STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Rapport S 265 (C13)

Leergang Besliskunde

Hoofdstuk IV

Limietstellingen

door

J. Kriens

(Bewerking van de hoofdstukken 10 en 11 van rapport S 200 (C8)
door Prof. Dr J. Hemelrijk)

Februari 1960

1. De theoretische wet van de grote aantallen.

Limietstellingen vormen een zeer belangrijk gedeelte van de waarschijnlijkheidsrekening en de mathematische statistiek. Gezien het karakter van deze leergang kunnen wij er niet lang bij blijven stilstaan en zullen wij ons moeten beperken tot het noemen van de voor ons belangrijkste stellingen. Een uitvoerige behandeling en bewijzen kan men vinden in ieder fundamenteel boek over statistiek en waarschijnlijkheidsrekening, bijvoorbeeld in de aan het eind van paragraaf 1, hoofdstuk I, genoemde boeken van M.G. KENDALL en H. CRAMÉR.

In paragraaf 2 van hoofdstuk II hebben wij de limiet van een rij getallen als volgt gedefinieerd:
Een oneindige rij getallen $a_1, a_2, \ldots$ convergeert naar de limiet a, wanneer bij ieder positief getal ε een getal N gevonden kan worden, zodanig dat voor $n > N$ geldt

$$|a_n - a| < \varepsilon$$

Dit betekent dus dat vanaf de index N de getallen a_n minder dan ε verwijderd zijn van de limiet a.

Een dergelijke definitie kan bij stochastische grootheden niet gegeven worden. Onderzoekt men bijvoorbeeld een rij stochastische grootheden $\underline{x}_1, \underline{x}_2, \ldots$, dan zal niet alleen $\underline{x}_n$ voor iedere n nog geheel verschillende waarden aan kunnen nemen, maar hetzelfde zal in het algemeen gelden voor iedere functie, die men van deze grootheden kan vormen. Men zal dus een ander, zwakker convergentiebegrip in moeten voeren, dan dat voor rijen getallen. Dit gebeurt door niet te eisen dat de stochastische grootheden $\underline{x}_n$ vanaf een bepaalde n alleen waarden dichtbij een bepaald getal, bijvoorbeeld μ aannemen, maar door slechts te eisen dat de kans op een waarde van $\underline{x}_n$, welke meer dan ε van μ verwijderd is, naar nul convergeert. Deze kans is $P[|\underline{x}_n - \mu| > \varepsilon]$ en wij zeggen nu dat $\underline{x}_n$ stochastisch naar μ convergeert voor $n \to \infty$, als deze kans voor $n \to \infty$ tot nul nadert; ook spreekt men wel van convergentie in waarschijnlijkheid. Stelling 1;1 bevat een eigenschap van dit type van convergentie.

Stelling 1;1.

Als $\underline{x}_1, \underline{x}_2, \ldots$ onafhankelijk verdeelde stochastische groot-

heden zijn, die alle dezelfde verwachting μ en variantie σ^2 bezitten, dan convergeert, voor $n \to \infty$, het gemiddelde

(1;1) $$\bar{X}_n \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} X_i$$

stochastisch naar μ, d.w.z dat voor iedere $\varepsilon > 0$ geldt

(1;2) $$\lim_{n \to \infty} P\left[|\bar{x}_n - \mu| > \varepsilon \right] = 0$$

Een speciaal geval van deze stelling wordt verkregen door voor de x_i het aantal successen (0 of 1) te nemen bij een experiment met kans p op succes. De stelling neemt dan de volgende vorm aan.

Stelling 1;2. (Theoretische wet der grote aantallen)

In een reeks van n onafhankelijke experimenten, ieder met kans p op succes, convergeert de fractie successen x/n voor $n \to \infty$ in waarschijnlijkheid naar p, d.w.z. voor iedere $\varepsilon > 0$ is

(1;3) $$\lim_{n \to \infty} P\left[|x/n - p| \geq \varepsilon \right] = 0$$

Opmerkingen

Deze stelling heeft een grote praktische betekenis. Het is de theoretische tegenhanger van de experimentele wet van de grote aantallen (zie paragraaf I,2). Het zou niet juist zijn te zeggen, dat deze laatste nu "bewezen" is, daar een stelling, die op de werkelijkheid betrekking heeft, nooit op grond van een axiomastelsel bewezen kan worden. Alleen experimentele bevestiging is daar geschikt voor. Wel kunnen wij nu echter opmerken, dat de uitermate eenvoudige axioma's, waarvan wij zijn uitgegaan, adaequaat blijken te zijn ter beschrijving van de nogal ingewikkelde experimentele wet van de grote aantallen, die wij fundamenteel voor statistische verschijnselen hebben genoemd. Dit betekent, dat de interpretatie van een kans p als - bij benadering - een fq in een lange reeks waarnemingen bevestigd is, terwijl bovendien dit succes van de theorie het vertrouwen in verdere resultaten - d.w.z. het vertrouwen dat de toepassingen daarvan in de praktijk bruikbaar zullen blijken te zijn - vergroot. De op blz. 20 van hoofdstuk I gemaakte opmerking, "dat in het kansmodel het fq als zodanig, met zijn statistische fluctuaties, op natuur-

lijke wijze weer ingevoerd kan worden", is hiermede waar gemaakt.

De praktische betekenis van stelling 1;1 is de volgende. Indien men van een stochastische grootheid $\underline{x}$ met verwachting μ en variantie σ^2, een aantal onafhankelijke waarnemingen verricht, kan de i^e waarneming voorgesteld worden door $\underline{x}_i$ en is de stelling toepasbaar, daar alle $\underline{x}_i$ nu dezelfde verdeling (nl. dezelfde verdeling als $\underline{x}$) bezitten. Formule (1;1) betekent nu, dat het r enkundige gemiddelde der waarnemingen, naarmate n groter wordt, steeds minder kans bezit om ver van μ verwijderd te zijn. Als men n maar groot genoeg maakt, kan men - behoudens een willekeurig kleine kans - zorgen willekeurig dicht bij μ terecht te komen.

Indien men bijv. een physische of chemische constante (gewicht, soortelijk gewicht, titer van een vloeistof, lading van een electron, enz.) nauwkeurig wenst te bepalen, kan men een aantal onafhankelijke waarnemingen daarvan verrichten. Ieder daarvan is aan meetfouten onderhevig en kan daarom als een stochastische grootheid beschouwd worden. Worden de waarnemingen op dezelfde wijze (door dezelfde personen en met dezelfde apparatuur) verricht, dan bezitten ze alle dezelfde kansverdeling. Worden bovendien de uitkomsten van latere waarnemingen niet beïnvloed door die van de voorafgaande - gevaar voor psychologische beïnvloeding is vaak in sterke mate aanwezig -, dan kunnen zij als stochastisch onafhankelijk beschouwd worden. In deze omstandigheden is de stelling toepasbaar. Veelal zal men n dan niet zeer groot kunnen (of willen) maken, maar ook bij kleine n is de variabiliteit van een gemiddelde van een aantal onafhankelijke waarnemingen aanzienlijk kleiner, dus de nauwkeurigheid aanzienlijk groter dan van één enkele waarneming.

De grootheid μ, waarbij $\bar{\underline{x}}_n$ in de buurt komt, hangt vaak mede van de waarnemingsapparatuur af. Men dient zich daarvan bij het verrichten van metingen steeds bewust te zijn (hetgeen niet altijd het geval is). Bij de bepaling van een titer van een vloeistof bijv. kunnen afwijkingen in de gebruikte buret en pipet systematische verschillen veroorzaken tussen μ en de titer, die men eigenlijk wenst te meten. Het is dan zinloos de nauwkeurigheid van de bepaling, door n op te voeren, zeer groot te maken,

daar de systematische fout dan gaat overheersen. Men kan daaraan tegemoet komen door de apparatuur te wisselen. Zolang de spreiding van de met verschillende apparaturen verkregen gemiddelden $\bar{x}_n$ nog groot zijn in vergelijking met de systematische verschillen tussen de bij de apparaturen behorende μ's, is vergroting van het aantal waarnemingen bij ieder der apparaturen in principe zinvol. Zijn de systematische verschillen groot, dan is grot r nauwkeurigheid alleen te bereiken door verbetering van de apparatuur.

Laat men verschillende waarnemers werken met dezelfde of verschillende apparatuur, maar zijn er geen (of nauwelijks) systematische verschillen, zodat μ voor alle waarnemingen gelijk is, dan is het van belang op te merken, dat stelling 1;1 ook blijft gelden als de varianties ongelijk zijn, mits deze begrensd zijn. Is $\sigma^2(x_i) \leqq \sigma_i^2$ voor alle i, dan is nl. het bewijs, met een kleine wijziging, aan te passen. De voorwaarden kunnen trouwens nog verder verzwakt worden, doch dit punt zullen wij hier buiten beschouwing laten.

De uitwerking van de hierboven geschetste gedachtengang behoort tot de in hoofdstuk V te behandelen schattingstheorie.

2. Functies van stochastische grootheden.

Reeds eerder is opgemerkt, dat een functie van één of meer stochastische grootheden zelf in het algemeen weer een stochastische grootheid is. Eén der belangrijkste voorbeelden hiervan is het gemiddelde van een aantal stochastische grootheden dat reeds in paragraaf 1 ter sprake is geweest.

De volgende stelling geeft ons een middel om de kansverdeling van een dergelijke functie te berekenen. Evenals in de tweede paragraaf van hoofdstuk III formuleren wij de stellingen voor tweedimensionale kansverdelingen. Ze zijn echter algemeen geldig en kunnen voor n-dimensionale kansverdelingen op analoge wijze toegepast worden als voor twee-dimensionale verdelingen.

Stelling 2;1.

Laten $\underline{x}_1$ en $\underline{x}_2$ een simultane kansverdeling bezitten met $f(x_1,x_2)$ als kansdichtheid en laat $\varphi(x_1,x_2)$ een functie van x_1 en x_2 zijn. Laat verder, in de twee-dimensionale ruimte R_2

met x_1 en x_2 als coördinaten, G_a het gebied voorstellen van al die punten, waarvoor geldt

(2;1) $\varphi(x_1, x_2) \leq a$

Dan is

(2;2) $$F_\varphi(a) = P[\varphi \leq a] = P[\varphi(x_1, x_2) \leq a] = P[P \leq G_a] = \iint_{G_a} f(x_1, x_2)\, dx_1\, dx_2$$

Hierin stelt $\underline{P}$ het stochastische punt $(\underline{x}_1, \underline{x}_2)$ voor. In geval van een discrete verdeling wordt in het laatste lid van (2;2) de tweevoudige integraal door een dubbele som vervangen. De stelling volgt uit (2;8) uit hoofdstuk III.

Hieronder volgen twee theoretische en één numerieke toepassing van deze stelling.

<u>Stelling 2;2.</u>

Als $\underline{x}$ een $N(\mu, \sigma)$-verdeelde grootheid is, dan is

(2;3) $$\underline{u} = \frac{x - \mu}{\sigma}$$

$N(0,1)$-verdeeld.

<u>Bewijs</u>

Het gebied G_a, waarin $u \leq a$ is, wordt gegeven door $\frac{x - \mu}{\sigma} \leq a$ of $x \leq \mu + a\sigma$. Of, direct opgeschreven,

$$F_{\underline{u}}(u) \overset{def}{=} P[\underline{u} \leq u] = P[\underline{x} \leq \mu + u\sigma] = F_x(\mu + u\sigma)$$

De verdelingsdichtheid $f_u(u)$ wordt hieruit verkregen door differentiëren (vgl. stelling 6;3 uit hoofdstuk II):

$$f_{\underline{u}}(u) = \frac{d}{du} F_{\underline{u}}(u) = \frac{d}{du} F_x(\mu + u\sigma) = \sigma f_x(\mu + u\sigma) = \frac{\sigma}{\sigma\sqrt{2\pi}} e^{\frac{(\mu + u\sigma - \mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

Volgens (1;18) uit hoofdstuk III is het laatste lid de verdelingsdichtheid van een normale verdeling met parameters $\mu=0$ en $\sigma=1$, dus een $N(0,1)$-verdeling.

In de voorbeelden 4;11 en 4;12 uit hoofdstuk III werd deze stelling reeds toegepast.

Voorbeeld 2;1.[1)]

Een tentamen bestaat uit 5 vragen, die goed of fout beantwoord kunnen worden. Voor ieder goed antwoord wordt 1 punt gegeven, voor een fout antwoord 0.

a) Candidaat A heeft zich zodanig op het tentamen voorbereid, dat hij bij iedere vraag een kans $\frac{2}{3}$ heeft op het geven van het goede antwoord. Bereken, in de onderstelling van onafhankelijkheid der antwoorden, de kansverdeling van zijn eindcijfer en verwachting en spreiding daarvan.

b) Doe hetzelfde voor candidaat B, die voor de eerste 3 vragen dezelfde kans op een goed antwoord heeft als A, maar voor de laatste twee slechts een kans $\frac{1}{2}$.

Oplossing

a) Geven wij het eindcijfer van candidaat A aan met $\underline{x}$, dan bezit $\underline{x}$ een binomiale verdeling met $p = \frac{2}{3}$ en $n = 5$. De kansen worden dus gegeven door (vgl. (1;6) uit hoofdstuk III):

$$P\left[\underline{x} = k\right] = \binom{5}{k}\left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{5-k}$$

terwijl de verwachting is $\mathcal{E}\underline{x} = \frac{10}{3}$ en de variantie $\sigma^2(\underline{x}) = \frac{10}{9}$ (vgl. tabel 4;I uit hoofdstuk III); de spreiding is dus $\sqrt{\frac{10}{9}} = 1{,}05$.

b) Voor candidaat B noemen wij $\underline{x}$ het aantal punten dat hij met het oplossen van de vragen 1, 2 of 3 kan bereiken en $\underline{y}$ het aantal punten, dat hij met het oplossen van de opgaven 4 en 5 kan bereiken. Het eindcijfer, voorgesteld door $\underline{z}$, is dan dus gelijk aan

$$\underline{z} = \underline{x} + \underline{y}$$

Nu is

$$P\left[\underline{x} = k\right] = \binom{3}{k}\left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{3-k}$$

en

$$P\left[\underline{y} = l\right] = \binom{2}{l}\left(\frac{1}{2}\right)^l \left(\frac{1}{2}\right)^{2-l} = \binom{2}{l}\frac{1}{4}$$

De verzamelingen G_a uit stelling 2;1 kunnen nu aangegeven worden in een twee-dimensionale figuur; zie figuur 2;1.

1) Dit is opgave 73 van het reeds eerder genoemde opgavenboekje van Prof. Dr J. HEMELRIJK en Ir DORALINE WABEKE.

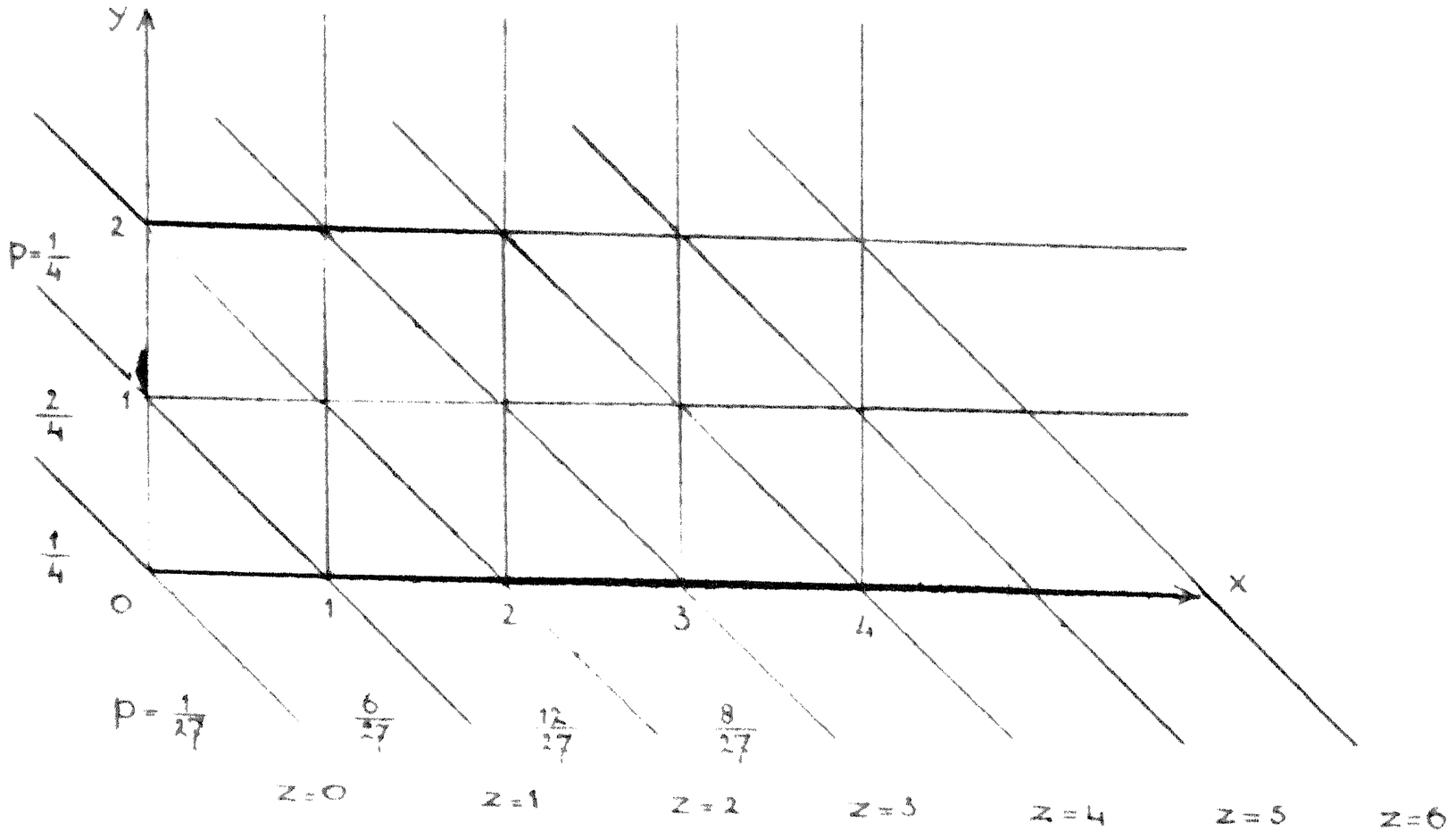


fig. 2;1

De kansverdeling van $\underline{z} = \underline{x}+\underline{y}$

We vinden op deze wijze

$$P[\underline{z}=0] = \frac{1}{4}\cdot\frac{1}{27} = \frac{1}{108}$$

$$P[\underline{z}=1] = \frac{2}{4}\cdot\frac{1}{27} + \frac{1}{4}\cdot\frac{6}{27} = \frac{8}{108}$$

$$P[\underline{z}=2] = \frac{1}{4}\cdot\frac{1}{27} + \frac{2}{4}\cdot\frac{6}{27} + \frac{1}{4}\cdot\frac{12}{27} = \frac{25}{108}$$

$$P[\underline{z}=3] = \frac{1}{4}\cdot\frac{6}{27} + \frac{2}{4}\cdot\frac{12}{27} + \frac{1}{4}\cdot\frac{8}{27} = \frac{38}{108}$$

$$P[\underline{z}=4] = \frac{1}{4}\cdot\frac{12}{27} + \frac{2}{4}\cdot\frac{8}{27} = \frac{28}{108}$$

$$P[\underline{z}=5] = \frac{1}{4}\cdot\frac{8}{27} = \frac{8}{108}\ .$$

De verwachting bedraagt

$$\mathcal{E}_{\underline{z}} = \mathcal{E}_{\underline{x}} + \mathcal{E}_{\underline{y}} = 3\,\tfrac{2}{3} + 2\,\tfrac{1}{2} = 3,$$

de variantie

$$\sigma^2(\underline{z}) = \sigma^2(\underline{x}) + \sigma^2(\underline{y}) = 3\,\tfrac{2}{3}\,\tfrac{1}{3} + 2\,\tfrac{1}{2}\,\tfrac{1}{2} = \tfrac{7}{6}$$

en de spreiding

$$\sigma(\underline{z}) = \sqrt{\tfrac{7}{6}} = 1{,}08$$

Een belangrijke toepassing van stelling 2;1 is tenslotte de

volgende.

Stelling 2;3.

Zijn $\underline{x}$ en $\underline{y}$ continu en onafhankelijk verdeelde stochastische grootheden met verdelingsfuncties F(x) en G(y) en verdelingsdichtheden f(x) en g(y), dan geldt voor verdelingsfunctie H(z) en verdelingsdichtheid h(z) van $\underline{z} = \underline{x}+\underline{y}$

$$(2;4) \qquad H(z) = \int_{-\infty}^{+\infty} f(x)\, G(z-x)\, dx$$

en

$$(2;5) \qquad h(z) = \int_{-\infty}^{+\infty} f(x)\, g(z-x)\, dx$$

Bewijs

Volgens (2;2) en stelling 3;3 uit hoofdstuk III is

$$H(z) = \iint_{x+y \leq z} f(x)\, g(y)\, dx\, dy = \int_{-\infty}^{+\infty} f(x)\, dx \int_{-\infty}^{z-x} g(y)\, dy = \int_{-\infty}^{+\infty} f(x)\, G(z-x)\, dx$$

Hieruit wordt h(z) verkregen door naar z te differentiëren (vgl. stelling 6;4 uit hoofdstuk II).

Deze stelling, die men bij berekeningen vaak moet toepassen, illustreren wij aan de hand van twee voorbeelden.

Stelling 2;4.

Zijn $\underline{x}$ en $\underline{y}$ onafhankelijk normaal verdeeld, dan is $\underline{z} = \underline{x}+\underline{y}$ eveneens normaal verdeeld.

Bewijs

De verdelingsdichtheden f(x) en g(y) zijn nu van de vorm

$$f(x) :: e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} \qquad \text{en} \qquad g(y) :: e^{-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2},$$

waarin :: betekent: op een constante factor na gelijk aan.

Het produkt f(x)g(z-x) heeft dus de vorm

$$f(x)\, g(z-x) :: e^{-\frac{1}{2}\left\{\left(\frac{x}{\sigma_1}\right)^2 - \frac{2\mu_1 x}{\sigma_1^2} + \left(\frac{z-x}{\sigma_2}\right)^2 - \frac{2\mu_2(z-x)}{\sigma_2^2}\right\}}.$$

De exponent kan (vergelijk blz. 14 van hoofdstuk III) gesplitst worden in twee delen, waarvan het ene een kwadratische vorm in z is (zonder x) en het tweede een volkomen kwadraat van

de vorm $(\alpha x+\beta z+\gamma)^2$. Het eerste deel kan als een e-macht voor de integraal van (2;5) gebracht worden en deze integraal zelf is een constante (substitueer daarin nl. $v=\alpha x+\beta z+\gamma$, dan veranderen de grenzen niet; deze blijven $-\infty$ en $+\infty$). Daaruit volgt echter de stelling.

Opmerkingen

Stelling 2;4 is direct (door inductie) uit te breiden tot meer dan 2 onafhankelijke normaal verdeelde grootheden.

De stelling geldt ook als deze grootheden niet stochastisch onafhankelijk zijn, doch een simultane normale verdeling bezitten. Het bewijs verloopt op geheel analoge wijze met behulp van de formules (2;9) en (2;2), waarin voor G_a het gebied $x+y \leq z$ genomen wordt.

Tevens is gemakkelijk in te zien, dat vermenigvuldiging van de stochastische grootheden met constante factoren de normaliteit van de simultane verdeling niet verstoort (vergelijk stelling 2;2), evenmin als het optellen van constante termen, zodat de volgende algemene stelling geldt.

Stelling 2;5.

Bezitten $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n$ een simultane normale verdeling, dan is ook iedere lineaire combinatie

(2;6) $\underline{z}=a_0+a_1\underline{x}_1+\ldots+a_n\underline{x}_n$

normaal verdeeld.

Stelling 2;6.

Zijn $\underline{x}$ en $\underline{y}$ onderling onafhankelijk verdeeld en bezit $\underline{x}$ een gamma-verdeling met parameters α en λ en $\underline{y}$ een exponentiële verdeling met parameter λ, dan heeft $\underline{z}=\underline{x}+\underline{y}$ een gamma-verdeling met parameters $\alpha+1$ en λ.

Bewijs

De verdelingsdichtheden van $\underline{x}$ en $\underline{y}$ zijn volgens (1;25) resp. (1;20) uit hoofdstuk III

$$\frac{\lambda^\alpha}{\Gamma(\alpha)}e^{-\lambda x}x^{\alpha-1} \quad \text{resp.} \quad \lambda e^{-\lambda y}.$$

Substitueren wij dit in (2;5), dan vinden wij met behulp van (5;15) en (5;30) uit hoofdstuk II

(2;6) $$h(z) = \int_0^z \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} \lambda e^{-\lambda(z-x)} dx =$$
$$\frac{\lambda^{\alpha+1}}{\Gamma(\alpha)} e^{-\lambda z} \int_0^z x^{\alpha-1} dx = \frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} e^{-\lambda z} z^\alpha,$$

hetgeen de verdelingsdichtheid van een gamma-verdeling met parameters $\alpha+1$ en λ voorstelt.

Opmerking

Uit stelling 2;6 volgt voor $\alpha = 1$, dat de som van twee exponentiële verdelingen met dezelfde parameter λ een gamma-verdeling bezit met parameters $\alpha=2$ en λ. Successievelijke toepassing van stelling 2;6 voor $\alpha=2,3,\ldots$ leidt tot de volgende stelling.

Stelling 2;7.

De som van k onderling onafhankelijke, exponentiëel verdeelde grootheden met parameter λ bezit een gamma-verdeling met parameters $\alpha = k$ en λ.

Ook deze stelling kan nog algemener gemaakt worden, doch daarop zullen wij hier niet ingaan. Wel merken wij op dat een stelling zoals 2;5 voor gamma-verdelingen niet geldt, evenmin als voor exponentiële verdelingen.

Wij besluiten deze paragraaf met een praktische toepassing van stelling 2;5.

Voorbeeld 2;2.[1)]

Een leverancier van flessen slaolie vermeldt als netto inhoud van zijn flessen: 350 gram olie. Een winkelier wenst deze bewering te controleren zonder de flessen te openen. Hij weegt daartoe een zeer groot aantal gevulde flessen en hij vindt voor dit gewicht een normale verdeling met gemiddelde 585,2 gram en spreiding 12,8 gram. Vervolgens weegt hij een groot aantal lege flessen die hij van zijn klanten teruggekregen heeft, met de bijbehorende sluitcapsules. Voor dit gewicht vindt hij eveneens een normale verdeling

1) Dit is opgave 67 uit het opgavenboekje van Prof. Dr J. HEMELRIJK en Ir DORALINE WABEKE.

met gemiddelde 228,3 gram en spreiding 11,3 gram.

Bereken, in de veronderstelling, dat het gewicht van de olie in de flessen normaal verdeeld is en onafhankelijk is van het gewicht van fles + sluiting:

a) de kansverdeling van het gewicht aan olie in de fles,
b) het percentage der flessen olie, die minder dan 350 gram olie bevatten,
c) de covariantie van het gewicht van een gevulde fles en dat van de olie, die erin zit.
d) Indien men nu, door zorgvuldiger vullen van de flessen, de spreiding van fles + inhoud, bij gelijkblijvend gemiddelde, zoveel kleiner wil maken, dat slechts 2,5% der flessen minder dan 350 gram olie bevat, tot welk bedrag zal men dan de spreiding der gevulde flessen moeten verminderen?

Oplossing

Geef het gewicht van de olieïnhoud aan met $\underline{x}$, van de flessen met $\underline{y}$ en van de gevulde flessen met $\underline{z}$. Gegeven is dat $\underline{y}$ $N(228,3;11,3)$ verdeeld is en $\underline{z}$ $N(585,2;12,8)$ verdeeld. Nu is $\underline{z} = \underline{x}+\underline{y}$ en dus is, mede wegens de onafhankelijkheid van $\underline{x}$ en $\underline{y}$

$$\mathcal{E}_{\underline{z}} = \mathcal{E}_{\underline{x}} + \mathcal{E}_{\underline{y}}$$

en

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

a) Volgens stelling 2.5 is $\underline{x}$ normaal verdeeld; de verwachting is

$$\mathcal{E}_{\underline{x}} = 585,2 - 228,3 = 356,9$$

en de variantie

$$\sigma^2(x) = (12,8)^2 - (11,3)^2 = 36,15 = (6,0)^2 ;$$

$\underline{x}$ is dus $N(356,9;6,0)$ verdeeld.

b) De kans op een fles met minder dan 350 gram olie bedraagt

$$P[\underline{x} \leq 350] = P\left[\frac{\underline{x} - 356,9}{6,0} \leq \frac{350 - 356,9}{6,0}\right] = P[\underline{u} \leq -1,15] = 0,1251$$

en dus bevat 12,5% van de flessen minder dan 350 gram olie.

c) Uit (4;40) van hoofdstuk III volgt

$$\sigma^2(\underline{y}) = \sigma^2(\underline{z}) + \sigma^2(\underline{x}) - 2\,\mathrm{cov}(\underline{x},\underline{z})$$

en dus is

$$\mathrm{cov}(\underline{x},\underline{z}) = \tfrac{1}{2}\left[(12{,}8)^2 + 36{,}15 - (11{,}3)^2\right] = 36{,}15$$

d) Laat de nieuwe spreiding van $\underline{x}$ bedragen $\sigma'(\underline{x})$, dan is de kans op een fles met minder dan 350 gram olie

$$P[\underline{x} \leq 350] = P\left[\underline{u} \leq \frac{350 - 356{,}9}{\sigma'(\underline{x})}\right].$$

Deze kans mag hoogstens 0,025 bedragen en dus moet voldaan zijn aan

$$\frac{350 - 356{,}9}{\sigma'(\underline{x})} = -1{,}96,$$

zodat $\sigma'(\underline{x})$ hoogstens 3,5 mag zijn.
Hieruit volgt voor de variantie van $\underline{z}$

$$\sigma^2(\underline{z}) = (\sigma'(\underline{x}))^2 + \sigma^2(\underline{y}) = (3{,}5)^2 + (11{,}3)^2 = (11{,}8)^2$$

en de spreiding der gevulde flessen moet dus tot 11,8 gram teruggebracht worden.

Opmerking

Deze opgave geeft een typerend beeld van enigszins primitieve toepassingen van de statistiek, zoals deze in de praktijk vaak voorkomen. Het aantal metingen wordt zo groot verondersteld, dat het verschil tussen de uit deze metingen gevonden gemiddelden en de daardoor geschatte verwachtingen (eigenlijk: gemiddelden over "alle" flessen olie van die fabrikant) te verwaarlozen is. Hetzelfde geldt voor de spreidingen. De normaliteit van de kansverdelingen is een onderstelling, die onderzocht wordt door de vorm van de frequentieverdeling van de metingen na te gaan. Vertoont deze een "redelijke" gelijkenis met een normale verdelingsdichtheid, dan wordt de normaliteit aanvaard. De uitkomsten van dit soort berekeningen dienen steeds als slechts bij benadering juist te worden beschouwd.

3. De centrale limietstelling.

In de vorige paragraaf hebben wij gezien (vgl. stelling 2;5) dat een som van normaal verdeelde grootheden weer normaal verdeeld

is. Met behulp van deze stelling kunnen allerlei vraagstukken op eenvoudige wijze opgelost worden. Voor grootheden met een andere verdeling dan de normale gaat dit in het algemeen niet zo eenvoudig. Vaak kan men dan gebruik maken van de volgende uiterst belangrijke stelling, die bekend staat onder de naam: centrale limietstelling en welke hier geformuleerd wordt voor onderling onafhankelijke stochastische grootheden.

Stelling 3;1

Laat $\underline{x}_1, \underline{x}_2, \ldots$ een rij stochastisch onafhankelijke grootheden voorstellen. Laat μ_i en σ_i verwachting en spreiding van $\underline{x}_i$ $(i = 1,2,\ldots)$ zijn en laat verder

$$\delta_i^3 \stackrel{def}{=} \mathcal{E}|\underline{x}_i - \mu_i|^3 \qquad 1) \tag{3;1}$$

en

$$\begin{cases} \underline{x}(n) \stackrel{def}{=} \sum_{i=1}^{n} \underline{x}_i, \\ \mu(n) \stackrel{def}{=} \sum_{i=1}^{n} \mu_i, \\ \sigma^2(n) \stackrel{def}{=} \sum_{i=1}^{n} \sigma_i^2, \\ \delta^3(n) \stackrel{def}{=} \sum_{i=1}^{n} \delta_i^3 \end{cases} \tag{3;2}$$

Indien nu de voorwaarde

$$\lim_{n\to\infty} \frac{\delta(n)}{\sigma(n)} = 0 \tag{3;3}$$

vervuld is, dan is de grootheid

$$\underline{x}(n)^* \stackrel{def}{=} \frac{\underline{x}(n) - \mu(n)}{\sigma(n)} \tag{3;4}$$

voor $n \to \infty$ asymptotisch $N(0,1)$ verdeeld, d.w.z. dat voor iedere a geldt:

$$\lim_{n\to\infty} P[\underline{x}(n)^* \leq a] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}u^2} du \tag{3;5}$$

Opmerkingen

De grootheid $\underline{x}(n)^*$ is de gestandaardiseerde van $\underline{x}(n)$, want deze heeft $\mu(n)$ als verwachting en $\sigma(n)$ als spreiding en wij kunnen de inhoud van de stelling dus ook als volgt formuleren:

---

1) Dit is het z.g. absolute derde moment.

onder bepaalde, vrij algemeen geldende voorwaarden bezit de gestandaardiseerde som van een rij stochastische grootheden $\underline{x}_1, \underline{x}_2, \ldots$ asymptotisch een N(0,1) verdeling. Vaak is men minder exact en dan zegt men dat "$\underline{x}(n)$ asymptotisch normaal verdeeld" is waarbij dus nog geen standaardisering toegepast is).

Behalve in de hier gegeven vorm bestaan er nog verschillende andere vormen van de centrale limietstelling; de verschillen hebben dan betrekking op de voorwaarden, waaronder de asymptotische normaliteit geldt.

Als $\underline{x}_1, \underline{x}_2, \ldots$ alle dezelfde verdeling bezitten, dan kunnen de voorwaarden verzwakt worden en wel behoeft men niet te onderstellen dat het derde absolute moment bestaat. De stelling krijgt dan de volgende vorm.

Stelling 3;2 (Centrale limietstelling voor onderling onafhankelijke, identiek verdeelde stochastische grootheden)

Zijn $\underline{x}_1, \underline{x}_2, \ldots$ onderling onafhankelijk verdeeld volgens dezelfde verdeling met verwachting μ en spreiding σ, dan is

$$(3;6) \qquad \frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma\sqrt{n}} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i - \mu}{\sigma/\sqrt{n}}$$

voor $n \to \infty$ asymptotisch N(0,1) verdeeld.

Uit deze stelling volgt direct, dat wanneer $\underline{x}_1, \underline{x}_2, \ldots$ een steekproef voorstellen uit een willekeurige verdeling met verwachting μ en spreiding σ, dat dan het gemiddelde

$$(3;7) \qquad \bar{\underline{x}} = \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i$$

asymptotisch normaal verdeeld is voor $n \to \infty$ met verwachting μ en spreiding $\sigma/\sqrt{n}$.

In deze leergang houden wij ons slechts bezig met verdelingen, waarvan alle momenten eindig zijn, zodat wij de stellingen 3;1 en 3;2 steeds zonder meer kunnen toepassen.

Een belangrijke toepassing van stelling 3;2 komt nu aan de orde.

Stelling 3;3 (Stelling van De Moivre en Laplace)

Bezit $\underline{x}$ een binomiale verdeling met parameters n en p, d.w.z. is voor $k = 0,1,\ldots,n$

$$(3;8) \qquad P[\underline{x}=k] = \binom{n}{k} p^k q^{n-k} \qquad (q=1-p),$$

dan is $\underline{x}$ voor $n \to \infty$ asymptotisch normaal verdeeld. Dus voor iedere a geldt

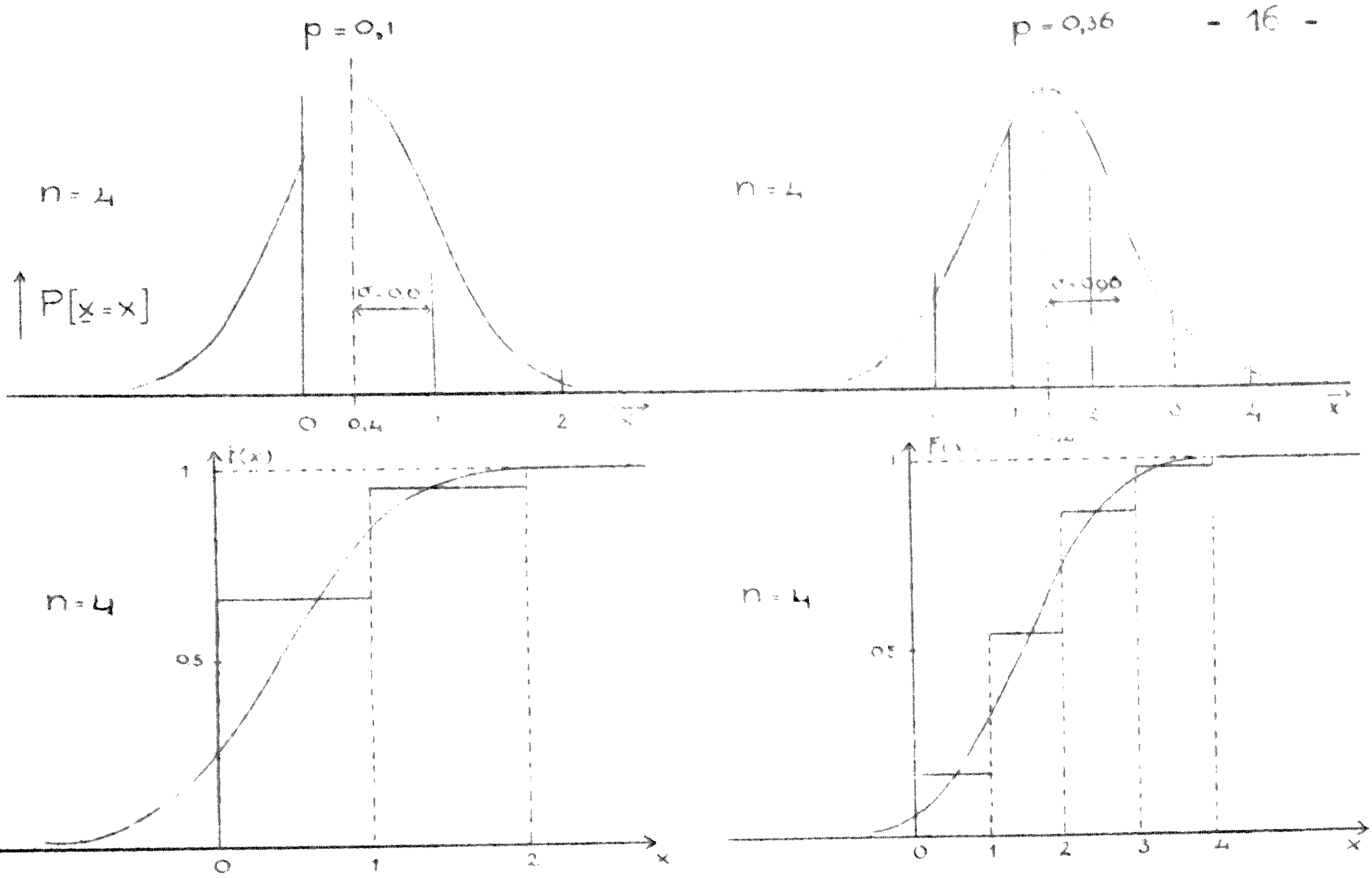
$$(3;9) \qquad \lim_{n\to\infty} P\left[\frac{\underline{x}-np}{\sqrt{npq}} \le a\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}u^2}\,du .$$

Bewijs

Op blz. 33 van hoofdstuk III hebben wij er reeds gebruik van gemaakt, dat $\underline{x}$ beschouwd kan worden als de som van n onafhankelijk verdeelde stochastische grootheden $\underline{x}_i$, die "het aantal successen bij het i^e experiment" voorstellen. Met behulp daarvan worden daar de formules $\mu(\underline{x}) = np$ en $\sigma^2(\underline{x}) = npq$ afgeleid. Daarmede is de stelling echter teruggebracht tot een speciaal geval van stelling 3;2, met $\mu = \mu(\underline{x}_i) = p$ en $\sigma^2 = \sigma^2(\underline{x}_i) = pq$

Op grond van deze stelling kan de normale verdeling gebruikt worden als benadering van de binomiale. Zijn de parameters van de binomiale verdeling n en p, dan is de verwachting np en de spreiding $\sqrt{npq}$. De normale verdeling met dezelfde verwachting en spreiding wordt nu de aangepaste normale verdeling genoemd.

In figuur 1;1 zijn voor p=0,1 en p=0,36 en voor n = 4,9,25,100 de kansen van de binomiale verdeling geschetst evenals de verdelingsdichtheid van de aangepaste normale verdeling. De x-schaal is daarbij steeds zo gekozen, dat de verdelingsdichtheid dezelfde vorm bezit (de spreidingen zijn dus in cm gelijk gemaakt). Verder zijn behalve voor n=100 ook de verdelingsfuncties getekend. Stelling 3;3 wordt geïllustreerd door het feit, dat de benadering duidelijk beter wordt bij toenemende n. Tevens blijkt uit de figuren, dat bij dezelfde waarde van n de benadering voor p=0,36 beter is dan voor p=0,1. De aanpassing is op zijn best voor $p=\frac{1}{2}$, hetgeen men reeds kan vermoeden op grond van het feit, dat de binomiale verdeling dan symmetrisch is, evenals de normale. Voor n=100 is het aantal stappen van de cumulatieve binomiale verdeling zo groot, dat de figuur op deze schaal niet meer getekend



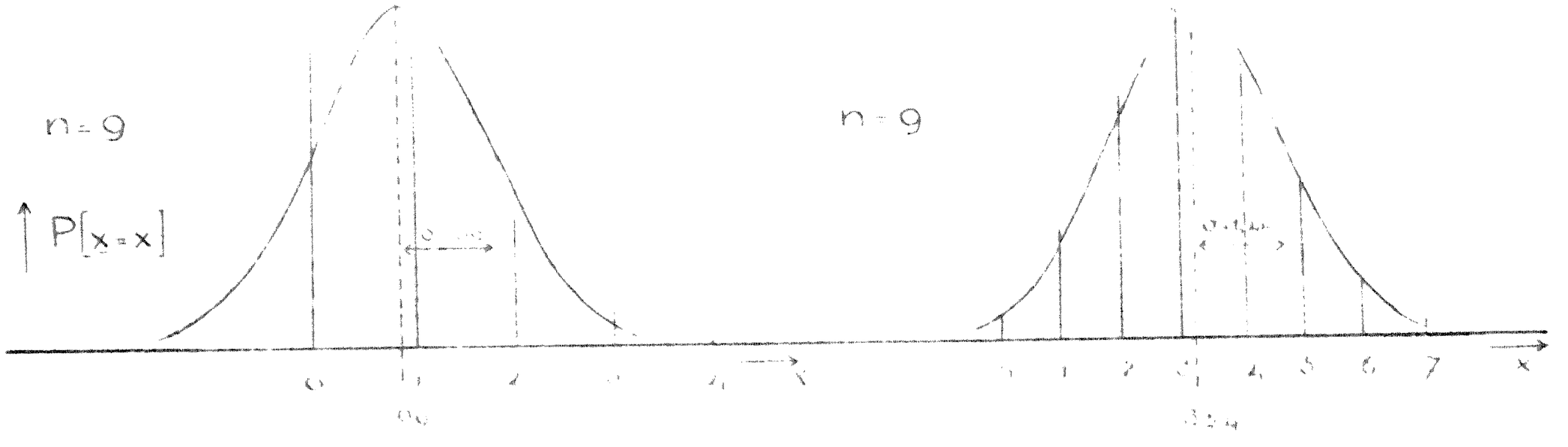


fig. 3.1

Convergentie van de binomiale verdeling naar de normale

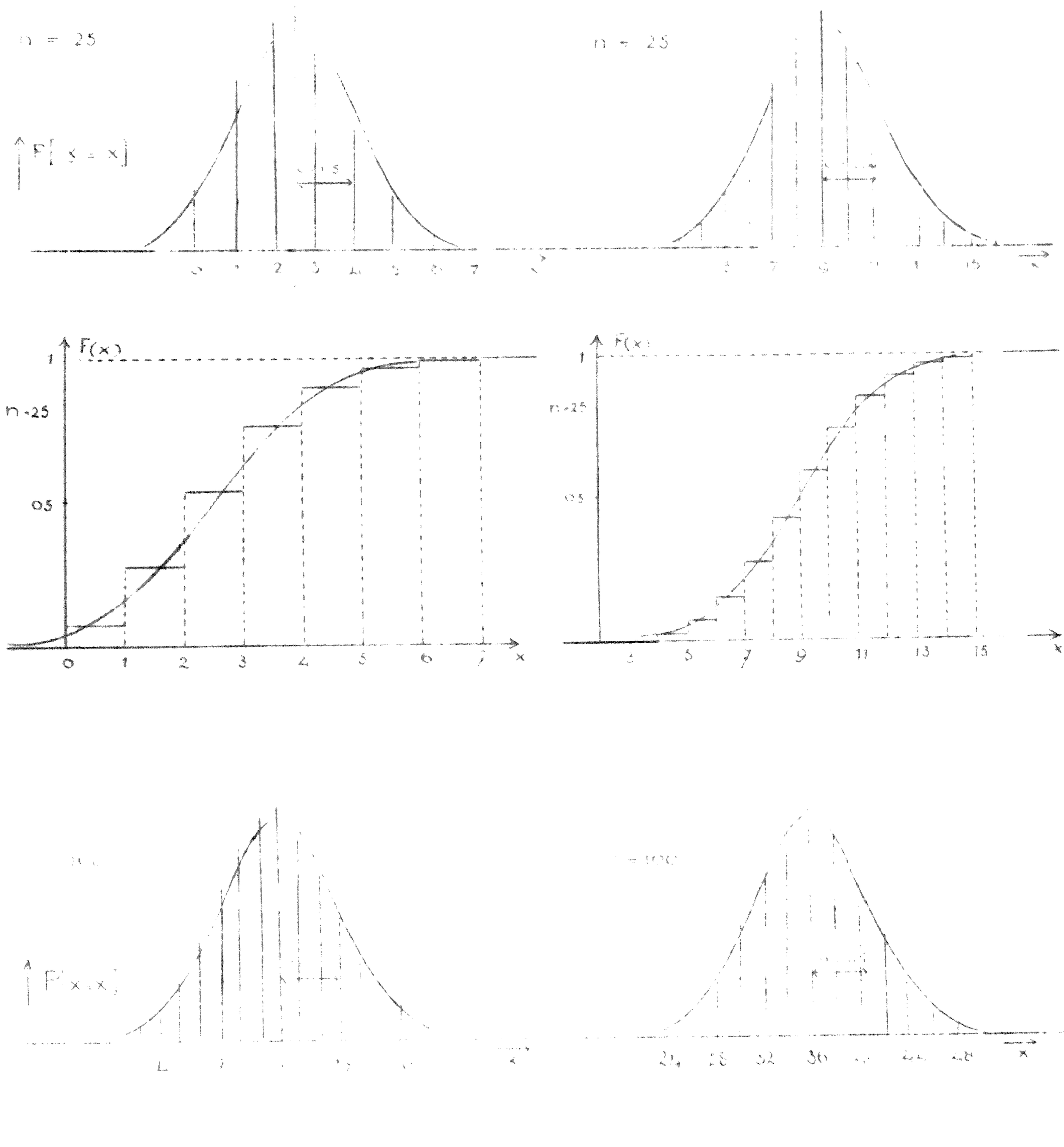


fig. 3.1 (vervolg)

Convergentie van de binomiale verdeling naar de normale

kan worden. Overigens blijkt reeds uit de figuur van de cumulatieve verdelingen bij p=0,36 en n=25 dat de aanpassing daar heel behoorlijk is; voor n=100 valt de trapfunctie in nog sterkere mate samen met de vloeiende lijn van de cumulatieve normale verdeling.

De stellingen 3;2 en 3;3 kunnen vaak in de praktijk gebruikt worden en daarom besluiten wij deze paragraaf met twee toepassingen.

Voorbeeld 3;1.

Van een verpakkingsmachine, die thee in pakjes van nominaal a gram verpakt is uit vroegere waarnemingen bekend dat de werkelijk verpakte hoeveelheid $\underline{x}$ een normale verdeling bezit met spreiding σ gram. De verwachting μ van de verdeling kan men regelen met een instellings-mechanisme.

Volgens de warenwet moet elk pakje minstens het nominale gewicht bevatten. Om te bereiken dat slechts een fractie α (met $\alpha < \frac{1}{2}$) van de pakjes minder dan a gram bevat zal men μ groter dan a moeten nemen.

Gevraagd wordt met hoeveel procent het overwicht μ-a verminderd kan worden indien de controle slechts zou eisen dat het gemiddelde van een aselecte steekproef van n pakjes minstens a gram moet bedragen, terwijl weer een kans α toegelaten wordt dat hieraan niet voldaan is?

Oplossing

Bij verwachting μ geldt

$$P[\underline{x} \leq a] = P\left[\frac{\underline{x}-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right] = P\left[\underline{u} \leq \frac{a-\mu}{\sigma}\right],$$

waarin $\underline{u}$ een N(0,1) verdeelde grootheid voorstelt. Deze kans moet gelijk zijn aan α. Als nu ξ_α gedefinieerd wordt door

$$P[\underline{u} \geq \xi_\alpha] = \alpha$$

dan moet dus gelden

$$\frac{a-\mu}{\sigma} = -\xi_\alpha \quad \text{of} \quad \mu = a + \xi_\alpha \sigma$$

Het overwicht moet dus $\xi_\alpha \sigma$ bedragen om ervoor te zorgen dat slechts een fractie α van de pakjes minder dan a gram bevat.

Stelt $\underline{\bar{x}}$ het gemiddelde gewicht van n aselect gekozen pakjes voor, dan volgt door toepassing van de stellingen 4;3 en 4;4 uit hoofdstuk III $\mathcal{E}\underline{\bar{x}} = \mathcal{E}\underline{x} = \mu$ en

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

Derhalve is

$$P[\underline{\bar{x}} \leq a] = P\left[\underline{u} \leq \frac{a-\mu}{\sigma}\sqrt{n}\right]$$

en deze kans is $=\alpha$ als

$$\frac{a-\mu}{\sigma}\sqrt{n} = -\xi_\alpha \quad \text{of} \quad \mu = a + \xi_\alpha \frac{\sigma}{\sqrt{n}}$$

Het overwicht moet nu dus $\xi_\alpha \frac{\sigma}{\sqrt{n}}$ bedragen. De besparing in overwicht is dus, in procenten van het oorspronkelijk overwicht

$$100\ \frac{\xi_\alpha \sigma - \xi_\alpha \frac{\sigma}{\sqrt{n}}}{\xi_\alpha \sigma} = 100\left(1 - \frac{1}{\sqrt{n}}\right)$$

Deze uitkomst is dus onafhankelijk van a, σ en α.

<u>Voorbeeld 3;2</u>.

In paragraaf 2 van hoofdstuk I hebben wij gezien, dat in 1954 in 6 wijken van Amsterdam tezamen op een totaal van 1927 geboorten 974 jongens voorkwamen. Wij vragen ons nu af, hoe groot de kans op een zo groot of nog groter aantal jongens zou zijn als bij ieder der geboorten de kans op een jongen gelijk is aan $\frac{1}{2}$.

Het aantal jongens, $\underline{x}$, heeft dan een binomiale verdeling met n=1927 en $p=\frac{1}{2}$, dus met

$$\mathcal{E}\underline{x} = np = 963{,}5 \quad \text{en} \quad \sigma(\underline{x}) = \sqrt{npq} = \tfrac{1}{2}\sqrt{1927} = 21{,}9$$

De gevraagde kans is

$$P\left[\underline{x} \geq 974 \mid n = 1927,\ p = \tfrac{1}{2}\right],$$

waarin nu achter de streep geen voorwaarden, maar gegevens staan aangegeven. Wij berekenen deze kans nu - bij benadering - met behulp van (3;9) op de volgende wijze (zonder de gegevens over n en p telkens weer te vermelden).

$$P\left[\underline{x} \geq 974\right] = P\left[\frac{\underline{x} - np}{\sqrt{npq}} \geq \frac{974 - np}{\sqrt{npq}}\right] \approx$$

$$\approx P\left[\underline{u} \geq \frac{974 - 963{,}5}{21{,}9}\right] = P\left[\underline{u} \geq 0{,}48\right].$$

waarin $\underline{u}$ een N(0,1) verdeelde grootheid voorstelt en $\approx$ betekent, dat wij met een benadering te maken hebben.

Volgens de tabel van de N(0,1) verdeling is de kans in het

laatste lid gelijk aan 0,32. Onze uitkomst is dus

$$P[x \geq 974 \mid n = 1927,\ p=\tfrac{1}{2}] \approx 0{,}32.$$

Het belang van kansen van dit type, die <u>overschrijdingskansen</u> genoemd worden, zullen wij later leren kennen.