaruit volgt:

$\left(\mathcal{E}_{\underline{s}'}\right)^2 < \mathcal{E}_{\underline{s}'^2} = \sigma^2$ of : $\mathcal{E}_{\underline{s}'} < \sigma$.

Als schatting van de spreiding geeft s' dus, gemiddeld over een groot aantal steekproeven, in het algemeen te kleine waarden Dit geldt dan a fortiori voor s, aangezien:

$s = \sqrt{\frac{n-1}{n}}\ s'$ (Vgl. formule (2.14.3)).

Opgaven:

2.14 a Is $\frac{1}{n}\sum_{i=1}^{n} x_i$ een zuivere schatting van $\mathcal{E}_{\underline{x}}$ en is $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ een zuivere schatting van $\mathcal{E}_{\underline{x}^2}$?

2.14.b Verandert de waarde van s'^2 in ons voorbeeld als alle steekproefwaarden met 100 vermeerderd worden? En als alle steekproefwaarden met 100 worden vermenigvuldigd?

2.14 c Bereken gemiddelde en variantie s'^2 bij de volgende reeks waarnemingen:

20,15; 19,95; 20,40; 20,16; 19,98; 20,03; 20,12; 19,85.

Opmerking: Men kan de berekening aanzienlijk vereenvoudigen door gebruik te maken van het feit, dat alle waarnemingen in de omgeving van 20 liggen.

2.15 Standaardisering van een stochastische variabele.

In par. 2.12 (Eigenschap 7) hebben wij gezien, dat als a een constante is, geldt:

$$\sigma^2\{a\,\underline{x}\} = a^2\,\sigma^2\{\underline{x}\}\ .$$

Indien wij nu voor a $\frac{1}{\sigma\{\underline{x}\}}$ kiezen, vinden wij, dat de variantie van

$$y = \frac{\underline{x}}{\sigma\{\underline{x}\}}$$

gelijk is aan 1. De spreiding van y is dan dus eveneens gelijk aan 1 Indien dus een stochastische variabele gedeeld wordt door haar spreiding, gaat zij over in een stochastische variabele met spreiding 1. Het delen door de spreiding, noemt men

standaardiseren, de verkregen stochastische variabele noemt men een gestandaardiseerde variabele.

In par. 2.9 hebben wij reeds gezien, dat men een stochastische variabele kan overvoeren in een variabele met verwachting 0, door haar te reduceren, d.w.z. te verminderen met haar verwachting. Bij dit reduceren verandert de spreiding niet; indien men een gereduceerde variabele dus door haar spreiding (die gelijk is aan de spreiding van de oorspronkelijke variabele) deelt, verkrijgt men een gereduceerde én gestandaardiseerde variabele, die een verwachting 0 én een spreiding 1 heeft. Men kan een willekeurige stochastische variabele x met verwachting μ en spreiding σ dus reduceren en standaardiseren, door over te gaan op de variabele

(2.15 1) $$y = \frac{x - \mu}{\sigma} .$$

De overgang van x op y komt overeen met een verandering van het nulpunt en van de schaal van het coördinatenstelsel. Door het reduceren wordt het punt, dat correspondeert met de verwachtingswaarde het nulpunt. Door het standaardiseren wordt de spreiding de schaaleenheid. Het zal blijken, dat de keuze van ee dergelijke schaal bijzonder pretttig is bij normale verdelingen en bovendien van belang is voor de definitie van de correlatiecoëfficiënt.

Men kan de waarnemingen $x_1, \ldots, x_n$ van een steekproef eveneens verminderen met hun gemiddelde $\bar{x}$ en delen door hun spreiding s'. Men verkrijgt dan een nieuwe reeks waarden: $y_1, \ldots, y_n$ met $$y_i = \frac{x_i - \bar{x}}{s'} ,$$ waarvan het gemiddelde 0 en de spreiding: $\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} y_i^2}$ gelijk is aan 1. Wij zullen dit toepassen bij het aanpassen van een normale verdeling aan een aantal steekproefwaarden.

Opgaven:

2.15.a Als x een binomiale verdeling heeft met parameters p en n, welke is dan de bij x behorende

1) gereduceerde variabele?
2) gestandaardiseerde variabele?
3) gereduceerde én gestandaardiseerde variabele?

2.15.b $\underline{x}$ is een stochastische variabele met verwachting 1 en spreiding 2. Gegeven zijn de volgende 10 waarnemingen van $\underline{x}$: +0,20; +0,44; +3,46; +2,80; -1,00; +1,68; +2,62; +0,80; +2,96; +2,10

1 Leid hieruit de 10 steekproefwaarden af van de gereduceerde en gestandaardiseerde variabele behorende bij $\underline{x}$.

2 Reduceer en standaardiseer de steekproefwaarden met hun eigen gemiddelde en spreiding

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 155

Cursus

Toegepaste Statistiek II

door

Ph. van Elteren en J. Kriens

III De Normale verdeling.

1955

3 De Normale verdeling.

3.1. $N(0,1)$-verdeling.

Normale verdelingen zijn verdelingen, die door reducering en standaardisering van de variabele overgaan in een bepaalde verdeling, die wij hier aanduiden als de $N(0,1)$-verdeling. Volgens de definitie is de $N(0,1)$-verdeling zelf ook een normale verdeling; dit wordt weergegeven door de letter N in het symbool $N(0,1)$; de cijfers 0 en 1 corresponderen met het gemiddelde en de spreiding van de verdeling. Een normale verdeling met gemiddelde μ en spreiding σ zal dus een $N(\mu,\sigma)$-verdeling genoemd worden (zie par. 3 2).

De verdelingsdichtheid van een $N(0,1)$-verdeling wordt gegeven door:

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

Hierin is π een uit de meetkunde bekend getal: de verhouding tussen omtrek en middellijn van een cirkel. Het getal e is bekend als het grondtal van het natuurlijk logarithmenstelsel.
Er geldt:

$$\pi = 3,14159265\ldots$$

$$e = 2,71828183\ldots$$

Men ziet aan formule (3.1.1), dat de $N(0,1)$-verdeling symmetrisch is t.o.v. 0 immers:

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

Hieruit volgt dat het gemiddelde van de verdeling inderdaad 0 is (zie par. 2.5). Het is iets moeilijker om aan te tonen, dat haar spreiding 1 is; dit zullen wij hier niet doen.

De verdelingsfunctie van de $N(0,1)$-verdeling wordt gegeven door:

$$P[\underline{x} \leqq x] = F(x) = \int_{-\infty}^{x} f(u)\,du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}u^2}\,du.$$

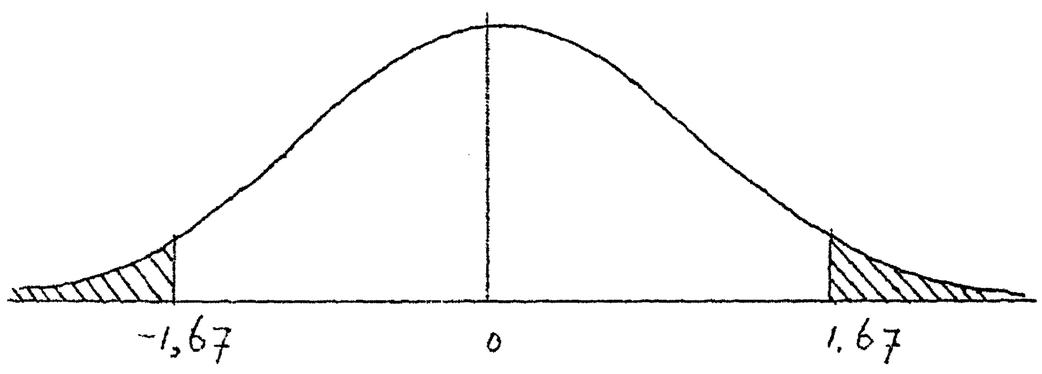
(Vgl. par. 1.5)

In deze syllabus wordt een tabel van $P[\underline{x} \geqq x] = 1 - F(x)$ voor $x = 0,00$; 0,01; 0,02; ..,3,49 ingelast. Wij willen het gebruik van deze tabel aan een aantal voorbeelden toelichten. Bij al deze voorbeelden stelt $\underline{x}$ een $N(0,1)$-verdeelde stochastische variabele voor.

a Zoek op: $P[x \geqq 2,25]$.
Oplossing: Men zoekt in de tabel het getal in rij 2,2 en kolom 5; men vindt 0122. Dit zijn de eerste 4 decimalen van de gezochte kans. Er geldt dus $P[x \geqq 2,25] = 0,0122$.

b Zoek op: $P[x \leqq 1,07]$.
Oplossing: Er geldt: $P[x \leqq 1,07] + P[x \geqq 1,07] = 1$. $(P[x = 1,07] = 0)$
Dus $P[x \leqq 1,07] = 1 - P[x \geqq 1,07] = 1 - 0,1423$ (rij 1,0, kolom 7)
$= 0,8577$.

c Zoek op: $P[x \leqq -1,67]$.
Oplossing: De verdeling is symmetrisch t.o.v 0 (zie figuur 3.1)
Er geldt dus: $P[x \leqq -1,67] = P[x \geqq 1,67] = 0,0475$.
Het streepje onder de 5 betekent, dat bij afronding op 3 decimalen voor de derde decimaal het dichtstbijzijnde oneven getal moet worden gekozen; dus hier: 0,047 en niet 0,048.



Figuur 3.1. Toelichting bij voorbeeld c. De tweezijdige overschrijdingskans is de gearceerde oppervlakte.

d Zoek op: $P[x \geqq -0,92]$.
Oplossing: $P[x \geqq -0,92] = P[x \leqq 0,92] = 1 - P[x \geqq 0,92] = 1 - 0,1788 = 0,8212$.

e Zoek op: $P[x \geqq 1,96$ of $x \leqq -1,96]$.
Oplossing: $P[x \geqq 1,96$ of $x \leqq -1,96] = P[x \geqq 1,96] + P[x \leqq -1,96] = 2\,P[x \geqq 1,96] =$
$2 \times 0,0250 = 0,0500$.

f Voor welke x is $P[x \geqq x] = 0,01$?
Oplossing: $P[x \geqq 2,32] = 0,0102$.
$P[x \geqq 2,33] = 0,0099$.

De gezochte waarde van x ligt dus tussen 2,32 en 2,33. Bij lineaire interpolatie vindt men 2,327; volgens nauwkeuriger tabellen is $x = 2,326$.

g Voor welke waarde van x is $P[x \geqq x$ of $x \leqq -x] = 0,10$?
Oplossing: $P[x \geqq x$ of $x \leqq -x] = 2\,P[x \geqq x] = 0,10$.
Dus $P[x \geqq x] = 0,05$.

Volgens de methode gevolgd onder f met lineaire interpolatie vindt men hieruit $x = 1,645$.

Opmerking De hier gegeven voorbeelden houden verband met het bepalen van overschrijdingskansen voor het toetsen van hypothesen. Bij de $N(0,1)$-verdeling is $P[\underline{x} \geqq x]$ de rechtseenzijdige en $P[\underline{x} \leqq x]$ de linkseenzijdige overschrijdingskans van x. De tweezijdige overschrijdingskans van x is de kans dat $\underline{x}$ een waarde aanneemt die minstens even ver van het gemiddelde van de verdeling (hier: 0) afwijkt als x; deze kans is hier dus:

$P[\underline{x} \geqq x \text{ of } \underline{x} \leqq -x] = 2\ P[\underline{x} \geqq x]$ als $x > 0$ en

$P[\underline{x} \geqq -x \text{ of } \underline{x} \leqq +x] = 2\ P[\underline{x} \geqq -x]$ als $x < 0$.

In de voorbeelden a, b, c en d zijn eenzijdige overschrijdingskansen bepaald, in voorbeeld e een tweezijdige overschrijdingskans. In de voorbeelden f en g is gezocht naar een waarde van x met gegeven rechtseenzijdige, resp. tweezijdige overschrijdingskans. Men lette erop, dat bij gegeven tweezijdige overschrijdingskans 2 waarden van x behoren; bij de kans 0,10 bijvoorbeeld de waarden: $x = 1{,}645$ en $x = -1{,}645$.

De overschrijdingskansen worden gebruikt als men wenst na te gaan of een bepaalde waarneming al dan niet afkomstig zou kunnen zijn uit een $N(0,1)$-verdeling. Men toetst dan de hypothese, dat een stochastische variabele $\underline{x}$, waarvan men een waarneming verricht heeft, de verdeling $N(0,1)$ heeft, waarbij de overschrijdingskans als criterium gebruikt wordt. Bij verscheidene toetsingsmethoden komt de bewerking uiteindelijk neer op het toetsen van een dergelijke hypothese. Dit is bij een voldoende groot aantal waarnemingen o.a. het geval bij de tekentoets en de toets van Wilcoxon, die beide behandeld zijn in de Cursus Toegepaste Statistiek I.

Opgaven:

3.1.a Bepaal (bij een $N(0,1)$-verdeling) de rechtseenzijdige overschrijdingskansen van $x = 0{,}84$; $2{,}34$ en $-3{,}00$. Rond de uitkomsten af tot in 3 decimalen nauwkeurig.

3.1.b Bepaal tevens de linkseenzijdige overschrijdingskansen bij de bovengenoemde waarden van x.

3.1.c Bepaal de tweezijdige overschrijdingskansen voor $x = 0$ en $x = -3{,}49$.

3.1.d Benader zo goed mogelijk de waarden van x met

1) de rechtseenzijdige overschrijdingskans 0,10,
2) de linkseenzijdige " 0,05,
3) de tweezijdige " 0,01.

3.2 $N(\mu,\sigma)$-verdeling.

Een $N(\mu,\sigma)$-verdeling is, zoals vermeld in par. 3.1, een normale verdeling met gemiddelde μ en spreiding σ en kan door reducering en standaardisering van de variabele herleid worden tot een $N(0,1)$-verdeling. Overschrijdingskansen e.d. met betrekking tot een $N(\mu,\sigma)$-verdeling kunnen dus bepaald worden met behulp van de tabel van de $N(0,1)$-verdeling. Wij geven hieronder enige voorbeelden:

a Bepaal $P[x \geqq 3,50]$ als x een $N(4,50\ ;\ 0,5)$ -verdeling heeft.

Oplossing:

$$P[x \geqq 3,50] = P\left[\frac{x-4,50}{0,5} \geqq \frac{3,50-4,50}{0,5}\right] = P[y \geqq -2,00],$$

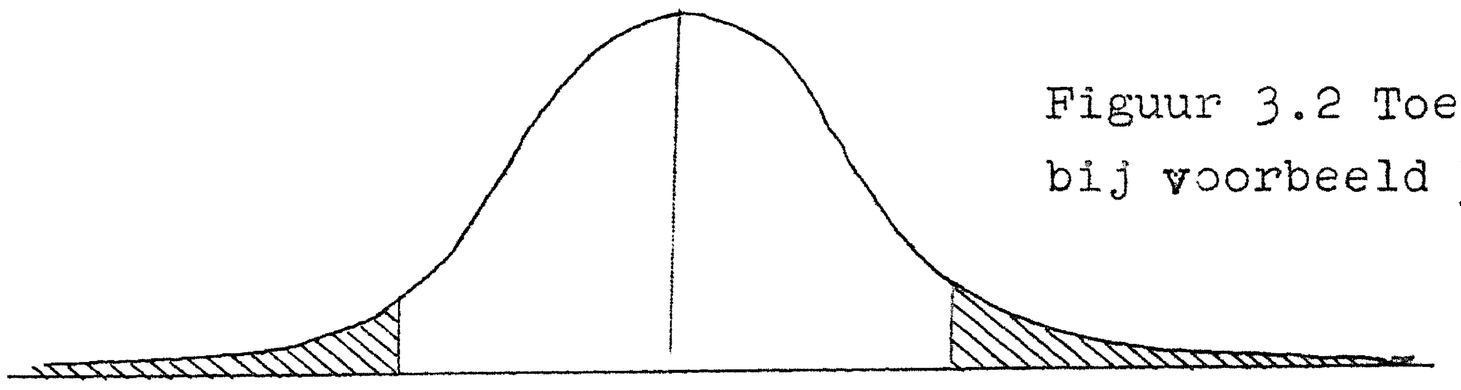
waarin y een $N(0,1)$-verdeling heeft. Wij vinden dus:

$$P[y \geqq -2,00] = 1 - P[y \geqq 2,00] = 0,9772.$$

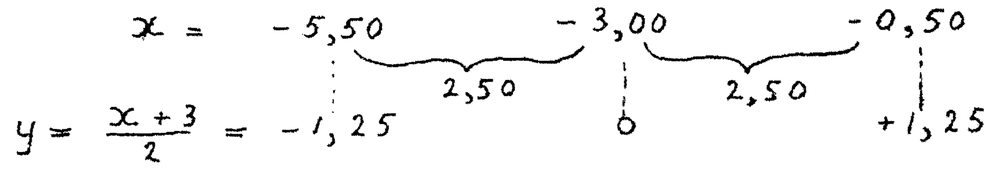
b Bepaal de tweezijdige overschrijdingskans van $x = -0,50$ bij een $N(-3,00\ ;2,00)$verdeling.

Oplossing: De tweezijdige overschrijdingskans van x is, zoals wij gezien hebben, de kans, dat x een waarde aanneemt, die minstens even ver van haar verwachtingswaarde μ (hier: - 3,00) afwijkt, als x (hier: -0,50). In dit geval wordt dit dus de kans:

$$P[x \geqq -0,50 \text{ of } x \leqq -5,50] \quad \text{(zie figuur 3.2)}$$



Figuur 3.2 Toelichting bij voorbeeld b.



Bij reducering en standaardisering van de variabele gaat deze kans over in de tweezijdige overschrijdingskans van de waarde $\frac{x-\mu}{\sigma}$ dus in: $P[y \geqq 1,25 \text{ of } y \leqq -1,25]$ $(y : N(0,1))$

$$= 2\,P[y \geqq 1,25] = 0,2112.$$

c Bepaal bij een $N(1,2)$-verdeling de waarden van $\underline{x}$ met een tweezijdige overschrijdingskans 0,05.

Oplossing: Als $y=\frac{1}{2}(x-1)$ heeft y een $N(0,1)$-verdeling en dus hebben $y_1=-1{,}96$ en $y_2=+1{,}96$ een tweezijdige overschrijdingskans 0,05 (vgl. 3.1 vraag e). De gezochte waarden van $\underline{x}$ worden dus gevonden uit:

$$\pm\frac{x_1-1}{2}=-1{,}96 \quad \text{dus} \quad x_1=-2{,}92 \quad \text{en}$$

$$\frac{x_2-1}{2}=+1{,}96 \quad \text{dus} \quad x_2=+4{,}92.$$

Opgaven:

3.2.a $\underline{x}$ is normaal verdeeld met verwachting 1,37 en variantie 1,44. Bepaal:

1) De linkseenzijdige overschrijdingskans van $x=0$.

2) De tweezijdige overschrijdingskans van $x=4$.

3) De waarde van x met een rechtseenzijdige overschrijdingskans 0,01.

4) De waarden van x met een tweezijdige overschrijdingskans 0,001.

3.3. Enige eigenschappen van de normale verdeling.

Als $\underline{x}$ een $N(\mu,\sigma)$-verdeling heeft en a en c zijn constanten, dan heeft $\underline{z}=a\underline{x}+c$ volgens de eigenschappen 1 en 2 (pag. 32) en 6 en 7 (pag. 46) de verwachting $a\mu+c$ en de spreiding $|a|\sigma$. Er geldt nu:

$$\frac{\underline{z}-\mathcal{E}\underline{z}}{\sigma(\underline{z})}=\frac{a\underline{x}+c-(a\mu+c)}{|a|\sigma}=\begin{cases}+\dfrac{\underline{x}-\mu}{\sigma} & \text{als } a>0\\ -\dfrac{\underline{x}-\mu}{\sigma} & \text{als } a<0.\end{cases}$$

Omdat $\underline{x}$ een $N(\mu,\sigma)$ verdeling heeft, heeft zowel $\frac{\underline{x}-\mu}{\sigma}$ als $-\frac{\underline{x}-\mu}{\sigma}$ een $N(0,1)$-verdeling Daaruit volgt dat $\frac{\underline{z}-\mathcal{E}\underline{z}}{\sigma(\underline{z})}$ een $N(0,1)$-verdeling heeft en dus $\underline{z}$ normaal verdeeld is. Wij vinden dus:

Eigenschap 1: Als $\underline{x}$ een $N(\mu,\sigma)$-verdeling heeft, heeft $a\underline{x}+c$ een $N(a\mu+c,\ |a|\sigma)$-verdeling.

Voorbeeld:

Gevraagd: de rechtseenzijdige overschrijdingskans van $\underline{z}=0{,}32$ als $\underline{z}=-2\underline{x}+4$, waarin $\underline{x}$ $N(-1,2)$ verdeeld is.

Oplossing:

$$\mathcal{E}\underline{z}=-2\,\mathcal{E}\underline{x}+4=(-2)(-1)+4=6,$$

$$\sigma(\underline{z})=|-2|\cdot\sigma(\underline{x})=2\times 2=4.$$

$\underline{z}$ heeft dus een $N(6,4)$-verdeling. Dus:

$$P[\underline{x}\geq 0{,}32]=P\left[y\geq\frac{0{,}32-6}{4}\right]=P[y\geq -1{,}42]$$

$$=1-0{,}0778=0{,}9222. \qquad (y: N(0,1))$$

Eigenschap 2: Als $\underline{x}_1$ en $\underline{x}_2$ onderling onafhankelijk normaal verdeeld zijn, dan is ook $\underline{x}_1 + \underline{x}_2$ normaal verdeeld.

Als gegeven is dat $\underline{x}_1$ $N(\mu_1, \sigma_1)$ en $\underline{x}_2$ $N(\mu_2, \sigma_2)$ verdeeld is, volgt nu uit de eigenschappen 3(pag. 36) en 8(pag. 47) gemakkelijk dat $\underline{x}_1 + \underline{x}_2$ een $N(\mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_2^2})$ -verdeling heeft.

Voorbeeld: $\underline{x}_1$ is $N(0,3)$-verdeeld
$\underline{x}_2$ is $N(-1,4)$ "
Bepaal: $P[\underline{x}_1 + \underline{x}_2 \geqq 0]$.

Oplossing: $\underline{x}_1 + \underline{x}_2$ is $N(-1, \sqrt{3^2+4^2})$-verdeeld. Dus $P[\underline{x}_1 + \underline{x}_2 \geqq 0] =$ $(y:N(0,1))$

$= P[y \geqq \frac{0+1}{5}] = P[y \geqq 0,2] = 0,4207.$

Wij zullen eigenschap 2 niet bewijzen, doch onmiddellijk overgaan op de consequenties ervan.

Door combinatie van de eigenschappen 1 en 2 kunnen wij de verdeling vinden van iedere lineaire combinatie van normaal verdeelde grootheden. Wij vinden dan de volgende eigenschap:

Eigenschap 3: Als $\underline{x}_1, \ldots\ldots, \underline{x}_n$ onderling onafhankelijk normaal verdeelde grootheden zijn, waarbij $\underline{x}_i$ een $N(\mu_i, \sigma_i)$ -verdeling heeft, dan heeft

$$\underline{z} = \sum_{i=1}^{n} a_i \underline{x}_i \qquad (a_i \text{ constant})$$

een normale verdeling met verwachting:

$$\mathcal{E}\,\underline{z} = \sum_{i=1}^{n} a_i \mu_i$$

en spreiding

$$\sigma(\underline{z}) = \sqrt{\sum_{i=1}^{n} a_i^2 \sigma_i^2}\,.$$

De uitdrukkingen voor $\mathcal{E}\,\underline{z}$ en $\sigma(\underline{z})$ zijn reeds afgeleid in het vorige hoofdstuk. Uit eigenschap 1 volgt dat de grootheden $a_i \underline{x}_i$ alle normaal verdeeld zijn; door herhaalde toepassing van eigenschap 2 volgt hieruit eigenschap 3.

De belangrijkste toepassing van eigenschap 3 hebben wij in het geval, dat $\underline{x}_1, \ldots\ldots, \underline{x}_n$ alle dezelfde $N(\mu, \sigma)$ -verdeling hebben en dat $a_1 = a_2 = \ldots\ldots = a_n = \frac{1}{n}$. Er geldt dan:

$$\underline{z} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i \qquad \text{met } \mathcal{E}\,\underline{z} = \mu \qquad \text{en} \qquad \sigma(\underline{z}) = \frac{\sigma}{\sqrt{n}}\,.$$

In dat geval geldt dus, dat $\underline{z}$ een $N(\mu, \frac{\sigma}{\sqrt{n}})$ -verdeling heeft. Wij vinden langs deze weg de verdeling van het gemiddelde van n waarnemingen uit een $N(\mu, \sigma)$-verdeling.

Voorbeeld:

Gegeven de volgende 5 waarnemingen:
2,95; 0,91; 0,49; -0,62; 1,13.
Gevraagd: de tweezijdige overschrijdingskans van het gemiddelde van deze waarnemingen, als aangenomen wordt dat het onderling onafhankelijke waarnemingen uit een $N(0,1)$-verdeling zijn.
Oplossing: Men vindt voor het gemiddelde: $\bar{x} = \frac{1}{5} \cdot 4,86 = 0,972$.
Als de waarnemingen afkomstig zijn uit een $N(0,1)$-verdeling, heeft $\bar{x}$ een $N(0, \frac{1}{\sqrt{5}})$-verdeling. De tweezijdige overschrijdingskans van 0,972 wordt dus:

$$2\,P[\bar{x} \geqq 0,972] = 2\,P[y \geqq 0,972\sqrt{5}] = 2\,P[y \geqq 2,173] \approx 0,03. \qquad (y:N(0,1))$$

Opgaven:
3.3.a. Geef de verdeling van $x_1 - x_2$ en van $4x_1 - 3x_2 - 1$ als x_1 en x_2 onderling onafhankelijk $N(\mu,1)$-verdeeld zijn.
3.3.b. Gegeven zijn 5 waarnemingen uit een $N(\mu,1)$-verdeling:
0,02; 1,46; 0,40; 0,69; -1,30.
Gevraagd wordt μ zó te bepalen,
1) dat de rechtseenzijdige overschrijdingskans van het gemiddelde van deze waarnemingen juist gelijk is aan 0,05,
2) dat de linkseenzijdige overschrijdingskans van het gemiddelde van deze waarnemingen juist gelijk is aan 0,05,
3) dat de tweezijdige overschrijdingskans van het gemiddelde juist gelijk is aan 0,05.

Geef in ieder van deze gevallen tevens aan voor welke waarden van μ de overschrijdingskans groter en voor welke zij kleiner is dan 0,05.

3.4 Centrale limietstelling.

In de vorige paragraaf hebben wij gezien, dat een lineaire combinatie van n onafhankelijke normaal verdeelde grootheden zelf normaal verdeeld is. Men kan nu aantonen, dat dit bij benadering onder bepaalde niet ernstige voorwaarden ook geldt voor een lineaire combinatie van onderling onafhankelijke grootheden met een willekeurige verdeling. Dit is een consequentie van de centrale limietstelling, één van de belangrijkste stellingen van de wiskundige statistiek, die wij als volgt kunnen formuleren:
Als $x_1, x_2, \ldots$ onderling onafhankelijke stochastische grootheden zijn met $\mathcal{E}\,x_i = \mu_i$ en $\sigma(x_i) = \sigma_i$, dan convergeert voor $n \to \infty$ de verdelingsfunctie van

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

naar de verdelingsfunctie van een $N(0,1)$-verdeling. Er is aan deze stelling nog een extra voorwaarde verbonden, die hierop neerkomt, dat de staarten van de samenstellende verdelingen niet te dik mogen zijn. Deze voorwaarde is in ieder geval vervuld, als de samenstellende variabelen verdelingen hebben, die alleen waarden aannemen, welke in een begrensd interval liggen.

Indien een stochastische variabele beschouwd kan worden als de som van een groot aantal onafhankelijke stochastische grootheden, zal zij volgens de Centrale limietstelling dus bij benadering normaal verdeeld zijn. Dit is bijvoorbeeld vaak het geval bij waarnemingsfouten, die optreden bij het meten van physische grootheden. Gauss heeft hiervoor onder bepaalde onderstellingen reeds afgeleid, dat zij normaal verdeeld zijn. De normale verdeling wordt in verband hiermee wel de "foutenwet" genoemd. Toch is het gevaarlijk, zonder nader onderzoek, deze "foutenwet" bij alle mogelijke soorten van waarnemingsfouten toe te passen. De voorwaarden nodig voor de geldigheid van de Centrale limietstelling zijn voor vele soorten van fouten niet vervuld en het is in ieder geval zeer moeilijk te verifiëren of bepaalde fouten samengesteld zijn uit een groot aantal onafhankelijke stochastische variabelen. Indien men een essentiëel gebruik wil maken van de normaliteit van de verdeling van de waarnemingsfouten verdient het aanbeveling dit eerst te onderzoeken. Wij zullen hier niet ingaan op de methoden waarmee dit gedaan kan worden.

Een voor de toepassingen belangrijk gevolg van de Centrale limietstelling is, dat de som van een groot aantal stochastische grootheden bij benadering normaal verdeeld is. Indien $\underline{x}_1, \ldots\ldots, \underline{x}_n$ alle dezelfde verdeling hebben met verwachting μ en spreiding σ, heeft volgens de Centrale limietstelling

$$\frac{\sum_{i=1}^{n} \underline{x}_i - n\mu}{\sigma\sqrt{n}}$$

een verdeling, die voor $n \to \infty$ convergeert naar een $N(0,1)$-verdeling.

Voorbeeld: Wij hebben in hoofdstuk 2 gezien dat een stochastische variabele $\underline{x}$, die binomiaal verdeeld is met parameters n en p, beschouwd kan worden als de som van n onafhankelijke stochastische grootheden $\underline{x}_i$, die alle dezelfde verdeling hebben. Deze verdeling is een alternatief met kansen: $P[\underline{x}_i = 1] = p$ en $P[\underline{x}_i = 0] = 1-p$ waaruit volgt: $\mathcal{E}\,\underline{x}_i = p$ en $\sigma(\underline{x}_i) = \sqrt{p(1-p)}$.

De verdeling van:

$$y = \frac{x - np}{\sqrt{np(1-p)}}$$

zal dus voor voldoend grote n bij benadering een $N(0,1)$-verdeling moeten zijn.

Het kan misschien bevreemding wekken, dat een discrete verdeling als de binomiale, hier benaderd wordt door een continue normale verdeling. Men moet hierbij echter bedenken, dat de waarden die x aanneemt op afstand 1 van elkaar liggen, doch dat de waarden die y aanneemt dichter bij elkaar liggen naarmate n groter is, doordat de spreiding van de binomiale verdeling waardoor $x - np$ gedeeld wordt om y te verkrijgen bij toenemende n voortdurend aangroeit. Als n toeneemt, krijgt dus de gereduceerde en gestandaardiseerde binomiale variabele y steeds meer een continu karakter. Wij komen op de benadering van een binomiale verdeling door een normale nog terug in paragraaf 3.6.

Uit het voorafgaande volgt, dat ook het gemiddelde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ van onafhankelijke stochastische variabelen $x_1, \ldots\ldots, x_n$ met dezelfde verdeling $(\mathcal{E}x_i = \mu\ ;\ \sigma(x_i) = \sigma)$ bij benadering normaal verdeeld zal zijn en wel heeft $\frac{x-\mu}{\sigma}\sqrt{n}$ bij benadering een $N(0,1)$-verdeling. Dit betekent, dat het gemiddelde van een grote steekproef van n onafhankelijke waarnemingen uit dezelfde verdeling bij benadering normaal verdeeld is. Dit geldt, zoals wij in de vorige paragraaf gezien hebben, exact voor een steekproef uit een normale verdeling

Deze eigenschap van het gemiddelde wordt in de practijk ook zeer vaak toegepast. Op grond hiervan kan men methoden, die strikt genomen alleen gelden voor normaal verdeelde grootheden, dikwijls ook toepassen op steekproeven uit verdelingen, die niet normaal zijn. Men moet hiermee echter ook voorzichtig zijn. De nauwkeurigheid van de benadering hangt namelijk niet alleen af van het aantal waarnemingen, doch ook van het karakter van de verdeling waaruit de steekproef genomen wordt. Vooral als deze verdeling sterk asymmetrisch is, zal een groot aantal waarnemingen vereist zijn vóór het gemiddelde daarvan redelijk normaal verdeeld is. Bovendien is het mogelijk, dat de verdeling niet voldoet aan de door ons niet geformuleerde voorwaarde voor de geldigheid der Centrale limietstelling.

Opgave:

3.4.a. Laat $\underline{x}$ een stochastische variabele zijn met een binomiale verdeling met parameters n en p . Geef een benadering voor de kans dat $\frac{\underline{x}}{n}$ meer dan 0,01 van p afwijkt voor de volgende gevallen:

$p = 0,5 \quad n = 10^2;\ 10^4;\ 10^5$

$p = 0,2 \quad n =$ " ; " ; " .

3.5 Combinatie van onafhankelijke toetsen.

Eigenschap 3 behandeld in par. 3.3 en de Centrale limietstelling vinden een belangrijke toepasing bij het combineren van onafhankelijke toetsen. Wij zullen deze methode behandelen aan de hand van een voorbeeld, ontleend aan de "Elementaire statistische opgaven met uitgewerkte oplossingen", Rapport S 158 van het Mathematisch Centrum van Prof.Dr J. Hemelrijk en Ir Doralien Wabeke (opgave 48). Wij maken hierbij gebruik van de toets van WILCOXON.Voor de beschrijving hiervan verwijzen wij naar hoofdstuk 8 van de Cursus Toegepaste Statistiek I (Rapport S 120).

Wij vermelden hier alleen enkele bijzonderheden, die in de Cursus Toegepaste Statistiek I niet behandeld konden worden, omdat de daarvoor vereiste theorie nog niet uiteengezet was. Als de U van Wilcoxon berekend wordt uit een steekproef van m en een van n waarnemingen, geldt, onder de hypothese H_o , dat deze steekproeven afkomstig zijn uit dezelfde wh-verdeling:

(3.5.1) $$\mathcal{E}\,\underline{U} = \tfrac{1}{2} m n \qquad \text{en}$$

(3.5.2) $$\sigma^2(\underline{U}) = \tfrac{1}{2} m n (m+n+1).$$

Formule (3.5.1) geldt algemeen, formule (3.5.2) geldt alléén, indien er in de steekproeven geen gelijke waarnemingen voorkomen; de exacte formule voor $\sigma^2(\underline{U})$ in het geval er wel gelijke waarnemingen voorkomen, hebben wij behandeld in par. 8.8 van bovengenoemde cursus.

Verder nadert, als m en n beide toenemen, de verdeling van $\underline{U}$ onder H_o tot een normale verdeling. Dit wil dus zeggen, dat voor voldoend grote m en n :

$$\underline{W} = \frac{\underline{U} - \frac{1}{2} m n}{\sigma(\underline{U})}$$

bij benadering $N(0,1)$-verdeeld is.

Wij hebben deze eigenschappen reeds in par. 8.7 van genoemde cursus toegepast voor de berekening van overschrijdingskansen. Er werd daarbij nog een zogenaamde continuïteitscorrectie toegepast; d.w.z. bij een gegeven waarde van u werd niet de overschrijdingskans van

$$\frac{u - \frac{1}{4}mn}{\sigma(u)}$$

doch van een quotiënt, waarbij de absolute waarde van de teller $\frac{1}{2}$ lager was, in de $N(0,1)$-tabel opgezocht. De betekenis hiervan wordt in par 3.6 nog toegelicht. De afleidingen van deze stellingen zullen wij hier niet geven.

De opgave luidt als volgt:

Door een bepaalde wijze van prepareren (methode A) hoopt een textielfabrikant zijn garens sterker te maken dan zij zonder toepassing daarvan zijn. Om dit te onderzoeken vervaardigt hij van twee soorten garens, die hij gewoonlijk in grote hoeveelheden produceert een partij met en zonder toepassing van methode A. Vervolgens past hij trekproeven op de zo verkregen garens toe en hij vindt de volgende numerieke resultaten (uitgedrukt in een of andere sterkte-eenheid):

Garensoort

I	II	
13,37	11,46	zonder A
12,66	10,69	
12,10	10,19	
13,20	10,34	
11,40	11,79	
12,97	11,52	met A
13,62	12,63	
12,54	10,59	
13,41	11,64	
13,57	11,70	

Ga met behulp van een toets met onbetrouwbaarheid $\alpha = 0,05$ na:
a) of de garensoorten I en II verschillend in sterkte zijn
b) of de fabrikant er wijs aan doet methode A in te voeren of niet.

Oplossing:

a) Met de toets van Wilcoxon toetsen wij de hypothese H_0, dat de garens I en II even sterk zijn. Deze toets voeren wij uit zowel met de waarnemingen der trekproeven van de met A behandelde, als de niet met A behandelde garens. Dit doen wij, omdat wij van te voren niet weten of A invloed heeft of niet; als A invloed heeft

kunnen wij niet meer aannemen, dat de 10 waarnemingen van b.v. garensoort I afkomstig zijn uit één verdeling. Als wij de niet met A behandelde garens rangschikken naar opklimmende waarde van de treksterkten verkrijgen wij het volgende resultaat:

10,19	10,34	10,69	11,40	11,46	11,79	12,10	12,66	13,20	13,37
II	II	II	I	II	II	I	I	I	I

Wij bepalen nu de toetsingsgrootheid u ; dit is het aantal keren, dat een II vóór een I staat; men vindt dan:

$$u = 5+5+5+3+3 = 23.$$

Volgens tabel 8.II van de Cursus Toegepaste Statistiek I is de linker kritieke waarde van de tweezijdige toets van Wilcoxon bij $\alpha = 0,05$ gelijk aan 2, volgens formule (8.6.1) uit dezelfde cursus is dus de rechter kritieke waarde: 25-2=23. Wij moeten bij $\alpha = 0,05$ dus de hypothese H_0 verwerpen en wij kunnen concluderen, dat zonder behandeling met A garensoort I systematisch sterker is dan garensoort II.

Wij passen nu ook de toets van Wilcoxon toe bij de met A behandelde garens. De naar opklimmende grootte gerangschikte treksterkten worden:

10,59	11,52	11,64	11,70	12,54	12,63	12,97	13,41	13,57	13,62
II	II	II	II	I	II	I	I	I	I

Men vindt hieruit $u = 4x5+1x4 = 24$

Ook in dit geval moet dus H_0 verworpen worden. Weer vindt men, dat I systematisch sterker is dan II.

Wij concluderen dus dat zowel voor de behandelde als voor de niet behandelde garens de hypothese: de garens zijn even sterk, verworpen moet worden ten gunste van de hypothese dat I sterker is dan II, behoudens een onbetrouwbaarheid 0,05.

b) Wij willen nu de hypothese H_0 toetsen, dat het prepareren van de garens geen invloed heeft op de sterkte. Wij hebben reeds geconstateerd dat I en II verschillen en kunnen dus de waarnemingen "zonder A" niet als één steekproef beschouwen en evenmin de de waarnemingen "met A". Wij zullen dus de toets van Wilcoxon afzonderlijk voor de garensoorten I en II moeten toepassen. Wij hebben dit hieronder gedaan; B betekent: zonder A. Garensoort I:

11,40	12,10	12,54	12,66	12,97	13,20	13,37	13,41	13,57	13,62
B	B	A	B	A	B	B	A	A	A

Wij vinden nu, als u het aantal keren is, dat een A vóór een B staat:: $u = 3+2 = 5$

Op grond hiervan kan H_0 niet verworpen worden.

Garensoort II.

10,19	10,34	10,59	10,69	11,46	11,52	11,64	11,70	11,79	12,63
B	B	A	B	B	A	A	A	B	A

$U = 3+3x1=6$.

Op grond hiervan kan H_o evenmin verworpen worden.

Voor de garens I en II kunnen wij dus de hypothese, dat toepassing van methode A geen invloed heeft op de sterkte, niet verwerpen. Wel constateren wij, dat in beide gevallen de toetsingsgrootheid U aanzienlijk lager is dan haar verwachtingswaarde: $\frac{1}{2}.5\ 5=12\frac{1}{2}$ (vgl. 3.5.1). De vraag is nu, of de combinatie van beide resultaten niet kan leiden tot verwerping van de hypothese H_o.

De hiervoor in aanmerking komende toetsingsgrootheid is $V = U_I + U_{II}$, als U_I de grootheid U voorstelt verkregen bij garensoort I en U_{II} de grootheid verkregen bij garensoort II. Men kan de verdeling van $\underline{V} = \underline{U}_I + \underline{U}_{II}$ onder de hypothese H_o berekenen, als men de verdelingen van $\underline{U}_I$ en $\underline{U}_{II}$ afzonderlijk kent en gebruik maakt van het feit, dat beide grootheden onderling onafhankelijk zijn. Wij zullen dit hier niet doen, doch gebruik maken van een benaderingsmethode, die berust op eigenschap 2, die een bijzonder geval is van eigenschap 3.

Het is bekend, dat de toetsingsgrootheid $\underline{U}$ van Wilcoxon onder de hypothese H_o een symmetrische discrete verdeling heeft (mits, zoals hier het geval is, er geen gelijke waarnemingen voorkomen). Als m en n, de aantallen waarnemingen van beide steekproeven, toenemen, nadert, zoals vermeld, de verdeling van $\underline{U}$ tot een normale verdeling. Bij $m=n=5$ is de benadering door middel van een normale verdeling nog niet erg nauwkeurig. Als $\underline{U}_I$ en $\underline{U}_{II}$ exact normaal verdeeld waren, zou $\underline{V}$ volgens eigenschap 2 ook exact normaal verdeeld zijn. Nu $\underline{U}_I$ en $\underline{U}_{II}$ slechts bij benadering normaal verdeeld zijn, zal dit ook voor $\underline{V}$ het geval zijn. Het blijkt, dat de normale benadering voor $\underline{V}$ veel beter voldoet dan voor $\underline{U}_I$ en $\underline{U}_{II}$ afzonderlijk. Dit behoeft ons niet te verwonderen, aangezien uit de Centrale limietstelling bekend is, dat zelfs de som van een aantal willekeurig verdeelde onafhankelijke stochastische grootheden bij benadering normaal verdeeld is.

Wij benaderen hier de verdeling van $\underline{V} = \underline{U}_I + \underline{U}_{II}$ onder H_o dus met

een normale verdeling. Wij moeten daartoe nog verwachting en spreiding van $\underline{V}$ kennen. Hiervoor geldt;

$$\mathcal{E}\,\underline{V} = \mathcal{E}\,\underline{U}_{I} + \mathcal{E}\,\underline{U}_{II} = 2 \times \tfrac{1}{2} \times 5 \times 5 = 25 \quad \text{(zie (3.5.1))}$$

$$\sigma^2(\underline{V}) = \sigma^2(\underline{U}_{I}) + \sigma^2(\underline{U}_{II}) = 2 \times \tfrac{1}{12} \times 5 \times 5 \times (5+5+1) = 45{,}83$$

dus $\sigma(\underline{V}) = 6{,}77$.

Wij zoeken nu in de tabel van de $N(0,1)$ de tweezijdige overschrijdingskans van

$$\frac{V - \mathcal{E}\,\underline{V} + \frac{1}{2}}{\sigma(\underline{V})} = \frac{5+6-25+\frac{1}{2}}{6{,}77} = -1{,}99.$$

De tweezijdige overschrijdingskans wordt:

$$\ell = 2 \times 0{,}0233 = 0{,}047.$$

Conclusie: Op grond van dit resultaat kunnen wij de hypothese H_o verwerpen. Aangezien zowel bij garensoort I als bij garensoort II bleek, dat het vaker voorkwam, dat een met A behandeld garen sterker was dan een niet met A behandeld garen dan omgekeerd, verwerpen wij H_o ten gunste van de hypothese dat methode A de garens sterker maakt. Het is dus aan te raden methode A in te voeren.

Opmerking 1: Indien het invoeren van methode A grote kosten met zich meebrengt, verdient het wellicht aanbeveling een grotere proef te verrichten om met grotere stelligheid tot een conclusie te kunnen komen.

Opmerking 2: De hier behandelde methode kan worden gebruikt voor de combinatie van willekeurig veel toetsen. Indien men bijvoorbeeld niet met 2 maar met n onderling verschillende garensoorten te doen heeft, en bij soort i steekproeven met m_i en n_i waarnemingen worden genomen, waarbij de toetsingsgrootheid U_i van Wilcoxon wordt berekend, dan wordt de toetsingsgrootheid van de combinatiemethode:

$$\underline{V} = \sum_{i=1}^{n} \underline{U}_i$$

met verwachting: $\mathcal{E}\,\underline{V} = \frac{1}{2} \sum_{i=1}^{n} m_i n_i$

en variantie

$$\sigma^2(\underline{V}) = \frac{1}{12} \sum_{i=1}^{n} m_i n_i (m_i + n_i + 1)$$

(afgezien van gelijke waarnemingen).

$\dfrac{\underline{V} - \mathcal{E}\,\underline{V}}{\sigma(\underline{V})}$ heeft dan bij benadering een $N(0,1)$-verdeling.

De methode kan ook worden toegepast voor de combinatie van andere toetsen, dan de toets van Wilcoxon. Een voorbeeld hiervan zullen wij in par. 3.6 behandelen.

Opgaven:

3.5.a. Hoe zou men de exacte verdeling van $U_I + U_{II}$ onder de hypothese H_0 berekenen, als de verdeling van U voor $m=n=5$ gegeven is?

3.5.b. Combineer bij het voorbeeld van de garens de toetsen voor het verschil tussen de garensoorten I en II, voor de gevallen zonder A en met A.

3.5.c. Veronderstel, dat de garenfabrikant de sterkteproeven ook nog bij een garensoort III heeft uitgevoerd met de volgende resultaten:

Zonder A: 10,31; 9,87; 10,03; 10,12; 10,45; 9,98,

met A: 10,08; 10,17; 10,15; 10,50; 9,93; 10,41.

Ga na:

1) of garensoort III in sterkte verschilt van garensoort I en garensoort II,
2) of bij garensoort III de hypothese verworpen kan worden, dat methode A geen invloed heeft op de sterkte,
3) combineer de toetsen voor het verschil tussen sterkten, verkregen met en zonder de methode A bij de 3 garensoorten.

Opmerking: Bij een onbetrouwbaarheidsdrempel 0,05 zijn de kritieke waarden van de tweezijdige toets van Wilcoxon voor $m=n=6$: $u_{\ell}=5$ en $u_{r}=31$.

3.6 Normale verdeling als benadering van discrete verdelingen.

Wij hebben in par. 1.6 gezien, dat men uit practische overwegingen vaak discrete verdelingen door continue benadert. Dit geschiedt vooral in de gevallen, dat men discrete verdelingen nodig heeft, die niet getabelleerd zijn en slechts ten koste van omvangrijke berekeningen verkregen kunnen worden. Bovendien zijn de benaderende continue verdelingen dikwijls veel gemakkelijker hanteerbaar dan de corresponderende exacte.

Een groot aantal discrete verdelingen kan goed benaderd worden met behulp van een normale verdeling. Dit is dan meestal een gevolg van het feit, dat deze verdelingen asymptotisch normaal zijn, hetgeen wil zeggen, dat de verdelingsfuncties van deze verdelingen naderen tot de verdelingsfunctie van een normale verdeling als een of meer parameters van de verdeling onbeperkt aangroeien. Dit is bijvoorbeeld het geval bij:

1) De binomiale verdeling, als het aantal experimenten n aangroeit,
2) De verdeling van de toetsingsgrootheid $\underline{U}$ van Wilcoxon onder de hypothese H_0, als de omvang van beide steekproeven (m en n) aangroeit,
3) De verdeling van de som van n onafhankelijke stochastische grootheden als n aangroeit.

In al deze gevallen benadert men de discrete verdeling steeds met een normale verdeling, die hetzelfde gemiddelde en dezelfde spreiding heeft.

Als voorbeeld geven wij hier de benadering van een binomiale verdeling met $n=10$ en $p=0,4$. Een indruk van deze benadering hebben wij reeds verkregen uit fig. 1.6 op blz 12. Voor deze binomiale verdeling geldt volgens hoofdstuk 2:

$$\mathcal{E}_{\underline{x}} = np = 10 \cdot 0,4 = 4 \qquad \text{en}$$

$$\sigma(\underline{x}) = \sqrt{np(1-p)} = \sqrt{10 \cdot 0,4 \cdot 0,6} = \sqrt{2,4} = 1,549.$$

Als benadering gebruiken wij dus een $N(4; 1,549)$-verdeling. Als wij nu de kans $P[\underline{x}=3]$ van de binomiale verdeling willen benaderen, kunnen wij niet de overeenkomstige kans van de normale verdeling gebruiken, want deze is steeds 0, omdat de normale verdeling continu is. Wij kiezen daarom de kans, dat de normaal verdeelde variabele, hier aan te geven met $\underline{x}'$ een waarde aanneemt in het interval $(2\frac{1}{2}, 3\frac{1}{2})$, d.w.z. een interval van de breedte 1 waarvan 3 het midden is. Wij benaderen nu $P[\underline{x}=3]$ ($\underline{x}$ binomiaal verdeeld met $p=0,4$ en $n=10$) met $P[2\frac{1}{2} < \underline{x}' < 3\frac{1}{2}]$ ($\underline{x}'$ $N(4; 1,549)$ -verdeeld). Deze laatste kans kan gemakkelijk berekend worden met de tabel van de $N(0,1)$-verdeling. Men vindt als y een $N(0,1)$ -variabele is:

$$P[2\tfrac{1}{2} < \underline{x}' < 3\tfrac{1}{2}] = P\left[\frac{2\frac{1}{2}-4}{1,549} < y < \frac{3\frac{1}{2}-4}{1,549}\right] = P[-0,968 < y < -0,323] =$$

$$= P[0,323 < y < 0,968] = P[y \geq 0,323] - P[y \geq 0,968] = 0,207.$$

Op analoge wijze kan men $P[\underline{x}=1], \ldots\ldots, P[\underline{x}=9]$ benaderen. Indien men wil bereiken dat de som van de benaderende whn 1 is, zal men $P[\underline{x}=0]$ nu moeten benaderen met $P[\underline{x}' < \frac{1}{2}]$ en $P[\underline{x}=10]$ met $P[\underline{x}' > 9\frac{1}{2}]$. De exacte en de benaderende whn van deze binomiale verdeling hebben wij gegeven in tabel 3.I.

Tabel 3.I.

Benadering van de Binomiale verdeling ($n = 10$ $p = 0,4$) met de aangepaste normale verdeling ($N(4; 1,549)$)

$P[\underline{x}=x]$ \ x	0	1	2	3	4	5	6	7	8	9	10
exact	0,006	0,040	0,121	0,210	0,251	0,201	0,111	0,042	0,011	0,002	0,000
benaderd	0,012	0,041	0,113	0,207	0,253	0,207	0,113	0,041	0,010	0,002	0,000

De linkseenzijdige overschrijdingskans $P[\underline{x} \leqq 3]$ is bij de binomiale verdeling gelijk aan : $P[\underline{x}=0] + P[\underline{x}=1] + P[\underline{x}=2] + P[\underline{x}=3]$. De som van de overeenkomstige kansen bij de normale verdeling wordt: $P[\underline{x}' < \frac{1}{2}] + P[\frac{1}{2} < \underline{x}' < 1\frac{1}{2}] + P[1\frac{1}{2} < \underline{x}' < 2\frac{1}{2}] + P[2\frac{1}{2} < \underline{x}' < 3\frac{1}{2}] = P[\underline{x}' < 3\frac{1}{2}]$. Wij zien dus, dat $P[\underline{x} \leqq 3] = 0{,}377$ wordt benaderd door: $P[\underline{x}' < 3\frac{1}{2}] = 0{,}373$. Evenzo wordt $P[\underline{x} \geqq 3]$ benaderd door $P[\underline{x}' > 2\frac{1}{2}]$. In het algemeen wordt de linkseenzijdige overschrijdingskans van x bij de binomiale verdeling benaderd door de linkseenzijdige overschrijdingskans van $x + \frac{1}{2}$ bij de normale verdeling en de rechtseenzijdige overschrijdingskans van x bij de binomiale verdeling benaderd door de rechtseenzijdige overschrijdingskans van $x - \frac{1}{2}$ bij de normale verdeling. Dus $P[\underline{x} \leqq x]$ door $P[\underline{x}' \leqq x + \frac{1}{2}]$; $P[\underline{x} \geqq x]$ door $P[\underline{x}' \geqq x - \frac{1}{2}]$.

De term $\frac{1}{2}$ die hier optreedt wordt continuïteitscorrectie genoemd, omdat zij voortkomt uit het feit, dat men een discrete verdeling benadert door een continue.

De continuïteitscorrectie bij een tweezijdige overschrijdingskans kan men direct uit het voorgaande afleiden, als men bedenkt, dat zij steeds een som is van een rechts- en een linkseenzijdige overschrijdingskans. Bij symmetrische verdelingen kan men als eenvoudige regel geven, dat men als benadering de tweezijdige overschrijdingskans volgens de normale verdeling gebruikt van een waarde, die $\frac{1}{2}$ dichter bij het gemiddelde van de verdeling ligt dan de waarde waarvan men de tweezijdige overschrijdingskans moet bepalen.

Voorbeeld: Benader bij de toets van Wilcoxon de tweezijdige overschrijdingskans van $U = 9$ als $m = 10$ en $n = 12$ voor het geval, dat er in de steekproeven geen gelijke waarnemingen voorkomen.

Oplossing: Volgens de formules (3.5.1) en (3.5.2) geldt onder H_0 :

$$\mathcal{E}\,\underline{U} = \tfrac{1}{2} \cdot 10 \cdot 12 = 60$$

$$\sigma^2(\underline{U}) = \tfrac{1}{12} \cdot 10 \cdot 12 \cdot 23 = 230 \quad ; \quad \sigma(\underline{U}) = \sqrt{230} = 15{,}17.$$

Wij gebruiken als benadering dus een $N(60;15{,}17)$-verdeling. Wij bepalen daarbij de tweezijdige overschrijdingskans van de waarde die $\frac{1}{2}$ dichter bij het gemiddelde ligt dan 9, dus van $9\frac{1}{2}$. Wij vinden, dat deze kans wordt:

$$2\,P\left[y \geqq \frac{50\frac{1}{2}}{15{,}17}\right] = 2\,P[y \geqq 3{,}329]$$

$$= 2 \times 0{,}0004 = 0{,}0008.$$

In het algemeen zal men de hier beschreven continuïteits-correctie toepassen, als het gaat om de benadering van discrete verdelingen, waarbij de variabele waarden aanneemt, die op afstand 1 van elkaar liggen. Als de afstand tussen de opeenvolgende waarden die worden aangenomen, niet één maar wel constant, b.v. c is, dan wordt eenzelfde continuïteitscorrectie ter grootte van $\frac{1}{2}c$ toegepast. Er zijn echter ook veel discrete verdelingen, waarbij tussen opeenvolgende waarden geen constant interval ligt. Dit is bijvoorbeeld het geval bij de verdeling van de toetsingsgrootheid van Wilcoxon, als wij te doen hebben met steekproeven, waarin groepen gelijke waarnemingen voorkomen. Het is in dergelijke gevallen veel moeilijker om een goede continuïteits-correctie aan te geven. Wij willen op de complicaties, die bij dergelijke onregelmatige verdelingen kunnen optreden hier verder niet ingaan. Wij moeten echter waarschuwen voor het gebruik van de normale benadering voor de verdeling van de toetsingsgrootheid $\underline{u}$ van WILCOXON, in het geval, dat in de steekproeven een groep gelijke waarnemingen voorkomt, waartoe het grootste gedeelte van het waarnemingsmateriaal behoort. Toepassing van de normale benadering kan dan tot geheel verkeerde conclusies leiden.

Opgaven:

3.6.a. Welke normale verdeling kan als benadering gebruikt worden voor de binomiale verdeling met $n=10$ en $p=0,5$? Bereken hierbij de exacte en de benaderde waarde voor $P[x=4]$, voor $P[x \geq 1]$ en voor de tweezijdige overschrijdingskans van $x=2$.

3.6.b. Welke normale verdeling kan als benadering gebruikt worden voor de som van 10 onafhankelijke variabelen, alle homogeen verdeeld tussen 0 en 1? Geef de benaderde waarde voor de tweezijdige overschrijdingskans van $x=9$ bij deze verdeling. Welke continuïteitscorrectie zou U hier gebruiken?

3.7 Aanpassing van een normale verdeling aan een steekproef.

In fig. 1.5 op blz. 11 hebben wij een histogram gegeven van de schedelbreedten in Engelse duimen van 1000 studenten uit Cambridge. Wij hebben in deze figuur tevens het histogram en de verdelingsdichtheidscurve der aangepaste normale verdeling getekend. In deze paragraaf willen wij aan de hand van hetzelfde voorbeeld uiteenzetten, hoe men een normale verdeling aan steekproefmateriaal aanpast.

Het principe van de aanpassing is zeer eenvoudig: men schat het gemiddelde en de spreiding van de verdeling uit de steekproef en kiest een normale verdeling, waarvan het gemiddelde en de spreiding gelijk zijn aan de geschatte waarden. De berekeningen, die hiervoor nodig zijn, vindt men in tabel 3.II.

In het genoemde voorbeeld was de steekproef ingedeeld in klassen naar de gemeten grootheid. De grenzen en de klassemerken van de klassen zijn in tabel 3.II resp. aangeduid met g_i en x_i. Het klassemerk x_i is het midden van de klasse, behalve bij de eenzijdig begrensde klassen 1 en 14, waar x_i op een afstand gelijk aan de helft van de breedte der overige klassen van de boven- resp. ondergrens gekozen is.

Wij berekenen nu uit de steekproef de grootheden $\bar{x}$ en s' zoals aangegeven bij de tabel. Dit zijn de schattingen van het gemiddelde resp. de spreiding van de verdeling. Wij gaan nu na wat de verwachtingswaarden zijn van de aantallen waarnemingen in de klassen, als men te doen heeft met een steekproef van 1000 waarnemingen uit een normale verdeling met gemiddelde $\bar{x} = 6,0615$ en spreiding $s' = 0,2012$. Deze verwachtingswaarden aangeduid met ν_i, berekenen wij als volgt:

$$\nu_i = 1000\, P[g_{i-1} \leqq x \leqq g_i] = 1000\, P\left[\frac{g_{i-1} - \bar{x}}{s'} \leqq y \leqq \frac{g_i - \bar{x}}{s'}\right],$$

waarin x een $N(\bar{x}, s')$- en y een $N(0,1)$-verdeelde variabele voorstelt. Voor de 1e klasse is $g_{i-1} = -\infty$ voor de laatste $g_i = +\infty$. Men vindt de gereduceerde en de gestandaardiseerde klasse-grenzen $\frac{g_i - \bar{x}}{s'}$ in tabel 3.II, evenals de gevonden waarden van ν_i. Van de ν_i zijn geen decimalen opgegeven, omdat deze, gezien de onnauwkeurigheid van de gegevens, weinig betrouwbaar zouden zijn. Tengevolge hiervan werd het totaal niet 1000 doch 999.

Indien wij de kolommen n_i en ν_i vergelijken, zien wij grote verschillen, zoals reeds bleek in fig. 1.5. Men kan zich afvragen of deze verschillen op een duidelijke afwijking van de normale verdeling wijzen. Dit kan voor een uitgebreid geklassificeerd steekproefmateriaal als het gegevene o.a. onderzocht worden met de chi-kwadraattoets voor aanpassing. Een behandeling van deze methode zou hier echter te ver voeren; bij toepassing ervan komt men tot de conclusie, dat de steekproef niet uit een normale verdeling afkomstig is, gezien de grootte van de overschrijdingskans kan deze conclusie met grote stelligheid getrokken worden.

Tabel 3.II

Aanpassing van een normale verdeling bij de frequentieverdeling van de schedelbreedten van 1000 studenten uit Cambridge.

Klasse No. i	g_i	x_i	n_i	$n_i x_i$	$n_i x_i^2$	$\frac{g_i - \bar{x}}{s'}$	ν_i
1	5,55	5,5	3	16,5	90,75	- 2,54	6
2	5,65	5,6	12	67,2	376,32	- 2,05	15
3	5,75	5,7	43	245,1	1397,07	- 1,55	40
4	5,85	5,8	80	464,0	2691,20	- 1,05	86
5	5,95	5,9	131	772,9	4560,11	- 0,55	141
6	6,05	6,0	236	1416,0	8496,00	- 0,06	188
7	6,15	6,1	185	1128,5	6883,85	0,44	194
8	6,25	6,2	142	880,4	5458,48	0,94	156
9	6,35	6,3	99	623,7	3929,31	1,43	97
10	6,45	6,4	37	236,8	1515,52	1,93	50
11	6,55	6,5	15	97,5	633,75	2,43	19
12	6,65	6,6	12	79,2	522,72	2,92	6
13	6,75	6,7	3	20,1	134,67	3,42	1
14	$+\infty$	6,8	2	13,6	92,48	∞	0
Totaal:	-	-	1000	6061,5	36782,23		1999

Toelichting:

g_i = bovengrens van i^e klasse

$x_i = g_i - 0,05$ = klassemerk, d.i. de waarde die men aan alle waarnemingen in een klasse toekent (bij klasse 14: $g_{i-1} + 0,05$)

n_i = aantal waarnemingen in i^e klasse

$\bar{x}$ = schatting van het gemiddelde der verdeling = $\frac{\sum_{i=1}^{14} n_i x_i}{1000}$;
d.i. het totaal van de kolom $n_i x_i$ gedeeld door 1000, dus $\bar{x} = 6,062$.

s' = schatting van de spreiding der verdeling. Wij gebruiken hiertoe de formules (2.14.3) en (2.14.4) en vinden:

$$s' = \sqrt{\frac{1}{999}\sum_{i=1}^{14} n_i (x_i - \bar{x})^2} = \sqrt{\frac{1}{999}\left(\sum_{i=1}^{14} n_i x_i^2 - 1000\,\bar{x}^2\right)} =$$

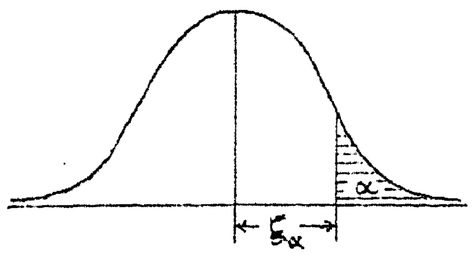
$$= \sqrt{\frac{1}{999}(36782,23 - 1000.(0,60615)^2} = \sqrt{0,040488} = 0,2012.$$

ν_i = aantal waarnemingen in i^e klasse volgens aangepaste normale verdeling (zie tekst)

De voornaamste afwijking, die de steekproef t.o.v. de aangepaste normale verdeling vertoont, is het feit, dat zij te "spits" is. De klasse No 6 is veel te zwaar bezet, terwijl de naburige klassen No 4,5,7 en 8 te weinig waarnemingen bevatten. Hierdoor is de verdeling duidelijk scheef geworden, want hoewel het gemiddelde van de steekproef in klasse No 7 valt, is de klasse No 6 aanzienlijk sterker bezet. Een dergelijke afwijking zou men ook kunnen constateren door berekening en toetsing van het 3e moment van de steekproef. Ook hierop kunnen wij hier niet nader ingaan.

TABEL VAN DE NORMALE VERDELING[1)]

Waarden van $\alpha.10^4$ voor $\xi_\alpha = 0{,}00(0{,}01)3{,}09$



met $\frac{1}{\sqrt{2\pi}} \int_{\xi_\alpha}^{\infty} e^{-\frac{x^2}{2}} dx = \alpha$.

ξ_α	0	1	2	3	4	5	6	7	8	9
0,0	5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
0,1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
0,2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
0,3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
0,4	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
0,5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
0,6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
0,7	2420	2389	2358	2327	2296	2266	2236	2206	2177	2148
0,8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
0,9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
1,0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1,1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
1,2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
1,3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
1,4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
1,5	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
1,6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
1,7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
1,8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
1,9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
2,0	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183
2,1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
2,2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
2,3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
2,4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
2,5	0062	0060	0059	0057	0055	0054	0052	0051	0049	0048
2,6	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
2,7	0035	0034	0033	0032	0031	0030	0029	0028	0027	0026
2,8	0026	0025	0024	0023	0023	0022	0021	0021	0020	0019
2,9	0019	0018	0018	0017	0016	0016	0015	0015	0014	0014
3,0	0013	0013	0013	0012	0012	0011	0011	0011	0010	0010
3,1	0010	0009	0009	0009	0008	0008	0008	0008	0007	0007
3,2	0007	0007	0006	0006	0006	0006	0006	0005	0005	0005
3,3	0005	0005	0005	0004	0004	0004	0004	0004	0004	0003
3,4	0003	0003	0003	0003	0003	0003	0003	0003	0003	0002

Een 5 betekent, dat bij afronding naar het dichtsbijgelegen oneven cijfer afgerond moet worden. Een 5 wordt naar het dichtsbijzijnde even cijfer afgerond.

[1)] Samengesteld met behulp van:

Tabel 1.5 uit: "Techniques of Statistical Analysis" van de Statistical Research Group, Columbia University, 1947.

Tabel I uit: "Tables of the error function and of its first twenty derivatives". The annals of the computation laboratory of Harvard University, Vol. XXIII, Harvard University Press, 1952.

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 155

Cursus

Toegepaste Statistiek II

door

Ph. van Elteren en J. Kriens

IV Correlatie

1955

4 Correlatie

4.1. Het begrip correlatie.

In par. 2.12 hebben wij kennis gemaakt met de covariantie van twee simultaan verdeelde stochastische variabelen $\underline{x}$ en $\underline{y}$. Deze was gedefinieerd als de verwachting $\mathcal{E}\,\tilde{\underline{x}}\tilde{\underline{y}}$ van het product der gereduceerde variabelen. Wij hebben daar reeds laten zien, dat $\mathcal{E}\,\tilde{\underline{x}}\tilde{\underline{y}} = 0$ is, als $\underline{x}$ en $\underline{y}$ onafhankelijk zijn.

Voor een discrete verdeling, waarbij $\underline{x}$ de waarden $x_1, \ldots, x_m$ en $\underline{y}$ de waarden $y_1, \ldots, y_n$ aanneemt, wordt $\mathcal{E}\,\tilde{\underline{x}}\tilde{\underline{y}}$ gegeven door:

$$(4.1.1) \qquad \mathcal{E}\,\tilde{\underline{x}}\tilde{\underline{y}} = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\,(x_i - \mathcal{E}\,\underline{x})(y_j - \mathcal{E}\,\underline{y}),$$

waarin $p_{ij} = P[\underline{x} = x_i \text{ én } \underline{y} = y_j]$ is.

Wij zien aan deze uitdrukking, dat $\mathcal{E}\,\tilde{\underline{x}}\tilde{\underline{y}}$ groot zal zijn, als grote waarden van $\underline{x}$ in het algemeen gepaard gaan met grote waarden van $\underline{y}$ en kleine waarden van $\underline{x}$ met kleine waarden van $\underline{y}$. Immers een grote waarde x_i van $\underline{x}$ gepaard met een grote waarde y_j van $\underline{y}$ betekent in het rechterlid van (4.1.1) een grote positieve waarde van $x_i - \mathcal{E}\,\underline{x}$ en van $y_j - \mathcal{E}\,\underline{y}$, dus ook een grote positieve waarde van $(x_i - \mathcal{E}\,\underline{x})(y_j - \mathcal{E}\,\underline{y})$; een kleine waarde van $\underline{x}$ gepaard met een kleine waarde van $\underline{y}$ betekent in het rechterlid van (4.1.1) een product van twee sterk negatieve factoren, dus ook een grote positieve bijdrage tot de som. Evenzo kunnen wij inzien, dat $\mathcal{E}\,\tilde{\underline{x}}\tilde{\underline{y}}$ sterk negatief zal zijn, als grote waarden van $\underline{x}$ in het algemeen gepaard gaan met kleine waarden van $\underline{y}$ en omgekeerd.

De hier beschreven vormen van afhankelijkheid van $\underline{x}$ en $\underline{y}$ noemt men correlatie. De covariantie van $\underline{x}$ en $\underline{y}$ is een maat voor hun correlatie; als $\mathcal{E}\,\tilde{\underline{x}}\tilde{\underline{y}}$ positief is, spreekt men van positieve, als $\mathcal{E}\,\tilde{\underline{x}}\tilde{\underline{y}}$ negatief is van negatieve correlatie. De correlatie is sterker (positief of negatief) naarmate de absolute waarde van $\mathcal{E}\,\tilde{\underline{x}}\tilde{\underline{y}}$ groter is.

Voorbeelden.

Grote mensen plegen in het algemeen zwaarder te zijn dan kleine. Als $\underline{x}$ dus de lengte en $\underline{y}$ het gewicht voorstelt van een persoon, die aselect gekozen is uit een mensenpopulatie, zal $\underline{x}$ positief gecorreleerd zijn met $\underline{y}$. De waterstand van b.v. de Rijn te Keulen, en de waterstand te Lobith nadat het water zich van Keulen naar Lobith verplaatst heeft, zijn positief gecorreleerd. De belasting, die men betaalt, is positief gecorreleerd met het inkomen etc.

Een negatieve correlatie bestaat b.v. tussen de bewolkingsgraad en het aantal uren zonneschijn per dag waargenomen; bij vele landbouwproducten tussen de hoeveelheid die in een jaar geoogst wordt en de prijs etc

Met nadruk zij er hier op gewezen, dat correlatie niet synoniem is met afhankelijkheid; twee grootheden kunnen sterk afhankelijk van elkaar zijn, zonder duidelijk gecorreleerd te zijn. Alle grootheden, die een zekere periodiciteit vertonen, zoals de temperatuur, de omzet van winkels, het ziekteverzuim van arbeiders e.d zullen in het algemeen afhankelijk zijn van de tijd, b.v het uur van de dag of de datum van het jaar waarop ze worden waargenomen. Men verwacht in een etmaal lagere temperaturen om 0 uur dan om 12 uur, en hogere temperaturen in Juli dan in December. Of men eventueel een correlatie vindt met de tijd hangt echter af van de periode waarin men de waarnemingen verricht. De temperatuur vertoont een positieve correlatie met de tijd in de ochtenduren, een negatieve in de namiddaguren; de daggemiddelden van de temperatuur vertonen een positieve correlatie met de tijd in de lente, een negatieve in de herfst. Beschouwt men echter een gehele dag, of een geheel jaar, dan zal men geen correlatie van betekenis vinden; de perioden van positieve correlatie zullen dan de perioden van negatieve correlatie compenseren

De tijdsvariabele behoeft in de voorafgaande voorbeelden niet stochastisch te zijn; dit is alleen het geval als de waarnemingen verricht worden op tijdstippen die aselect volgens een of andere wh-verdeling gekozen zijn. Als de waarnemingstijdstippen te voren vastgesteld zijn, is de tijdsvariabele niet stochastisch. Men kan in dergelijke gevallen toch de termen "afhankelijk", "gecorreleerd" enz. gebruiken om het verband tussen de stochastische en de niet stochastische grootheid te kenschetsen. De stochastische variabele y is afhankelijk van de niet stochastische variabele x, als de verdeling van y niet voor alle mogelijke waarden van x dezelfde is; y is b.v. positief gecorreleerd met x als y in het algemeen grotere waarden aanneemt naarmate x toeneemt. Als maat voor een dergelijke correlatie, kan men een analogon van de covariantie definiëren; wij zullen hier niet dieper op ingaan en in de volgende paragrafen alleen de correlatie tussen twee stochastische variabelen beschouwen. De daarin behandelde theorie kan echter zonder grote moeilijkheden ook opgesteld worden voor de correlatie tussen een stochastische en een niet stochastische variabele

4 2. De correlatiecoëfficiënt.

De covariantie van $\underline{x}$ en y is nog geen geschikte maat voor hun correlatie. Indien m.n.b.v. door een schaalverandering de de spreidingen van $\underline{x}$ en y vergroot, zonder hun simultane verdeling eventueel te wijzigen, zal de covariantie naar evenredigheid toenemen. Men kan dit bezwaar ondervangen door de covariantie door de spreidingen van $\underline{x}$ en y te delen; men verkrijgt dan de zogenaamde correlatiecoëfficiënt $\rho(\underline{x}, y)$, die bij een schaal-wijziging niet verandert. Als voldoende bekend is op welke variabelen de correlatiecoëfficiënt betrekking heeft, gebruikt men ook wel het symbool ρ zonder meer.
Er geldt dus:

$$(4.2.1) \qquad \rho = \rho(x, y) = \frac{\mathcal{E}\, \tilde{x}\, \tilde{y}}{\sigma(x)\,\sigma(y)} .$$

Deze maat is nul, als $\mathcal{E}\, \tilde{x}\, \tilde{y} = 0$ is (mits $\sigma(\underline{x})$ en $\sigma(y) > 0$ zijn). Zij is minstens -1 en hoogstens $+1$. Immers: er geldt:

$$0 \leqq \mathcal{E}\left(\frac{\tilde{x}}{\sigma(x)} + \frac{\tilde{y}}{\sigma(y)}\right)^2 = \frac{\mathcal{E}\,\tilde{x}^2}{\sigma^2(x)} + \frac{\mathcal{E}\,\tilde{y}^2}{\sigma^2(y)} + \frac{2\,\mathcal{E}\,\tilde{x}\,\tilde{y}}{\sigma(x)\,\sigma(y)} = 1 + 1 + 2\rho = 2 + 2\rho,$$

dus: $\rho \geqq -1$.

Bovendien is

$$0 \leqq \mathcal{E}\left(\frac{\tilde{x}}{\sigma(x)} - \frac{\tilde{y}}{\sigma(y)}\right)^2 = 2 - 2\rho, \qquad \text{dus} \qquad \rho \leqq +1 .$$

De correlatiecoëfficiënt is gelijk aan -1, als $\mathcal{E}\left(\frac{\tilde{x}}{\sigma(x)} + \frac{\tilde{y}}{\sigma(y)}\right)^2 = 0$ is, dit is alleen mogelijk, als geldt:

$$(4.2.2) \qquad \frac{\tilde{x}}{\sigma(x)} + \frac{\tilde{y}}{\sigma(y)} = 0 .$$

Evenzo is zij gelijk aan $+1$ als:

$$(4.2.3) \qquad \frac{\tilde{x}}{\sigma(x)} - \frac{\tilde{y}}{\sigma(y)} = 0 .$$

Als men $\tilde{x}$ uitzet tegen $\tilde{y}$, stellen zowel (4.2.2) als (4.2.3) vergelijkingen voor van rechte lijnen, die door de oorsprong gaan. Indien men x uitzet tegen y zijn het rechte lijnen door het punt $(\mathcal{E}_x, \mathcal{E}_y)$; (4.2.2) stelt een dalende lijn voor, d.w.z. een lijn, waarbij de waarden van y afnemen naarmate x toeneemt, (4.2.3) stelt een stijgende lijn voor.

Wij zien dus, dat ρ in absolute waarde maximaal gelijk is aan 1, en dat deze waarde alleen bereikt wordt, in gevallen dat de verdeling geheel op een rechte lijn geconcentreerd is. Omgekeerd kan men bewijzen, dat $|\rho|$ steeds gelijk is aan 1, als de

verdeling geconcentreerd is op een willekeurige lijn die niet evenwijdig is aan één van de coördinaat-assen. In dat geval bestaat namelijk tussen de stochastische variabelen $\underline{x}$ en y het volgende verband:

$$y = \alpha \underline{x} + \beta ,$$

dus

$$\mathcal{E}\, y = \alpha\, \mathcal{E}\, \underline{x} + \beta$$

$$\tilde{y} = \alpha\, \tilde{\underline{x}} .$$

$$\mathcal{E}\, \tilde{y}^2 = \alpha\, \mathcal{E}\, \tilde{\underline{x}}^2$$

$$\mathcal{E}\, \tilde{\underline{x}}\, \tilde{y} = \mathcal{E}\, \alpha\, \tilde{\underline{x}}^2 = \alpha\, \mathcal{E}\, \tilde{\underline{x}}^2 .$$

Hieruit volgt:

$$\rho = \frac{\mathcal{E}\, \tilde{\underline{x}}\, \tilde{y}}{\sqrt{\mathcal{E}\, \tilde{\underline{x}}^2 . \mathcal{E}\, \tilde{y}^2}} = \frac{\alpha\, \mathcal{E}\, \tilde{\underline{x}}^2}{\sqrt{\alpha^2 . \mathcal{E}\, \tilde{\underline{x}}^2 . \mathcal{E}\, \tilde{\underline{x}}^2}} = \frac{\alpha}{|\alpha|}\, \frac{\mathcal{E}\, \tilde{\underline{x}}^2}{\mathcal{E}\, \tilde{\underline{x}}^2} = \begin{cases} +1 \text{ als } \alpha > 0. \\ -1 \text{ als } \alpha < 0. \end{cases}$$

De absolute waarde van de correlatiecoëfficiënt is groot, als de verdeling een sterke concentratie van de waarschijnlijkheid in de buurt van een rechte lijn vertoont.

Voor de berekening van ρ in de praktijk kunnen wij dikwijls met voordeel van de volgende formule gebruik maken:

$$\mathcal{E}\, \tilde{\underline{x}}\, \tilde{y} = \mathcal{E}\, \underline{x}\, y - \mathcal{E}\, \underline{x} . \mathcal{E}\, y . \tag{4.2.4}$$

De definitie van $\mathcal{E}\, \underline{x}\, y$ wordt voor een discrete verdeling in de notatie van het begin van par. 4.1

$$\mathcal{E}\, \underline{x}\, y = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij}\, x_i\, y_j . \tag{4.2.5}$$

De formule (4.2.4) is analoog aan de formule (2.11.2) voor de variantie, en wordt op overeenkomstige wijze bewezen.

4.3. Voorbeelden.

Voorbeeld 1. Wij berekenen $\rho(\underline{x}, y)$ bij de volgende tweedimensionale verdeling:

$$P\left[\underline{x} = -5 ,\ y = 3\right] = \frac{1}{4} .$$

$$P\left[\underline{x} = -2 ,\ y = 6\right] = \frac{1}{2} .$$

$$P\left[\underline{x} = 1 ,\ y = 3\right] = \frac{1}{4} .$$

Wij hebben deze verdeling weergegeven in fig. 4.1

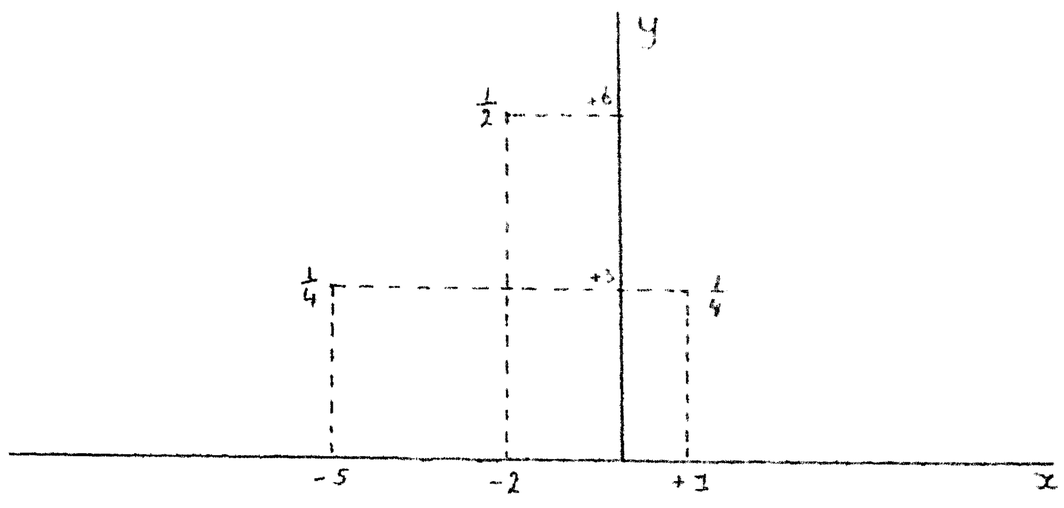


Fig. 4.1 (ad voorbeeld 1)

De verdeling is symmetrisch ten opzichte van de lijn $x = -2$. Daaruit volgt, dat $\mathcal{E}\,\underline{x} = -2$ is. Verder is:

$$\mathcal{E}\,y = \tfrac{1}{4}\cdot 3 + \tfrac{1}{2}\cdot 6 + \tfrac{1}{4}\cdot 3 = 4\tfrac{1}{2}.$$

Dus is:

$$\mathcal{E}\,\tilde{\underline{x}}\,\tilde{y} = \tfrac{1}{4}(-5+2)(3-4\tfrac{1}{2}) + \tfrac{1}{2}(-2+2)(6-4\tfrac{1}{2}) + \tfrac{1}{4}(1+2)(3-4\tfrac{1}{2}) =$$

$$= \tfrac{1}{4}(+4\tfrac{1}{2}) + \tfrac{1}{2}\cdot 0 + \tfrac{1}{4}(-4\tfrac{1}{2}) = 0.$$

Dit vinden wij ook als wij formule (4.2.5) toepassen:

$$\mathcal{E}\,\underline{x}\,y = \tfrac{1}{4}(-15) + \tfrac{1}{2}(-12) + \tfrac{1}{4}(+3) = -9.$$

$$\mathcal{E}\,\underline{x}\cdot\mathcal{E}\,y = -2\cdot 4\tfrac{1}{2} = -9.$$

Dus $$\mathcal{E}\,\tilde{\underline{x}}\,\tilde{y} = \mathcal{E}\,\underline{x}\,y - \mathcal{E}\,\underline{x}\;\mathcal{E}\,y = 0.$$

Aangezien $\rho = \dfrac{\mathcal{E}\,\tilde{\underline{x}}\,\tilde{y}}{\sigma(\underline{x})\cdot\sigma(y)}$ volgt hieruit, dat $\rho = 0$ is

Wij zien dus, dat $\underline{x}$ en y ongecorreleerd zijn, doch niet onafhankelijk; de punten vormen immers geen rechthoekig rooster. Uit het feit, dat de correlatiecoëfficiënt nul is, kan men dus niet concluderen, dat de beschouwde variabelen onafhankelijk zijn. Wij hebben daarop ook in par. 4.1 reeds de aandacht gevestigd.

Wij kunnen dit voorbeeld ook zien als een bijzonder geval van een twee-dimensionale verdeling, die symmetrisch is ten opzichte van een lijn evenwijdig aan één der coördinaatassen (hier de lijn $x = -2$). Indien dit het geval is, zal de covariantie van $\underline{x}$ en y steeds nul zijn. Laat de simultane verdeling

van $\underline{x}$ en y b.v symmetrisch zijn ten opzichte van de lijn $x = c$ (evenwijdig aan de y as). Het symmetriepunt van het punt (x, y) is nu $(c - x, y)$; de verdelingen van $(\underline{x}, y)$ en $(c - \underline{x}, y)$ zijn dus hetzelfde. Zij nu $\underline{u} = c - \underline{x}$, dan geldt dus:

$$\mathcal{E}\, \underline{u}\, y = \mathcal{E}\, \underline{x}\, y \,. \tag{4.3.1}$$

Tevens weten wij echter dat: $\mathcal{E}\, \underline{u} = c - \mathcal{E}\, \underline{x}$ is, dus is

$$\widetilde{\underline{u}} = \underline{u} - \mathcal{E}\, \underline{u} = c - \underline{x} - (c - \mathcal{E}\, \underline{x}) = -(\underline{x} - \mathcal{E}\, \underline{x}) = -\widetilde{\underline{x}} \,.$$

Hieruit volgt:

$$\mathcal{E}\, \widetilde{\underline{u}}\, \widetilde{y} = -\mathcal{E}\, \underline{x}\, y \,. \tag{4.3.2}$$

Uit de combinatie van (4.3.1) en (4.3.2) volgt dat $\mathcal{E}\, \widetilde{\underline{x}}\, \widetilde{y} = 0$ is. Wij hebben hier dus de volgende stelling bewezen:

Als de simultane verdeling van $\underline{x}$ en y symmetrisch is ten opzichte van een lijn evenwijdig aan een der coördinaatassen is de covariantie, dus ook de correlatie-coefficient van $\underline{x}$ en y gelijk aan nul.

Voorbeeld 2: Wij beschouwen het geval van fig 2.3 op pag. 36. (Trekking van twee loten zonder teruglegging uit een vaas met loten genummerd $1, \ldots, 5$).

Hier geldt:

$$\mathcal{E}\, \underline{x} = \mathcal{E}\, y = 3 \,,$$

$$\mathcal{E}\, \underline{x}^2 = \mathcal{E}\, y^2 = \tfrac{1}{5}\left(5^2 + 4^2 + 3^2 + 2^2 + 1^2\right) = 11 \,,$$

$$\mathcal{E}\, \widetilde{\underline{x}}^2 = \mathcal{E}\, \widetilde{y}^2 = \mathcal{E}\, \underline{x}^2 - (\mathcal{E}\, \underline{x})^2 = 11 - 3^2 = 2 \,.$$

De simultane verdeling van $\underline{x}$ en y kan gegeven worden in het volgende schema:

y \ x	1	2	3	4	5
5	5	10	15	20	×
4	4	8	12	×	20
3	3	6	×	12	15
2	2	×	6	8	10
1	×	2	3	4	5

De vakjes bevatten de waarden van xy of een kruis; de laatsten hebben een wh 0, de overige een wh $\frac{1}{20}$. Wij vinden zodoende:

$$\mathcal{E}\,\underline{x}\,\underline{y} = \tfrac{1}{20}\cdot 2\,(2+3+4+5+6+8+10+12+15+20) = 8\tfrac{1}{2},$$

$$\mathcal{E}\,\underline{\tilde{x}}\,\underline{\tilde{y}} = \mathcal{E}\,\underline{x}\,\underline{y} - \mathcal{E}\,\underline{x}\,.\,\mathcal{E}\,\underline{y} = 8\tfrac{1}{2} - 9 = -\tfrac{1}{2},$$

$$\rho(\underline{x}\ \underline{y}) = \frac{\mathcal{E}\,\underline{\tilde{x}}\,\underline{\tilde{y}}}{\sqrt{\mathcal{E}\,\underline{\tilde{x}}^2\ \mathcal{E}\,\underline{\tilde{y}}^2}} = \frac{-\frac{1}{2}}{2} = -\frac{1}{4}.$$

De correlatie-coëfficiënt is hier dus negatief. Dit was hier ook te verwachten. Door het feit, dat wij hier te doen hebben met een steekproef zonder teruglegging is voor y de waarde die $\underline{x}$ aanneemt niet beschikbaar; naarmate dit een hogere waarde is, zal y gemiddeld kleiner zijn.

Voorbeeld 3:

Wij nemen aan, dat er tussen $\underline{x}$ en y het volgende verband bestaat:

$$\underline{y} = \alpha\,\underline{x} + \beta + \underline{u}, \tag{4.3.3}$$

waarin α en β constanten zijn en $\underline{u}$ een stochastische variabele is, die niet van $\underline{x}$ afhangt. Dit zou zich b.v. kunnen voordoen in de volgende situatie: de variabele $\underline{x}$ stelt een stroomsterkte van een electrische stroom voor, die een bepaalde stochastische variatie vertoont; y stelt de wijzerstand van een galvanometer voor waarmee de stroom wordt gemeten; er wordt aangenomen, dat y lineair afhangt van $\underline{x}$, doch dat er bij de meting nog een zekere door u weergegeven waarnemingsfout optreedt, die niet afhangt van de stroomsterkte. Het verband tussen y, $\underline{x}$ en $\underline{u}$ wordt dan door een vergelijking van de gedaante (4.3.3) gegeven.

Wij gaan nu de variantie van y uitdrukken in die van $\underline{x}$ en $\underline{u}$. Wij vinden:

$$\mathcal{E}\,\underline{y} = \alpha\,\mathcal{E}\,\underline{x} + \beta + \mathcal{E}\,\underline{u}.$$

Dus:

$$\underline{\tilde{y}} = \underline{y} - \mathcal{E}\,\underline{y} = \alpha\,\underline{\tilde{x}} + \underline{\tilde{u}} \tag{4.3.4}$$

en

$$\sigma^2(y) = \mathcal{E}\,\underline{\tilde{y}}^2 = \alpha^2\,.\,\mathcal{E}\,\underline{\tilde{x}}^2 + \mathcal{E}\,\underline{\tilde{u}}^2 + 2\alpha\,\mathcal{E}\,\underline{\tilde{x}}\,\underline{\tilde{u}}.$$

De laatste term is echter 0 omdat $\underline{x}$ en $\underline{u}$ onafhankelijk zijn. Er geldt dus:

$$\sigma^2(y) = \alpha^2\,\sigma^2(\underline{x}) + \sigma^2(\underline{u}). \tag{4.3.5}$$

Dit resultaat was ook direct af te leiden uit Eigenschap 8 van par. 2.12.

Wij hebben nu de variantie van de metingen van de galvanometer gesplitst in een component $\alpha^2\sigma^2(\underline{x})$ die voortvloeit uit het lineaire verband met de stroomsterkte en een component $\sigma^2(\underline{u})$ afkomstig van de waarnemingsfout.

Uit (4 3 4) kunnen wij eveneens afleiden:

$$\mathcal{E}\,\underline{\tilde{x}}\,\underline{\tilde{y}} = \mathcal{E}\,\underline{\tilde{x}}(\alpha\,\underline{\tilde{x}} + \underline{\tilde{u}}) =$$

$$= \alpha\,\mathcal{E}\,\underline{\tilde{x}}^2 + \mathcal{E}\,\underline{\tilde{x}}\,\underline{\tilde{u}} = \alpha\,\mathcal{E}\,\underline{\tilde{x}}^2.$$

Dus:

$$\rho(\underline{x}\,y) = \frac{\mathcal{E}\,\underline{\tilde{x}}\,\tilde{y}}{\sigma(\underline{x})\,\sigma(y)} = \alpha\,\frac{\sigma(\underline{x})}{\sigma(y)},$$

of:

$$\alpha = \rho\,\frac{\sigma(y)}{\sigma(\underline{x})}, \tag{4.3.6}$$

en

$$\alpha^2\,\sigma^2(\underline{x}) = \rho^2\,\sigma^2(y).$$

In verband met (4.3.5) geldt dan:

$$\sigma^2(\underline{u}) = \sigma^2(y) - \alpha^2\,\sigma^2(\underline{x}) = (1-\rho^2)\,\sigma^2(y),$$

zodat tenslotte:

$$\frac{\alpha^2\sigma^2(\underline{x})}{\sigma^2(y)} = \rho^2 \qquad \frac{\sigma^2(\underline{u})}{\sigma^2(y)} = 1-\rho^2.$$

Wij zien dus, dat een fractie ρ^2 van de variantie in de metingen te wijten is aan het verband met de stroomsterkte en een fractie $1-\rho^2$ aan de variantie van de meetfout; dit voorbeeld illustreert dus het verband dat er bestaat tussen de correlatiecoëfficiënt en de mate waarin de simultane verdeling van $\underline{x}$ en y om een rechte lijn geconcentreerd is.

De studie van een verband van de vorm (4.3.3) en van meer gecompliceerde relaties van dezelfde aard vormt een belangrijk onderdeel van de wiskundige statistiek, welke bekend staat als de regressierekening. Daarin worden o.a. methoden gegeven om de coëfficiënten α en β, de variantie van $\underline{u}$ e.d. te schatten, als waarnemingen van $\underline{x}$ en de corresponderende waarnemingen van y gegeven zijn. Wij zullen op de regressierekening niet dieper ingaan. Uit het voorafgaande kan men wel vermoeden, dat de correlatiecoëfficiënt een belangrijke rol zal spelen; tussen de correlatie- en de regressierekening bestaat dan ook een nauw verband.

Opgave:

4.3.a. Bereken de correlatiecoëfficiënt van $\underline{x}$ en $\underline{y}$ bij tweedimensionale verdelingen, waarbij de whn van de drie punten A, B en C met hieronder gegeven coördinaten $\frac{1}{3}$ zijn:

1) A(-1,0) B(0,1) C(1,2)
2) A(-1,0) B(0,-2) C(1,4)
3) A(-1,0) B(0,0) C(0,1)
4) A(-1,0) B(0,1) C(1,0)
5) A(-1,0) B(0,0) C(0,1).

4.4. Passing in de machinebouw.

Wij geven nu een uitvoeriger voorbeeld, ontleend aan een artikel van J. SITTIG in Statistica.[11)]

Wij beschouwen twee onderdelen A en B van een machine, die in serie worden vervaardigd. A bevat een as, met voorgeschreven diameter α, die moet passen in een boring B met voorgeschreven diameter β. Daardoor is ook een voorgeschreven speling $\delta = \beta - \alpha$ bepaald.

In de practijk zullen de diameters van de onderdelen, die vervaardigd worden niet exact gelijk zijn aan α of β, doch een zekere variatie vertonen. Wij beschouwen daarom de diameter van A als een stochastische grootheid $\underline{a}$, de diameter van B als een stochastische grootheid $\underline{b}$. Wij onderstellen, dat de onderdelen gemiddeld zuiver gedraaid worden, d.w.z. $\mathcal{E}\,\underline{a} = \alpha$, $\mathcal{E}\,\underline{b} = \beta$. In dat geval kan men de spreidingen $\sigma^2(\underline{a})$ en $\sigma^2(\underline{b})$ als maten beschouwen voor de nauwkeurigheid van de machines, die de onderdelen vervaardigen. Deze nauwkeurigheid zal groter zijn, naarmate de spreiding kleiner is (vgl. par 2.11).

Bij de constructie van de machine, waarvan A en B onderdelen zijn, zal het ons niet in de eerste plaats interesseren of deze onderdelen zelf nauwkeurig vervaardigd worden, maar wel of de voorgeschreven speling δ behoorlijk wordt benaderd. Bij een willekeurig paar onderdelen A en B, zal een speling optreden, die beschouwd kan worden als een waarneming van een stochastische grootheid $\underline{d} = \underline{b} - \underline{a}$. De spreiding $\sigma^2(\underline{d})$ van $\underline{d}$ zal men dan kunnen beschouwen als een nauwkeurigheidsmaat voor de speling; immers er geldt steeds: $\mathcal{E}\,\underline{d} = \mathcal{E}\,\underline{b} - \mathcal{E}\,\underline{a} = \delta$ (zie formule (2.8.3)). De nauwkeurigheid van speling zullen wij "passing" noemen, de passing is beter naarmate $\sigma(\underline{d})$ kleiner is.

---

11) J. SITTIG, Over toleranties en passingen in de machinebouw, Statistica 1 (1946) p. 11-27.

Voor $\sigma(\underline{\alpha})$ geldt de volgende formule:

$$(4.4.1) \qquad \sigma^2(\underline{\alpha}) = \mathcal{E}(\tilde{\underline{a}} - \tilde{\underline{b}})^2 = \mathcal{E}\tilde{\underline{a}}^2 + \mathcal{E}\tilde{\underline{b}}^2 - 2\mathcal{E}\tilde{\underline{a}}\tilde{\underline{b}} = \sigma^2(\underline{a}) + \sigma^2(\underline{b}) - 2\rho\sigma(\underline{a})\sigma(\underline{b}),$$

waarin ρ de correlatiecoëfficiënt van $\underline{a}$ en $\underline{b}$ voorstelt.

In (4.4.1) zijn $\sigma^2(\underline{a})$ en $\sigma^2(\underline{b})$ grootheden, die bepaald worden door de nauwkeurigheid, waarmee A en B afzonderlijk vervaardigd worden, ρ is een maat voor de correlatie die tussen $\underline{a}$ en $\underline{b}$ aanwezig is. Deze correlatie hangt af van de wijze waarop onderdelen A en B gecombineerd worden. Zoals wij gezien hebben in par. 4.2 is $\rho = 0$ als $\underline{a}$ en $\underline{b}$ onafhankelijk zijn. Indien men steeds blindelings een A trekt uit een grote partij van deze onderdelen en eveneens blindelings een B uit een partij onderdelen B en deze combineert, dan mag men onderstellen, dat $\underline{a}$ en $\underline{b}$ onderling onafhankelijk zijn en zal dus gelden:

$$(4.4.2) \qquad \sigma^2(\underline{\alpha}) = \sigma^2(\underline{a}) + \sigma^2(\underline{b}).$$

Het blijkt uit (4.4.1) dat men de $\sigma(\underline{\alpha})$ kan verkleinen, dus de passing kan verbeteren door ρ positief te maken. Dit zal het geval zijn, indien $\underline{b}$ in het algemeen grotere waarden aanneemt naarmate de door $\underline{a}$ aangenomen waarde groter is, dus indien men aan grotere assen in het algemeen grotere boringen toevoegt. Dit kan b.v. bereikt worden door de onderdelen A en B te sorteren naar hun respectievelijke diameters alvorens ze te combineren. Het effect van een dergelijke sortering is verrassend. J. SITTIG heeft dit toegelicht aan de hand van een voorbeeld. Hij onderstelt daartoe, dat $\sigma(\underline{a}) = \sigma(\underline{b})$ is. Dan geldt:

$$(4.4.3) \qquad \sigma^2(\underline{\alpha}) = 2\sigma^2(1 - \rho^2)$$

Verder onderstelt SITTIG, dat men de onderdelen sorteert in twee klassen met als criterium: diameter groter of kleiner dan het gemiddelde. Men combineert dus steeds assen met diameter groter resp. kleiner dan α met boringen met diameter groter resp. kleiner dan β, doch men gaat overigens aselect te werk. Indien nu $\underline{a}$ en $\underline{b}$ beide normaal verdeeld zijn, wordt de correlatiecoëfficiënt van $\underline{a}$ en $\underline{b}$, zoals SITTIG bewijst, gelijk aan $\frac{2}{\pi} \approx 0{,}64$, waarna uit (4.4.3) volgt:

$$\sigma(\underline{\alpha}) = 0{,}855\,\sigma,$$

in plaats van $\sigma\sqrt{2} = 1{,}41\,\sigma$ bij niet sorteren.

Wij zien dus, dat wij door deze eenvoudige sortering de passing sterk kunnen verbeteren. De spreiding van de speling is immers al kleiner geworden dan de spreiding in de diameters van de assen en de boringen afzonderlijk. Omgekeerd kunnen wij, indien te voren reeds een voldoende passing bereikt werd, door invoering van deze sortering eventueel kosten besparen door een geringere precisie te eisen bij het vervaardigen van de onderdelen A en B.

4.5. Schatting van de correlatiecoëfficiënt.

In de meeste gevallen, dat men met correlatieverschijnselen te doen heeft, zal men de corresponderende tweedimensionale verdeling niet kennen; zoals gebruikelijk zal men dan trachten iets over de correlatie te weten te komen uit steekproefmateriaal. Dit is mogelijk als men beschikt over een reeks waarnemingen (x_i, y_i) $(i=1, \ldots, n)$ van de tweedimensionale verdeling van x en y. Wij kunnen dan de schatting r van de correlatiecoëfficiënt, die gewoonlijk gebruikt wordt, vinden uit de volgende formule:

$$(4.5.1) \qquad r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{j=1}^{n} (y_j - \bar{y})^2}} \qquad \left(\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \; \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j\right)$$

De grootheid r kan veelal het gemakkelijkste als volgt berekend worden: Men bepaalt:

$$(4.5.2) \qquad \begin{cases} S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \cdot \sum_{j=1}^{n} y_j \\ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i\right)^2 \\ S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} y_i\right)^2 \end{cases}$$

en berekent daarna

$$(4.5.3) \qquad r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \; .$$

Wij geven hieronder een eenvoudig rekenvoorbeeld (zie tabel 4.I). In de eerste kolom vindt men de waarnemingen van x, in de tweede op gelijke hoogte de bijbehorende waarnemingen van y; de overige kolommen bevatten de berekende waarden van x_i^2, $x_i y_i$ en y_i^2 die voor de bepaling van de sommen voorkomende in S_{xx}, S_{xy} en S_{yy} nodig zijn.

Ook de grootheid r kan alleen tussen -1 en +1 variëren. Indien er tussen de waarnemingen van $\underline{x}$ en die van y een exact lineair verband bestaat, zal $|r| = 1$ zijn. In het algemeen zal r niet $= 0$ zijn, als $\underline{x}$ en y onafhankelijk zijn; r is namelijk een statistische grootheid, die geheel bepaald wordt door de steekproefwaarden.

Tabel 4.I

Voorbeeld van de berekening van r volgens (4.5.3)

x	y	x^2	xy	y^2
5	3	25	15	9
7	-2	49	-14	4
6	0	36	0	0
3	5	9	15	25
4	-1	16	- 4	1
2	-2	4	- 4	4
-1	4	1	- 4	16
$\Sigma x_i = 26$	$\Sigma y_i = 7$	$\Sigma x_i^2 = 140$	$\Sigma x_i y_i = 4$	$\Sigma y_i^2 = 59$

$$S_{xy} = \Sigma x_i y_i - \frac{1}{n} \Sigma x_i \Sigma y_i = 4 - \frac{1}{7} \cdot 26 \cdot 7 = -22$$

$$S_{xx} = \Sigma x_i^2 - \frac{1}{n} (\Sigma x_i)^2 = 140 - \frac{1}{7} \cdot 26^2 = \frac{304}{7}$$

$$S_{yy} = \Sigma y_i^2 - \frac{1}{n} (\Sigma y_i)^2 = 59 - \frac{1}{7} \cdot 7^2 = 52$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-22}{\sqrt{\frac{304 \cdot 52}{7}}} = -0,463.$$

Beschouwen wij de verzameling van alle mogelijke steekproeven van n waarnemingen behorende bij dezelfde tweedimensionale verdeling, dan zal $\underline{r}$ over deze vz. een waarschijnlijkheidsverdeling hebben. Deze wh-verdeling is van dien aard, dat $\underline{r}$ zeer waarschijnlijk dicht bij ρ zal komen te liggen, als n groot is. Hoe groot n moet zijn, opdat $\underline{r}$ met een bepaalde waarschijnlijkheid hoogstens een bepaald bedrag van ρ zal verschillen, hangt echter af van het karakter van de tweedimensionale verdeling van $\underline{x}$ en y.

De wh-verdeling van $\underline{r}$ is voor enkele speciale gevallen berekend o.a. als $\underline{x}$ en $\underline{y}$ simultaan een zogenaamde tweedimensionale normale verdeling hebben. Indien men te voren weet, dat $\underline{x}$ en $\underline{y}$ exact of althans bij benadering tweedimensionaal normaal verdeeld zijn, kan men als $\underline{r}$ berekend is bepaalde wh-uitspraken omtrent ρ doen (b.v.: een onbetrouwbaarheidsinterval voor ρ afleiden), en daaruit het een en ander concluderen omtrent de correlatie tussen $\underline{x}$ en $\underline{y}$.

In de practijk is het vooral van belang om de hypothese $\rho = 0$ te kunnen toetsen. Indien deze hypothese verworpen wordt, kan men concluderen, dat $\underline{x}$ en $\underline{y}$ niet onafhankelijk zijn en dat er een correlatie tussen deze grootheden bestaat.

Wij zullen hieronder een correlatiemaat bespreken, waarmee men de hypothese, dat $\underline{x}$ en $\underline{y}$ onafhankelijk zijn kan toetsen, zonder te onderstellen, dat $\underline{x}$ en $\underline{y}$ simultaan een tweedimensionale verdeling hebben.

4.6. Rangcorrelatiecoëfficiënt van KENDALL.

Stel, dat een aantal jongens gerangschikt is volgens hun bekwaamheid in wiskunde en muziek. De jongens duiden wij aan met de letters A, B, ..., J, de rangschikking geven wij weer met rangnummers, b.v. naar opklimmende bekwaamheid; in gevallen waarin dat voorkomt, zullen wij aan jongens met gelijke bekwaamheid het gemiddelde toekennen van de rangnummers die ze zouden krijgen als ze in bekwaamheid verschilden. Ter vereenvoudiging van het betoog zullen wij de mogelijkheid van gelijke rangnummers buiten beschouwing laten.

Laat het resultaat van de rangschikking het volgende zijn:[12)]

Jongens:	A	B	C	D	E	F	G	H	I	J
Wiskunde	7	4	3	10	6	2	9	8	1	5
Muziek	5	7	3	10	1	9	6	2	8	4.

Wij willen weten of er uit dit materiaal iets blijkt van verwantschap tussen aanleg voor wiskunde en aanleg voor muziek.

Daartoe beschouwen wij alle paren jongens die wij uit de verzameling kunnen vormen (in dit geval $\frac{1}{2}.10.9 = 45$ paren). Wij kennen nu aan een paar het getal +1 of -1 toe al naar gelang de rangnummers van dat paar voor wiskunde en muziek gelijke of tegengestelde volgorde hebben. Zo verkrijgt in ons voorbeeld het paar AB het getal -1, omdat $7 > 4$ doch $5 < 7$ is, evenzo krijgt het paar AC het getal +1 , FH -1, CD +1, etc.

---

12) M.G. KENDALL, Rankcorrelation method, London (1948) p.3.

(Zo het paar in tenminste een van beide rangschikkingen gelijke rangnummers heeft, kent men een 0 toe). De algebraïsche som van de aan de paren toegekende getallen stellen wij voor door S. Wij kunnen deze grootheid als volgt berekenen.

Wij brengen door permutatie van de jongens een van beide rijen in de "natuurlijke" volgorde. (Door dergelijke permutaties worden uiteraard de getallen die wij de paren moeten toekennen niet veranderd.) Wij vinden dan:

Jongens:	I	F	C	B	J	E	A	H	G	D
Wiskunde	1	2	3	4	5	6	7	8	9	10
Muziek	8	9	3	7	4	1	5	2	6	10

Wij tellen dan in de laatste rij hoeveel malen achter ieder cijfer een hoger cijfer voorkomt. Dit is het totaal aantal paren waaraan een +1 moet worden toegekend. (Positief correlerende paren.) Hun aantal wordt:

$$P = 2 + 1 + 5 + 1 + 3 + 4 + 2 + 2 + 1 = 21.$$

De overige paren zijn negatief correlerend; wij vinden deze door te tellen hoeveeel malen achter ieder cijfer in de tweede rij een lager cijfer voorkomt. Hun aantal wordt:

$$Q = 7 + 7 + 2 + 5 + 2 + 0 + 1 + 0 + 0 = 24.$$

Inderdaad is $P + Q = 21 + 24 = 45$ het totaal aantal paren en

$$S = P - Q = -3.$$

Het is deze S die wij als maat voor de correlatie nader wensen te beschouwen. Allereerst kunnen wij opmerken, dat S in ons voorbeeld maximaal 45 had kunnen zijn, nl. indien alle paren een +1 hadden gehad. Dit is het geval als de rijen rangnummers volkomen identiek zijn, b.v.:

Jongens:	A	B	C	D	E	F	G	H	I	J
Wiskunde	7	4	3	10	6	2	9	8	1	5
Muziek	7	4	3	10	6	2	9	8	1	5.

Als alle paren een -1 hadden gehad, was S = - 45 geweest; dit is het geval bij volkomen tegengestelde rangschikkingen, b.v.:

Jongens:	A	B	C	D	E	F	G	H	I	J
Wiskunde	7	4	3	10	6	2	9	8	1	5
Muziek	4	7	8	1	5	9	2	3	10	6.

In het algemeen zal bij n jongens S maximaal $\binom{n}{2} = \frac{1}{2}n(n-1)$ en minimaal $-\frac{1}{2}n(n-1)$ zijn. Desgewenst kunnen wij hier een coëfficiënt definiëren, die tussen -1 en +1 kan variëren, nl.:

$$\tau = \frac{2S}{n(n-1)} \qquad (4.6.1)$$

In ons voorbeeld is $\tau = -\frac{3}{45} = -\frac{1}{15} = -0,067$.
Deze grootheid heet de rangcorrelatiecoëfficiënt van KENDALL. Het is een stochastische grootheid, evenals $\varkappa$, want zij is gebaseerd op de steekproefwaarden. Uit de wijze waarop zij berekend wordt blijkt, dat zij een maat is voor de correlatie, zoals wij dat begrip in 4.1 hebben ingevoerd; een hoge absolute waarde van τ zal overeenstemmen met een sterke correlatie, en het teken van τ zal corresponderen met het teken van de correlatie. Tevens blijkt hieruit, dat τ alléén afhangt van de grootteorde der waarnemingen, niet van hun waarden zelf. Een voordeel hiervan is, dat τ berekend kan worden in gevallen, waarin x en y geen meetbare grootheden zijn, doch waarin de waarnemingen van deze grootheden wel naar opklimmende of afdalende grootte gerangschikt kunnen worden. Dat was in ons voorbeeld feitelijk het geval. Om daarbij te kunnen berekenen, zal men een of andere maat voor bekwaamheid moeten bedenken; in de keuze van deze maat zit echter een element van willekeur.

Omdat τ slechts afhangt van de grootteorde der waarnemingen, is deze coëfficiënt in meer gevallen gelijk aan +1 of -1 dan $\varkappa$; dit is bij $\varkappa$ alleen het geval, als de waargenomen punten (x_i, y_i) op een rechte lijn liggen. Bij τ is het voldoende, dat de grootteorde der x_i dezelfde is als die der y_i $(\tau = +1)$ of daaraan juist tegengesteld is $(\tau = -1)$.

De rangcorrelatiecoëfficiënt van KENDALL kan ook berekend worden in gevallen, dat wij wel met meetbare grootheden te doen hebben. Dit geschiedt dan door de waarnemingen van b.v. x naar opklimmende grootte te rangschikken en de overeenkomstige waarnemingen van y eronder te plaatsen. Men kan verder volkomen te werk gaan als boven beschreven is; om vergissingen te voorkomen, kan men de gegevens ook eerst door rangnummers vervangen (vgl het voorbeeld in par. 4.9).

In par. 4.1 noemden wij reeds de mogelijkheid om correlatie tussen een stochastische en een niet-stochastische grootheid te beschouwen. Een groot voordeel van de rangcorrelatiecoëfficiënt is, dat de daarop gebaseerde toetsingsmethode toepasbaar is

zowel in het geval,dat men met een stochastische en een niet stochastische grootheid te doen heeft. Wij zullen van beide gevallen een voorbeeld geven (zie par. 4.8 resp. 4.9).

Opgave:

4.6.a. De aantallen aanwezige deelnemers aan de cursus Toegepaste Statistiek te Den Haag in 1953-1954 waren op de opeenvolgende cursusavonden:

25, 24, 20, 19, 18, 15, 17, 17, 15, 13, 12, 10, 12.

Bereken S en τ voor de correlatie tussen deze aantallen en de tijd.

Aanwijzing:

Men moet de aantallen correleren met de rij 1,2,...,13, de tijdsvolgorde der waarnemingen. Omdat niet alle aantallen verschillend zijn, moet aan enige paren een nul worden toegekend.

4.7. Toets gebaseerd op de rangcorrelatiecoëfficiënt τ.

Met de rangcorrelatiecoëfficiënt van KENDALL kunnen wij bij het voorbeeld van par. 4.6 de hypothese H_o toetsen, dat de rangschikking der jongens volgens het criterium muziek onafhankelijk is van de rangschikking volgens het criterium wiskunde. Als H_o juist is, zijn alle mogelijke volgorden van een van beide rangnummerrijen even waarschijnlijk, en onafhankelijk van de volgorde in de andere rij. Dit geldt ook indien één van beide rangschikkingen gebaseerd is op waarden van een niet-stochastische variabele.

Door de interpretatie, die wij aan onze hypothese H_o gegeven hebben zijn wij in staat de verdeling van S of τ onder deze hypothese te berekenen.

Laten wij dit toelichten aan een zeer eenvoudig voorbeeld, waar de rijen bestaan uit 3 rangnummers. Omdat wij de volgorde in de eerste rij onafhankelijk onderstellen van die in de tweede rij, kunnen wij de volgorde in de eerste rij constant nemen. Wij hebben dan de volgende mogelijkheden:

$$\left.\begin{matrix}1\ 2\ 3\\1\ 2\ 3\end{matrix}\right\} \quad S = +3 \qquad \left.\begin{matrix}1\ 2\ 3\\2\ 3\ 1\end{matrix}\right\} \quad S = -1$$

$$\left.\begin{matrix}1\ 2\ 3\\1\ 3\ 2\end{matrix}\right\} \quad S = +1 \qquad \left.\begin{matrix}1\ 2\ 3\\3\ 1\ 2\end{matrix}\right\} \quad S = -1$$

$$\left.\begin{matrix}1\ 2\ 3\\2\ 1\ 3\end{matrix}\right\} \quad S = +1 \qquad \left.\begin{matrix}1\ 2\ 3\\3\ 2\ 1\end{matrix}\right\} \quad S = -3\,.$$

Er zijn in totaal 6 mogelijkheden, die alle onder de hypothese H_o even waarschijnlijk zijn. Wij vinden dus:

$$P[\underline{S} = 3] = \frac{1}{6} \qquad P[\underline{S} = 1] = \frac{1}{3} \qquad P[\underline{S} = -1] = \frac{1}{3} \qquad P[\underline{S} = -3] = \frac{1}{6}$$

waarmee overeenkomt:

$$P[\underline{\tau} = 1] = \frac{1}{6} \qquad P[\underline{\tau} = \frac{1}{3}] = \frac{1}{3} \qquad P[\underline{\tau} = -\frac{1}{3}] = \frac{1}{3} \qquad P[\underline{\tau} = -1] = \frac{1}{6}.$$

Volgens ditzelfde principe kunnen wij ook bij grotere waarden van n de wh-verdeling van $\underline{S}$ of $\underline{\tau}$ opstellen.

Bij deze verdelingen zijn de whn van waarden van $\underline{S}$ kleiner naarmate zij verder van 0 verwijderd zijn; indien de beschouwde grootheden gecorreleerd zijn, zullen de ver van 0 verwijderde waarden van S vaker optreden, en wel de positieve waarden bij positieve en de negatieve waarden bij negatieve correlatie. Wij verwerpen daarom de hypothese H_o, zodra de kans op het optreden van de gevonden of een verder van S = 0 verwijderde waarde, (dus de tweezijdige overschrijdingskans) niet groter is dan de gekozen onbetrouwbaarheidsdrempel α; wij hebben in tabel 4.II de zogenaamde kritieke waarden van S gegeven, bij 4 verschillende waarden van α en voor $n = 4, 5, \ldots, 20$. Deze tabel is ontleend aan een publicatie in het tijdschrift Statistica.[13] De absolute waarde van S moet minstens even groot zijn als daar aangegeven is, om tot verwerping van H_o te kunnen komen.

In ons voorbeeld van par. 4.6 vonden wij bij $n = 10$ $S = -3$. De rechtse kritieke waarde van de tweezijdige toets is volgens tabel 4.II bij de grootste onbetrouwbaarheidsdrempel $\alpha = 0,10$ nog gelijk aan 21. De verdeling van $\underline{S}$ is symmetrisch ten opzichte van 0, dus de linker kritieke waarde is -21; de gevonden waarde ligt tussen deze kritieke waarden in en H_o kan niet verworpen worden.

---

13) L. Kaarsemaker en A. van Wijngaarden, Tables for use in rankcorrelation, Statistica 7 (1953) p. 41-54.

Tabel 4.II

Rechter kritieke waarden van S_α van $\underline{S}$ voor de tweezijdige rangcorrelatietoets.

n	$\alpha = 0{,}01$	$\alpha = 0{,}02$	$\alpha = 0{,}05$	$\alpha = 0{,}10$
4	-	-	-	6
5	-	10	10	8
6	15	13	13	11
7	19	17	15	13
8	22	20	18	16
9	26	24	20	18
10	29	27	23	21
11	33	31	27	23
12	38	36	30	26
13	44	40	34	28
14	47	43	37	33
15	53	49	41	35
16	58	52	46	38
17	64	58	50	42
18	69	63	53	45
19	75	67	57	49
20	80	72	62	52

De tabel bevat de kleinste waarde S_α, die $\underline{S}$ kan aannemen, waarvoor $2P\left[\underline{S} \geqq S_\alpha\right] \leqq \alpha$.

Indien wij bij dezelfde waarde van n, $S = +25$ vinden, kunnen wij H_o bij $\alpha = 0{,}10$ of $0{,}05$ wel verwerpen; men kan dan besluiten, dat de onderzochte grootheden positief gecorreleerd zijn. Bij een waarde van $S = -28$ kan men H_o verwerpen bij $\alpha = 0{,}10$, $0{,}05$ en $0{,}02$; men besluit dan tot negatieve correlatie.

Wij hebben hier de tweezijdige vorm van de toets gegeven. Men kan desgewenst ook eenzijdig toetsen, analoog aan hetgeen gedaan is bij de tekentoets en de toets van Wilcoxon in de cursus Toegepaste Statistiek I. Tabel 4.II geeft ook kritieke waarden van de rechts eenzijdige toets, mits men de opgegeven waarden van α halveert. De kritieke waarden van de links eenzijdige toets zijn het tegengestelde van die van de rechtseenzijdige toets.

Voor alle waarden van n zijn de verwachting $\mathcal{E}\,\underline{S}$ en de spreiding σ van $\underline{S}$ onder de hypothese H_0 bekend en wel is:

(4.7.1) $$\mathcal{E}\,\underline{S} = 0$$

en

(4.7.2) $$\sigma = \sqrt{\tfrac{1}{18}\,n(n-1)(2n+5)}.$$

Bij ons voorbeeld ($n = 10$) is dus:

$$\sigma = \sqrt{\tfrac{1}{18}\cdot 10.9.25} = \sqrt{125} = 11,18.$$

Strikt genomen geldt deze formule (evenals tabel 4.II en onze bewering, dat de verdeling van $\underline{S}$ onder H_0 symmetrisch is) alleen indien er geen gelijke waarnemingen in de steekproeven voorkomen, dus indien er geen paren zijn, waaraan bij de berekening van S een nul moet worden toegekend. De afwijkingen zijn echter van weinig betekenis zolang de groepen van gelijke waarnemingen relatief klein zijn (zoals b.v. in Opgave 4.9.a). Onder deze restrictie kan tevens worden aangetoond, dat $\underline{S}/\sigma$ voor grote waarden van n onder H_0 bij benadering een $N(0,1)$-verdeling heeft. Hiermee kan men de tweezijdige overschrijdingskans van een gevonden resultaat benaderen. Men berekent dan:

$$u = \frac{|S| - 1}{\sigma},$$

waarin $|S|$ de absolute waarde van S en de term -1 een continuïteitscorrectie is. Daarna bepaalt men met een tabel van de $N(0,1)$-verdeling de tweezijdige overschrijdingskans k van u. Men vindt zodoende bij ons voorbeeld ($S = -3$; $\sigma = 11,18$)

$$u = \frac{3-1}{11,18} = 0,179$$

dus volgens de tabel van de normale verdeling $k = 0,86$.

Men kan deze benaderingsmethode ook gebruiken om de toets toe te passen voor waarden van α en van $n \geqq 10$ die niet in tabel 4.I voorkomen. Men verwerpt dan H_0 zodra $k \leqq \alpha$. Indien men b.v. bij $n = 30$ vindt $S = 80$ en men toetst met $\alpha = 0,05$ tweezijdig, dan berekent men:

$$\sigma = \sqrt{\tfrac{1}{18}\cdot 30.29\cdot 62} = 54,74$$

$$u = \frac{79}{54,74} = 1,443.$$

De tweezijdige overschrijdingskans hiervan wordt volgens de tabel van de $N(0,1)$-verdeling $0,149 \geqq 0,05$. H_0 kan dus niet verworpen worden.

Opgave:

4.7.a. Pas de rangcorrelatie-toets toe op de resultaten van opgave 4.6.a (onbetrouwbaarheidsdrempel $\alpha = 0,01$). Geef de benadering voor de tweezijdige overschrijdingskans.

4.8. Daling van het geboortecijfer en fractie der jongensgeboorten.

In tabel 4.III is x het aantal geboorten per jaar in Engeland en Wales in de jaren 1910-1919 en y de verhouding van het aantal jongens-geboorten tot het totale aantal geboorten in dezelfde jaren.

De rangcorrelatiecoëfficiënt tussen x en y berekenen wij als volgt:

Tabel 4.III

Jaar	**x** = Aantal geb. in Engeland en Wales		Perunage jongensgeboorten	
	$10^{-3}.x$	Rangn.	y	Rangn.
1910	897	10	0,5098	5
1911	881	8	95	3
1912	873	6	99	6
1913	882	9	93	2
1914	879	7	87	1
1915	815	5	97	4
1916	786	4	119	9
1917	668	2	108	7
1918	663	1	117	8
1919	692	3	0,5145	10

Rangnummers:

naar x: 1 2 3 4 5 6 7 8 9 10

naar y: 8 7 10 9 4 6 1 3 2 5

P = 2 + 2 + 0 + 0 + 2 + 0 + 3 + 1 + 1 = 11

Q = 7 + 6 + 7 + 6 + 3 + 4 + 0 + 1 + 0 = 34

S=P - Q = - 5 - 4 - 7 - 6 - 1 - 4 + 3 + 1 = -23

Hierin zou men een aanwijzing kunnen zien voor een negatieve correlatie tussen het aantal geboorten en het perunage van de jongensgeboorten in Engeland en Wales. Het is echter zeer gevaarlijk zonder meer tot een verband tussen twee grootheden te besluiten op grond van een geconstateerde correlatie tussen twee cijferreeksen. Wij hebben hier kennelijk te doen met een geval van "indirecte correlatie" (Eng.: spurious correlation). Het ligt voor de hand de geconstateerde negatieve correlatie toe te schrijven aan:

1e Een sterk negatieve correlatie tussen de geboortecijfers en de jaartallen van de jaren waarop zij betrekking hebben. Dit is een in West Europa algemeen geconstateerd verschijnsel, dat hier nog versterkt wordt door de geboortedaling in de oorlogsjaren 1914 - 1918. ($S = -35$, H_o wordt dus reeds verworpen bij een onbetrouwbaarheidsdrempel 0,01)

2e Een zwak positieve correlatie tussen genoemde jaartallen en het perunage van de jongensgeboorten. Wij vinden $S = +21$. De benaderde tweezijdige overschrijdingskans hiervan is 0,074.

In deze correlatie manifesteert zich een stijging van het perunage in de oorlogsjaren, een eveneens bekend verschijnsel.

De hier optredende indirecte correlatie is dus te wijten aan het feit, dat het materiaal inhomogeen is ten aanzien van factoren, die op beide waargenomen grootheden van invloed zijn. Soms kunnen wij dergelijke indirecte correlaties ontmaskeren door het materiaal in meer homogene groepen te splitsen. In ons geval zouden wij kunnen splitsen in de jaren 1910 - 1914 en 1915 - 1919. Men vindt dan voor de eerste gropp $S = 0$ in de tweede groep $S = -2$; er is dus in de deelgroepen van correlatie niets overgebleven.

Opgave:

4.8.a. Verifieer de resultaten van de rangcorrelatietoetsen opgegeven onder 1e en 2e.

4.9. Combinatie van rangcorrelatietoetsen.

Wij bespreken de methode van het combineren van rangcorrelatietoetsen aan de hand van cijfers betreffende de "Bankers-deposits" bij de "Bank of England" op de Woensdagen van het jaar 1945 (zie tabel 4.IV).

Er zijn economische redenen om te veronderstellen, dat deze bedragen in de loop van iedere maand geleidelijk stijgen, om aan het einde snel af te nemen. Dit effect zullen wij niet gemakkelijk vinden door de rangcorrelatietoets toe te passen op de cijfers van een geheel jaar tezamen, omdat de rangcorrelatiecoëfficiënt niet gevoelig is voor periodiciteit. Wij hebben daarom de rangcorrelatiecoëfficiënten tussen data en deposito-bedragen maand voor maand afzonderlijk berekend (zie Tabel 4.IV).

Het blijkt, dat de S van Kendall in 10 van de 12 maanden positief is. Indien wij echter de hypothese H_0 toetsen, dat het depositobedrag onafhankelijk is van de Woensdag in de maand, waarop het wordt waargenomen, is de overschrijdingskans overal groter dan 0,05.

Wij kunnen toch tot een duidelijke conclusie komen, indien wij de toetsen combineren volgens het principe, dat in par. 3.5 behandeld is. Wij zouden daarbij als toetsingsgrootheid de som van de gevonden waarden van S kunnen nemen. In verband met het feit, dat de grootheden S niet alle uit hetzelfde aantal waarnemingen berekend zijn verdient het om theoretische, hier niet nader te vermelden, redenen de voorkeur de som der rangcorrelatiecoëfficiënten τ te kiezen; als τ_i de rangcorrelatiecoëfficiënt van KENDALL voor de i^e maand is, wordt onze toetsingsgrootheid

$$(4.9.1) \qquad \underline{T} = \sum_{i=1}^{12} \underline{\tau}_i .$$

Nu geldt onder de hypothese H_0 voor iedere i $\mathcal{E}\,\underline{\tau}_i = 0$, hieruit volgt:

$$(4.9.2) \qquad \mathcal{E}\,\underline{T} = 0 .$$

Verder is onder de hypothese H_0 als n_i het aantal waarnemingen in de i^e maand voorstelt

$$\sigma^2(\underline{S}_i) = \frac{1}{18}\, n_i (n_i - 1)(2n_i + 5),$$

dus:

$$(4.9.3) \qquad \sigma^2(\underline{\tau}_i) = \sigma^2\left(\frac{2\,\underline{S}_i}{n_i(n_i-1)}\right) = \frac{2}{9}\,\frac{2n_i + 5}{n_i(n_i-1)}$$

(Vgl formule 2.12.2 blz.46).

Tabel 4.IV

"Bankers deposits" bij de "Bank of England" op de Woensdagen van het jaar 1945 in millioenen Engelse Ponden.

Woens-dagen	Bedrag	Rangn.	S en $\tilde{\tau}$	Woens-dagen	Bedrag	Rangn.	S en $\tilde{\tau}$
3 Jan.	251	5		4 Juli	251	4	
10 "	222	4		11 "	219	2	
17 "	214	2	- 6	18 "	211	1	- 2
24 "	208	1	- 6/10	25 "	229	3	- 1/3
31 "	215	3					
7 Febr.	177	1		1 Aug.	221	3	
14 "	184	2		8 "	203	1	
21 "	197	3	+ 6	15 "	211	2	+ 6
28 "	208	4	+ 1	22 "	227	4	+ 6/10
				29 "	238	5	
7 Maart	204	2		5 Sept.	215	1	
14 "	191	1	+ 4	12 "	219	2	+ 6
21 "	218	3	+ 2/3	19 "	248	3	+ 1
28 "	219	4		26 "	279	4	
4 April	188	2		3 Oct.	233	3	
11 "	185	1		10 "	212	1	
18 "	191	3	+ 4	17 "	232	2	+ 6
25 "	230	4	+ 2/3	24 "	242	4	+ 6/10
				31 "	244	5	
2 Mei	192	3		7 Nov.	221	2	
9 "	176	1		14 "	225	3	
16 "	185	2	+ 4	21 "	218	1	+ 2
23 "	219	5	+ 4/10	28 "	250	4	+ 1/3
30 "	212	4					
6 Juni	181	1		5 Dec.	219	2	
13 "	201	2	+ 6	12 "	217	1	+ 4
20 "	211	3	+ 1	19 "	229	3	+ 2/3
27 "	262	4		26 "	274	4	

Voor $n_i = 5$ geldt dus:

$$\sigma^2(\tilde{\tau}_i) = \frac{2}{9} \cdot \frac{15}{5.4} = \frac{1}{6},$$

en voor $n_i = 4$

$$\sigma^2(\underline{\xi}_i) = \frac{2}{9} \cdot \frac{11}{4.3} = \frac{11}{54} .$$

Indien wij nu verder in onze getoetste hypothese opnemen, dat de bankers-deposits in de verschillende maanden onderling onafhankelijk zijn, dan geldt:

$$(4.9.4) \qquad \sigma^2(\underline{I}) = \sum_{i=1}^{12} \sigma^2(\underline{\xi}_i) .$$

Nu zijn er 4 maanden met $n_i = 5$, dus $\sigma^2(\underline{\xi}_i) = \frac{1}{6}$ en 8 maanden met $n_i = 4$, dus $\sigma^2(\underline{\xi}_i) = \frac{11}{54}$.
Wij vinden:

$$\sigma^2(\underline{I}) = 4 \cdot \frac{1}{6} + 8 \cdot 54 = 2\frac{8}{27} .$$

Dus is

$$\sigma(I) = \sqrt{2\frac{8}{27}} = \sqrt{2,2963} = 1,52 .$$

Onder de getoetste hypothese zal volgens par. 3.5 $I/\sigma(I)$ bij benadering een $N(0,1)$-verdeling hebben. Wij vonden: $T = 6$, dus $I/\sigma(I) = 3,44$. Hierbij behoort een tweezijdige overschrijdingskans $\alpha = 0,0006$; wij kunnen dus de hypothese verwerpen, dat de deposito's onafhankelijk zijn van de datum. Uit de gevonden waarde van T blijkt, dat er binnen de maanden een neiging is tot een stijgend verloop.

Inhoud Cursus Toegepaste Statistiek II (Rapport S 155)

Oplossingen der opgaven in rapport S 155, Cursus Toegepaste Statistiek II.

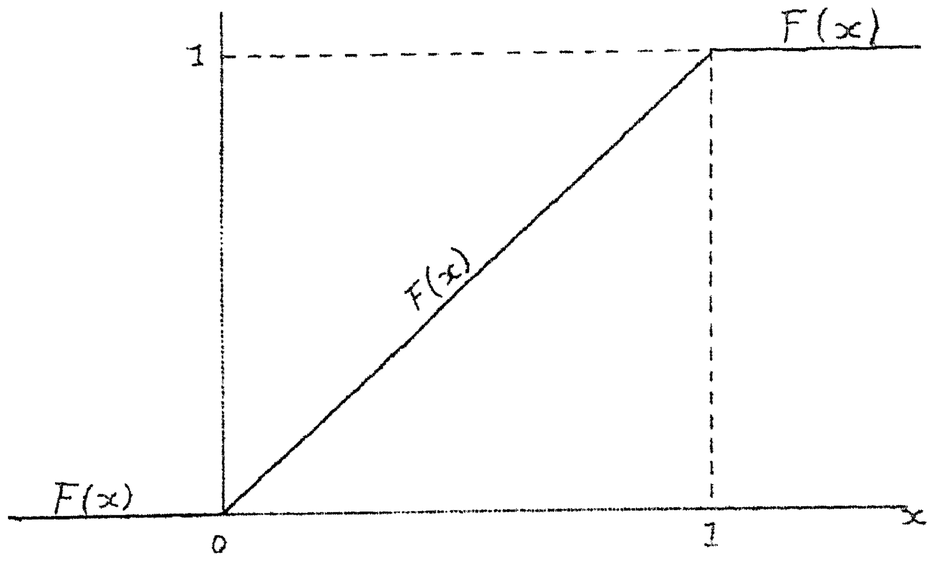
1.3.a (Blz.5)

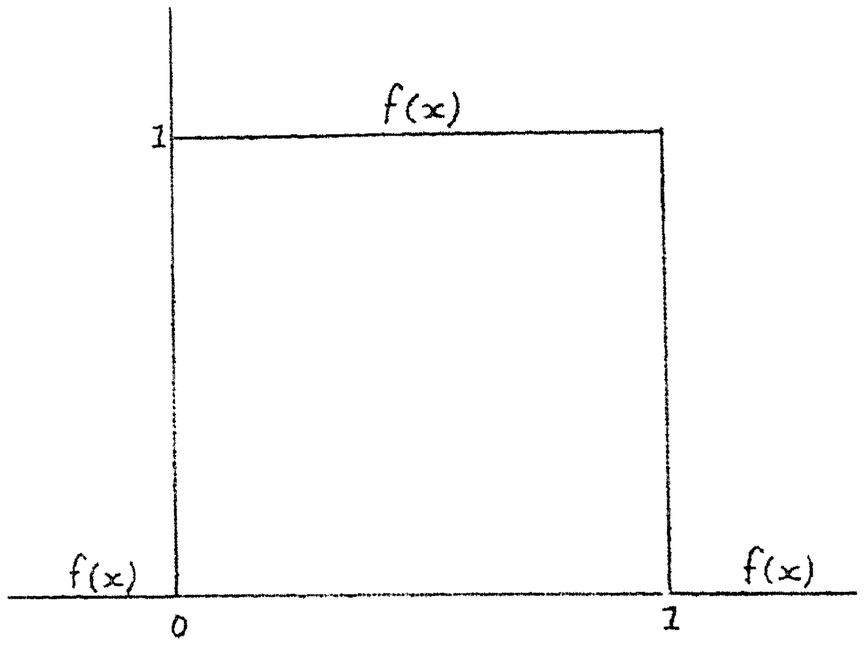
	n = 4	n = 5
x	$P[\underline{x} = x]$	$P[\underline{x} = x]$
0	q^4	q^5
1	$4pq^3$	$5pq^4$
2	$6p^2q^2$	$10p^2q^3$
3	$4p^3q$	$10p^3q^2$
4	p^4	$5p^4q$
5		p^5

1.4.a (Blz.7)

x	$P[\underline{x} = x]$	F(x)
0	0,168	0,168
1	0,360	0,528
2	0,309	0,837
3	0,132	0,969
4	0,028	0,998
5	0,002	1

1.5.a (Blz.9)





1.5.b (Blz.10)

1) Als $0 \leq x \leq 1$: $f(x) = x$; $F(x) = \frac{1}{2}x^2$.

2) Als $1 \leq x \leq 2$: $f(x) = 2 - x$; $F(x) = 1 - \frac{1}{2}(2-x)^2$.

3) $P[\underline{x} > x] = 1 - F(x) = 0,05 \longrightarrow \frac{1}{2}(2-x)^2 = 0,05 \longrightarrow x = 2 - \frac{1}{10}\sqrt{10} = 1,68.$

4) Volgens de errata moet de opgave luiden:

$$P[2 - x \leq \underline{x} \leq +x] = 0,98.$$

Aangezien de verdeling symmetrisch is t.o.v. $x = 1$ geldt:

$$P[2 - x \leq \underline{x} \leq +x] = 1 - P[\underline{x} \leq 2 - x \text{ of } \underline{x} \geq x] =$$

$= 1 - 2\,P[\underline{x} \geq x] = 0,98$. Dus $P[\underline{x} \geq x] = 0,01$ of $\frac{1}{2}(2-x)^2 = 0,01$,

$$x = 2 - \frac{1}{10}\sqrt{2} = 1,86.$$

1.9.a (Blz.18)

x+y	2	3	4	5	6	7	8	9	10	11	12
x-y	-5	-4	-3	-2	-1	0	1	2	3	4	5
kans	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

1.9.b (Blz.19) x+y: als in 1.9.a;

x-y	-5	-4	-3	-2	-1	0
kans	$\frac{2}{36}$	$\frac{4}{36}$	$\frac{6}{36}$	$\frac{8}{36}$	$\frac{10}{36}$	$\frac{6}{36}$

1.10.a 1) $\underline{x}$ en y zijn <u>onafhankelijk</u>, $f(x,y) = g(x) \cdot g(y)$
waarbij geldt: $g(x) = 1$ voor: $0 \leqq x \leqq 1$
$= 0$ voor: $x < 0$ en $x > 1$.

2) $\underline{x}$ en y zijn <u>afhankelijk</u>, want de grenzen van het gebied waarbinnen $f(x,y) > 0$ is zijn niet evenwijdig aan de coördinaatassen.

3) $\underline{x}$ en y zijn <u>afhankelijk</u> (beschouw de verticale doorsneden door de pyramide evenwijdig aan de coördinaatassen).

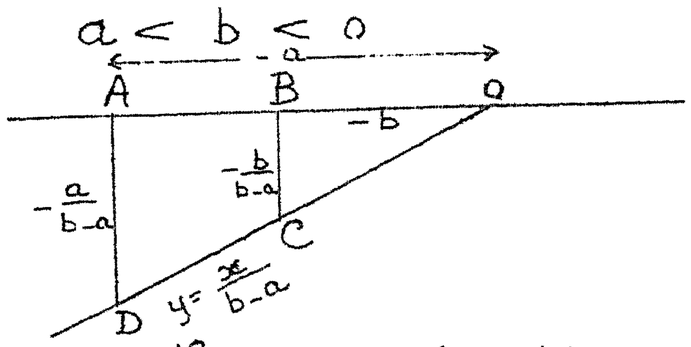
2.4.a (Blz.28) $p, 2p, 4p, 5p$ algemeen: np.

2.4.b (Blz.30) $\mathcal{E}\,\underline{x} = 7$ (formule 2.4 1).

2.5.a (Blz.31) Ja; als $\underline{x}$ de waarde $x_i \neq 0$ aanneemt, neemt zij ook de waarde $-x_i$ aan met dezelfde kans, de corresponderende termen in de uitdrukking voor $\mathcal{E}\,\underline{x}$ vallen tegen elkaar weg.

2.5.b (Blz.31) $f(x) = \frac{1}{b-a}$ dus $g(x) = \frac{x}{b-a}$ voor $a \leqq x \leqq b$
$= 0$ dus $= 0$ voor $x < a$ en $x > b$.

1. $a < b < 0$



Dus $\mathcal{E}\,\underline{x} = \frac{1}{2}(a+b)$.

$\sigma_+ = 0$
$\sigma_- = \sigma_{opp.\ trap.\ ABCD} =$
$= \frac{1}{2} AB.(BC + AD) =$
$= \frac{1}{2}(-a+b)\left(-\frac{a}{b-a} - \frac{b}{b-a}\right)$
$= -\frac{1}{2}(a+b)$.

2. $a < 0 < b$ wordt σ_+ de opp. van een driehoek met basis b en hoogte $\frac{b}{b-a}$, σ_- de opp. van een driehoek met basis $-a$ en hoogte $\frac{-a}{b-a}$; $\mathcal{E}\,\underline{x} = \frac{1}{2}(a+b)$.

3. $0 < a < b$ wordt $\sigma_- = 0, \sigma_+$ de opp. van een trapezium met hoogte $b-a$ en evenwijdige zijden $\frac{b}{b-a}$ en $\frac{a}{b-a}$; $\mathcal{E}\,\underline{x} = \frac{1}{2}(a+b)$.

2.6.a (Blz.32) Nee; de nieten worden niet met f 1000,-- verhoogd.

2.6.b (Blz.33) $\mathcal{E}(a\underline{x}+c) = \mathcal{E}\,a\underline{x} + c = a\,\mathcal{E}\,\underline{x} + c = a\mu + c$.

2.6.c (Blz.33) $\mathcal{E}\,\underline{x} = 0$; $\mathcal{E}\,a\underline{x} = 0$; $\mathcal{E}(\underline{x}+c) = c$; $\mathcal{E}(a\underline{x}+c) = c$.
Als $\underline{x}$ symmetrisch verdeeld is ten opzichte van $x=c$ geldt dus $\mathcal{E}\,\underline{x} = c$.

2.8.a (Blz.40) $\mathcal{E}\,\underline{x} = 0$ (Het betreft hier de som van n stochastische variabelen die alle symmetrisch zijn t.o.v. 0).

2.10.a (Blz.43) $\mathcal{E}\,\underline{x}^2 = 54\frac{5}{6}$ (formule 2.10.1).

2.11.a (Blz.46)

$\mathcal{E}\,\underline{x}^2 = 54\frac{5}{6}$ (opgave 2.10.a)

$\mathcal{E}\,\underline{x} = 7$ (opgave 2.4.b) : $(\mathcal{E}\,\underline{x})^2 = 49$

$\sigma^2(\underline{x}) = 5\frac{5}{6}$ (formule 2.11.2)

2.12.a (Blz.48) $\mathcal{E}\,\underline{x} = \mathcal{E}\,y = \frac{1}{2}$; $\mathcal{E}(\underline{x}+y) = 1$ (formule 2.7.1)

In geval 1: $\sigma^2(\underline{x}) = \sigma^2(y) = \frac{1}{12}$ (Blz.45)

$\sigma^2(\underline{x}+y) = \sigma^2(\underline{x}) + \sigma^2(y) = \frac{1}{6}$.

In geval 2: $\underline{x} + y = 2\underline{x}$

$\sigma^2(\underline{x}+y) = 4\sigma^2(\underline{x}) = \frac{1}{3}$.

2.13.a (Blz.51) $\sigma^2(\overline{\underline{x}}') = 0$; de steekproef omvat de gehele populatie; $\overline{\underline{x}}'$ is dus niet stochastisch meer.

2.13.b (Blz.51) $\sigma^2(\overline{\underline{x}}') = \frac{900}{100.999}\,\sigma^2(\underline{x})$. (formule 2.13.10)

$\sigma^2(\overline{\underline{x}}) = \frac{\sigma^2(\underline{x})}{m}$ bij m appels met teruglegging.

Los dus m op uit: $\frac{1}{m} = \frac{900}{100.999}$ dus $m = 111$

Bij $n = 500$: $\frac{1}{m} = \frac{500}{500.999}$. Dus $m = 999$.

Bij $n = 800$: $\frac{1}{m} = \frac{200}{800.999}$. Dus $m = 3996$.

2.13.c (Blz.51) Als de spreiding van de gewichten der appels σ is, en de gemiddelden van alle waarnemingen in gevallen 1, 2, 3 en 4 worden voorgesteld door $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, $\bar{x}_4$ geldt:

$\sigma(\bar{x}_1) = \frac{1}{\sqrt{20}}\sigma = 0,224\sigma$ (formule 2.13.6)

$\sigma(\bar{x}_2) = \sqrt{\frac{1}{4}\cdot\frac{1}{5}}\sigma = 0,224\sigma$ (tweemaal formule 2.13.6)

$\sigma(\bar{x}_3) = \sqrt{\frac{1}{4}\cdot\frac{1000-5}{5.999}}\sigma = 0,223\sigma$ (formule 2.13.6 en 2.13.10)

$\sigma(\bar{x}_4) = \sqrt{\frac{1000-20}{20.999}}\sigma = 0,221\sigma$ (formule 2.13.6)

De vierde methode is dus de nauwkeurigste, doch het verschil met de andere methoden is nog gering.

2.14.a (Blz.55) Beide schattingen zijn zuiver.

2.14.b In het eerste geval niet, in het tweede geval wordt s'^2 met 10.000 vermenigvuldigd.

2.14.c (Blz.55)

$\sum_{i=1}^{8}(x_i - 20) = 0,64$; $\sum_{i=1}^{8} x_i = 160,64$; $\bar{x} = 20,08$.

$\sum_{i=1}^{8}(x_i - 20)^2 = 0,2488$.

$\sum_{i=1}^{8}(x_i - \bar{x})^2 = \sum_{i=1}^{8}(x_i - 20)^2 - \frac{\{\sum_{i=1}^{8}(x_i - 20)\}^2}{8} = 0,2488 - 0,0512 = 0,1976$.

$s'^2 = \frac{1}{7}\sum_{i=1}^{8}(x_i - \bar{x})^2 = 0,0282$.

2.15.a (Blz.56) 1) $\underline{x} - np$. 2) $\frac{\underline{x}}{\sqrt{np(1-p)}}$. 3) $\frac{\underline{x} - np}{\sqrt{np(1-p)}}$.

2.15.b (Blz.57) 1. De waarnemingen van $\frac{\underline{\bar{x}}}{\sigma(\underline{x})}$ worden : $\frac{x_i - 1}{2}$; men vindt:

- 0,40	- 0,28	1,23	0,90	- 1
+ 0,34	+ 0,81	- 0,10	+ 0,98	+ 0,55

2. $\bar{x} = 1,606$; $s'^2 = 2,0835$; $s' = 1,4434$.

De gereduceerde en gestandaardiseerde steekproefwaarden zijn : $\frac{x_i - \bar{x}}{s'}$; men vindt:

- 0,97	- 0,81	+ 1,28	+ 0,83	- 1,81
+ 0,05	+ 0,70	- 0,56	+ 0,94	+ 0,34

3.1.a (Blz.60) 0,200; 0,010; 0,999.

3.1.b (Blz.60) 0,800; 0,990; 0,001.

3.1.c (Blz.60) 1; 0,0004.

3.1.d (Blz.60) 1) $x = 1,282$.

2) $x = -1,645$.

3) $x = 2,575$.

3.2.a (Blz.62) Variantie is 1,44; spreiding is dus 1,2; y is $N(0,1)$-verdeeld.

1) $P[\underline{x} \leq 0] = P\left[y \leq \frac{-1,37}{1,2}\right] = 0,127$.

2) $2\,P[\underline{x} \geq 4] = 2\,P\left[y \geq \frac{4-1,37}{1,2}\right] = 0,028$.

3) $P[\underline{x} \geq x] = P\left[y \geq \frac{x-1,37}{1,2}\right] = 0,01 \longrightarrow x = 4,16$.

4) Noem de te bepalen waarden x_1 en x_2, dan worden deze gevonden uit:

$$P\left[y \geq \frac{x_1 - 1,37}{1,2}\right] = 0,0005 \text{ en } P\left[y \leq \frac{x_2 - 1,37}{1,2}\right] = 0,0005$$

of uit: $\frac{x_1 - 1,37}{1,2} = +3,3 \longrightarrow x_1 = 5,33$.

$$\frac{x_2 - 1,37}{1,2} = -3,3 \longrightarrow x_2 = -2,59.$$

3.3.a (Blz.64) $\underline{x}_1 - \underline{x}_2$: $N(0, \sqrt{2})$.

$4\underline{x}_1 - 3\underline{x}_2 - 1$: $N(\mu - 1, \sqrt{4^2+3^2}) = N(\mu - 1, 5)$.

3.3.b (Blz.64) $\overline{x} = 0,254$; y is $N(0,1)$-verdeeld.

1) $P[\overline{x} \geq 0,254] = P[y \geq (0,254 - \mu)\sqrt{5}] = 0,05 \longrightarrow (0,254 - \mu)\sqrt{5} = 1,645 \longrightarrow \mu = -0,482$.

De overschrijdingskans is groter dan 0,05 als $\mu > -0,482$.

2) Analoog: $\mu = 0,990$; overschrijdingskans groter dan 0,05 als $\mu < 0,990$.

3) $P[\overline{x} \geq 0,254] = P[y \geq (0,254 - \mu_1)\sqrt{5}] = 0,025 \longrightarrow \mu_1 = -0,623$.

$P[\overline{x} \leq 0,254] = P[y \leq (0,254 - \mu_2)\sqrt{5}] = 0,025 \longrightarrow \mu_2 = 1,131$.

De overschrijdingskans is groter dan 0,05 als

$-0,623 < \mu < 1,131$.

3.4.a (Blz.67)

$$P\left[\left|\frac{x}{n} - p\right| \geq 0,01\right] = P[|x - np| \geq 0,01n]$$

$$\approx P\left[|y| \geq \frac{0,01n}{\sqrt{np(1-p)}}\right] \quad (y : N(0,1)\text{-verdeeld}).$$

Bij toepassing van een continuïteitscorrectie (zie par.3.6) kiest men : $P\left[|y| \geq \frac{0,01n - 0,5}{\sqrt{np(1-p)}}\right]$.

Men vindt:

a) Zonder continuïteitscorrectie:

p \ n	10^2	10^4	10^6
0,5	0,84	0,046	0,000
0,2	0,80	0,012	0,000

b) Met continuïteitscorrectie:

p \ n	10^2	10^4	10^6
0,5	0,92	0,046	0,000
0,2	0,90	0,013	0,000

3.5.a (Blz.72) Men kan te werk gaan op een wijze analoog aan hetgeen in par. 1.9 over de verdeling van $\underline{x} + \underline{y}$ is geschreven.

3.5.b (Blz.72) Analoog aan de combinatie van u_I en u_{II}; De overschrijdingskans wordt 0,001.

3.5 c (Blz.72) 1) Zonder A;
Vergelijking van I en III: $u = 30 > u_\alpha$ (zie errata); I sterker dan III.
Vergelijking van II en III: $u = 27 = u_\alpha$; II sterker dan III.

Met A:
Vergelijking van I en III: $u = 30 > u_\alpha$; I sterker dan III.
Vergelijking van II en III: $u = 30 > u_\alpha$; II sterker dan III.

2) Vergelijking "zonder A" en "met A" bij III: $u = 13$; $u_\alpha = 5$; de hypothese kan dus niet verworpen worden.

3) $V = u_I + u_{II} + u_{III} = 5 + 6 + 13 = 24.$

$$\mathcal{E}\,\underline{V} = \mathcal{E}\,\underline{u}_I + \mathcal{E}\,\underline{u}_{II} + \mathcal{E}\,\underline{u}_{III} = \tfrac{1}{2}(5\times5 + 5\times5 + 6\times6) = 43.$$

$$\sigma^2(\underline{V}) = \sigma^2(\underline{u}_I) + \sigma^2(\underline{u}_{II}) + \sigma^2(\underline{u}_{III}) = 45,83 + \tfrac{1}{12}\times6\times6\times13 = 84,83;$$

$$\sigma(\underline{V}) = 9,21. \quad (\sigma^2(\underline{u}_I) + \sigma^2(\underline{u}_{II}) = 45,83, \text{ zie blz.71})$$

Nu is: $\dfrac{V - \mathcal{E}\,\underline{V} + \frac{1}{2}}{\sigma(\underline{V})} = \dfrac{24 - 43 + \frac{1}{2}}{9,21} = -2,01.$

Tweezijdige overschrijdingskans 0,0444.

Conclusie: methode A maakt de garens sterker.

3.6.a (Blz.75) De normale verdeling met $\mu = np = 5$ en $\sigma = \sqrt{np(1-p)} = \sqrt{2,5} = 1,581$.

$P[\underline{x} = 4]$(exact) $= \binom{n}{4} p^4 (1-p)^6 = \binom{10}{4} 0,5^{10} =$

$= \frac{10.9.8.7}{1.2.3.4} \cdot \frac{1}{2^{10}} = 0,205.$

$P[\underline{x} = 4]$(benaderd) $= P\left[\frac{3\frac{1}{2} - 5}{1,581} \leqq y \leqq \frac{4\frac{1}{2} - 5}{1,581}\right]$ $(y : N(0,1))$

$= P[-0,949 \leqq y \leqq -0,316] = 0,205.$

Analoog vindt men:

$P[\underline{x} \geqq 1]$(exact) $= 1 - P[\underline{x} = 0] = 0,999.$

$P[\underline{x} \geqq 1]$(benaderd) $= 1 - P\left[y \leqq \frac{\frac{1}{2} - 5}{1,581}\right] = 0,998.$

$2\,P[\underline{x} \leqq 2]$(exact) $= 2 \cdot 0,0547 = 0,109.$

$2\,P[\underline{x} \leqq 2]$(benaderd) $= 2 \cdot P\left[y \leqq \frac{2\frac{1}{2} - 5}{1,581}\right] = 0,114.$

3.6.b (Blz.75) $\mathcal{E}\,\underline{x} = 10 \cdot \frac{1}{2} = 5$ (Blz. 45 bovenaan en Eig. 4 blz. 37).

$\sigma^2(\underline{x}) = 10 \cdot \frac{1}{12} = \frac{5}{6}$ (Blz. 45 bovenaan en Eig. 9 blz. 48).

Dus is de aangepaste normale verdeling : $N(5, \sqrt{\frac{5}{6}})$.

$2\,P[\underline{x} \geqq 9] = 2\,P\left[y \geqq \frac{9-5}{\sqrt{\frac{5}{6}}}\right] = 2\,P[y \geqq 4,382] << 0,0004.$

$(y : N(0,1)$-verdeeld)

4.3.a (Blz.86) 1) +1 . 2) -1 . 3) $-\frac{1}{2}$. 4) 0 . 5) $+\frac{1}{2}$.

4.6.a (Blz.93) $S = -69$. $r = -\frac{2.69}{13.12} = -\frac{23}{26} = -0,88.$

4.7.a (Blz.97) Bij $\alpha = 0,01$ is de linker kritieke waarde -44 (vgl. tab. 4.II).
Er is dus negatieve correlatie.

Normale benadering: $\sigma = \sqrt{\frac{1}{18} \cdot 13.12.31} = 16,39.$

$\frac{|s| - 1}{\sigma} = 4,149.$ Dus overschrijdingskans $<< 0,0004.$

Errata bij de cursus Toegepaste Statistiek II (Rapport S 155)

Hoofdstuk I

Blz.

2 Voetnoot, regel 2 v.o.:"uikomst"; moet zijn:"uitkomst".

10 Vraagstuk 1.5.b. Punt 4

$$P[-x \leqq \underline{x} \leqq +x] = 0{,}9\delta$$

moet zijn:

$$P[-x \leqq \underline{x} \leqq +x] = 0{,}9\delta$$

12 Fig. 1.6. Onder histogram binomiale verdeling toevoegen:

$n = 10$; $p = 0{,}4$.

15 Regel 14 v.o. achter "vertonen" toevoegen: "of niet geleidelijk verlopen".

20 Regel 3 v.o. De zin:"In dat geval...aan de y-as" moet vervangen worden door: "In dat geval zijn dus de niveaux van overeenkomstige punten van het berglandschap in twee doorsneden evenwijdig aan de x-as evenredig. Dit geldt mutatis mutandis ook voor twee verticale doorsneden evenwijdig aan de y-as."

21 Regel 2 v.b., laatste gedeelte. Er moet staan:

"en/of $k(y)$ nul is en"

Regel 4 en 5 v.o. "de hoofdstukken" moet zijn "het hoodstuk".

Hoofdstuk II

28 Regel 14 v.b. 3e term "$2.3.pq^2$ " moet zijn "$2.3.p^2q$ ".

30 Titel par. 2.5 moet zijn:

"De verwachting van <u>een</u> continue verdeling."

31 Regel 9 v.o. laatste woord: dis-

" 7 v.o. " " : dus

" 6 v.o. " " : de

" 6 v.o. aan het begin: $\mathcal{E}\,\underline{x} = \sum_i p_i x_i$

" 5 v.o. laatste formule $\mathcal{E}\,\underline{x} = \int_{-\infty}^{+\infty} x\, f(x)\, dx$

" 4 v.o. laatste woord: gemiddel-

" 3 v.o. " " : gelden

32 Regel 16 v.b. "$\mathcal{E}\underline{x}$ " moet zijn "$\mathcal{E}(\underline{x}+c)$ "

37 Regel 14 v.b. "geschiedt" moet zijn "geschied".

40 Regel 10 v.b. "zie par. 2.14" moet zijn "zie par. 2.13"

42 In fig. 2.4 moet het vlakdeel tussen de parabool, de x-as en de lijn $y = 1$ gearceerd worden.

46 Formule (2.12.2) " $\sigma\{a\underline{x}\} = a\sigma\{\underline{x}\}$ " moet zijn

" $\sigma\{a\underline{x}\} = |a|\,\sigma\{\underline{x}\}$ "

Blz.

50 Regel 14 v.o. formule (2.13.8): " $\sigma\{x\}$ " moet zijn " $\sigma^2\{x\}$ "

51 Regel 2 v.b. $\sqrt{\frac{N-n}{n(N-1)}}$ moet zijn $\sqrt{\frac{N-n}{N-1}}$

Hoofdstuk III

61 Regel 7 v.b. "$P[y \geqq 2{,}00]$" moet zijn "$P[y \geqq -2{,}00]$"

67 Regel 11 v.o. " $\frac{1}{2}mn(m+n+1)$" moet zijn: "$\frac{1}{12}mn(m+n+1)$"

69 Regel 9 v.b. " $= 5 + 5 + 5 + 3 + 3 = 23$" moet zijn:
$= 5 + 5 + 5 + 4 + 4 = 23$".

72 Regel 17 v.o. toevoegen: "en voor $m = 6$ $n = 5$ $u_2 = 3$ $u_n = 27$"

77 Tabel 3.II Totaal laatste kolom moet zijn: "999"
Regel 12 v.o. Haakje toe te voegen achter de uitdrukking onder het eerste wortelteken

Tabel normale verdeling regel 2 v.b.: " $\xi_\alpha = 0{,}00(0{,}01)3{,}09$"
moet zijn "$\xi_\alpha = 0{,}00(0{,}01)3{,}49$".

Hoofdstuk IV

78 Regel 2 v.o.: "De belasting, die en betaalt" moet zijn:
"De belasting, die men betaalt".

80 Regel 4 v.b. "eventueel" schrappen.
Regel 7 v.b. Lees: "te delen door de spreidingen van x en y ".

81 Regel 8 v.b. " $\mathcal{E}\tilde{y}^2 = \alpha\,\mathcal{E}\tilde{x}^2$ " moet zijn " $\mathcal{E}\tilde{y}^2 = \alpha^2\,\mathcal{E}\tilde{x}^2$ "

83 Regel 3, 4, 6, en 7 v.b. In de formules "c" te vervangen door "2c"

84 Regel 13 v.o. "u" moet zijn "$\underline{u}$".

86 Opgave 4.3.a 2) te lezen: "A(-1,0) B(0,-2) C(1,-4)"
Regel 15 v.o. " $\sigma^2(a)$ en $\sigma^2(b)$ " moet zijn " $\sigma\{a\}$ en $\sigma\{b\}$ "
Regel 5 v.o. " $\sigma^2(d)$ " moet zijn " $\sigma\{d\}$ "

90 Regel 6 v.b. "onbetrouwbaarheidsinterval" moet zijn "betrouwbaarheidsinterval"
Voetnoot: "method" moet zijn "methods"

92 Regel 19 v.b. "Om daarbij te kunnen berekenen" moet zijn "Om k daarbij te kunnen berekenen".

93 Regel 7 v.o. "waar" moet zijn "waarbij".

98 Regel 11 v.o. "fiet" moet zijn "feit".
"gropp" moet zijn "groep".

101 Regel 2 v.b. moet zijn: $\sigma^2(\bar{z}_i) = \frac{2}{9} \cdot \frac{13}{4.3} = \frac{13}{54}$
Regel 10 v.b. moet zijn: $\sigma^2(\bar{I}) = 4 \cdot \frac{1}{6} + 8 \cdot \frac{13}{54} = 2\frac{16}{27}$
12 v.b. moet zijn: $\sigma(\bar{I}) = \sqrt{2\frac{16}{27}} = \frac{14{,}49}{9} = 1{,}61$
5 v.o. $\bar{I}/\sigma(\bar{I}) = 3{,}44$ moet zijn: $\bar{I}/\sigma(\bar{I}) = 3{,}73$
4 v.o. $k = 0{,}0006$ " " $k = 0{,}0002$
