STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Rapport S 265 (C 13)

Leergang Besliskunde

Hoofdstuk V

Schattingstheorie

door

J. Kriens

(Bewerking van hoofdstuk 12 van rapport S 200(C8)
door Prof. Dr J. Hemelrijk)

April 1960

1. Inleiding

De schattingstheorie is een belangrijk onderdeel van de statistiek, waarvan wij hier wegens de grote uitgebreidheid van het onderwerp slechts een klein deel kunnen behandelen. Wij geven eerst een aantal voorbeelden met een intuïtief voor de hand liggende oplossing en behandelen daarna het principe van een algemene theorie, waarop deze oplossingen gebaseerd kunnen worden.

Voorbeeld 1;1. Onderzoek van een eigenschap van een serieprodukt

Eén der belangrijkste eigenschappen van veel produkten, zoals gloeilampen, onderdelen van machines en apparaten, is de levensduur. Het is dan ook van groot belang te beschikken over een methode, waarmee een goed inzicht in deze eigenschap verkregen kan worden. In dit voorbeeld zullen wij een methode beschrijven, geïllustreerd aan de hand van gloeilampen.

De levensduur van een gloeilamp kan men bepalen door hem te laten branden tot hij doorbrandt. Uiteraard is deze methode ongeschikt wanneer men een nog in de handel te brengen partij wil onderzoeken en daarom is men in dit geval gedwongen de eigenschappen van een bepaald soort gloeilampen met behulp van steekproeven te onderzoeken.

Wij beschouwen de levensduur als een stochastische variabele $\underline{x}$, die voor iedere lamp een bepaalde waarde aanneemt. De bij een steekproef van lampen gevonden waarden vormen dan een reeks van stochastisch onafhankelijke waarnemingen van $\underline{x}$. Veelal wordt bij dergelijke problemen verondersteld, dat $\underline{x}$ een (bij benadering) normale verdeling bezit. Soms geldt dit niet voor $\underline{x}$ zelf, maar wel voor log $\underline{x}$. In hoofdstuk VI, paragraaf 4, zullen wij ingaan op de vraag hoe een dergelijke onderstelling op grond van waarnemingen onderzocht kan worden. Hier hanteren wij deze alsof zij reeds onderzocht en bevestigd is.

Is nu $\underline{x}$ normaal verdeeld, dan is de verdeling volledig vastgelegd, wanneer de parameters $\mu = \mathcal{E}\underline{x}$ en $\sigma^2 = \sigma^2(\underline{x})$ bekend zijn en het schattingsprobleem is dus deze twee parameters uit een reeks waarnemingen te schatten.

Laat de steekproef van waarnemingen voorgesteld worden door de onafhankelijke grootheden

(1;1) $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n,$

alle met dezelfde verdeling als $\underline{x}$. In principe kan nu iedere functie $\underline{t}$ van $\underline{x}_1, \ldots, \underline{x}_n$ gekozen worden om een onbekende parameter θ te schatten. Van deze vrijheid maakt men gebruik door te eisen dat de schatting $\underline{t}$ één of meer gunstige eigenschappen bezit. Als zodanig geeft men veelal de voorkeur aan schattingen die zuiver zijn en asymptotisch raak, begrippen waarvan wij nu eerst de definities laten volgen. In deze definities schrijven wij in plaats van $\underline{t}$ het symbool $\underline{t}_n$ om duidelijk uit te laten komen dat de schatting gebaseerd is op n waarnemingen.

Definitie 1;1

Een schatting $\underline{t}_n$ van een onbekende parameter θ wordt zuiver genoemd als zijn verwachting gelijk is aan deze onbekende parameter

(1;2) $$\mathcal{E}\underline{t}_n = \theta \ .$$

Definitie 1;2

Een schatting $\underline{t}_n$ van een onbekende parameter θ wordt asymptotisch raak genoemd, als hij voor een onbegrensd stijgend aantal waarnemingen stochastisch convergeert naar die onbekende parameter:

(1;3) $$\lim_{n\to\infty} P\left[\,|\underline{t}_n - \theta| > \varepsilon\,\right] = 0 \qquad \text{voor iedere } \varepsilon > 0.$$

Het begrip zuiver hangt dus niet samen met de steekproefomvang en betekent ruw geformuleerd dat de schatting gemiddeld goed is. Asymptotisch raak is een eigenschap die wel afhangt van de steekproefomvang.

Wij trachten nu voor beide onbekende parameters van de normale verdeling schattingen te vinden, die zowel zuiver zijn als asymptotisch raak.

Het ligt voor de hand als schatting voor het gemiddelde μ te nemen het steekproefgemiddelde, dus

(1;4) $$\overline{\underline{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i \ .$$

Deze schatting is inderdaad zuiver, zoals wij reeds gezien hebben in voorbeeld 3;1 van hoofdstuk IV. Bovendien volgt uit stelling 1;1 van hetzelfde hoofdstuk dat $\underline{\bar{x}}$ een asymptotisch rake schatting is van μ.

Als <u>schatting voor de variantie σ^2</u> onderzoeken wij de grootheid

$$(1;5) \qquad \underline{s}'^2 = \frac{1}{n}\sum_{i=1}^{n}(\underline{x}_i-\underline{\bar{x}})^2$$

op zijn merites.

De verwachting van $\underline{s}'^2$ berekenen wij als volgt.

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

$$= \sum_{i=1}^{n}(\underline{x}_i-\mu)^2-2(\underline{\bar{x}}-\mu)\sum_{i=1}^{n}(\underline{x}_i-\mu)+n(\underline{\bar{x}}-\mu)^2.$$

Hierin is

$$\sum_{i=1}^{n}(\underline{x}_i-\mu) = n(\underline{\bar{x}}-\mu),$$

zodat

$$\sum_{i=1}^{n}(\underline{x}_i-\underline{\bar{x}})^2 = \sum_{i=1}^{n}(\underline{x}_i-\mu)^2-n(\underline{\bar{x}}-\mu)^2.$$

Daar

$$\mathcal{E}(\underline{x}_i-\mu)^2 = \sigma^2 \quad \text{en} \quad \mathcal{E}(\underline{\bar{x}}-\mu)^2 = \sigma^2(\underline{\bar{x}}) = \frac{\sigma^2}{n},$$

geldt

$$(1;6) \qquad \mathcal{E}\underline{s}'^2 = \frac{n-1}{n}\sigma^2.$$

De schatting $\underline{s}'^2$ is dus geen zuivere schatting van σ^2, maar

$$(1;7) \qquad \underline{s}^2 \overset{\text{def}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(\underline{x}_i-\underline{\bar{x}})^2$$

is dat wel. Voor praktische toepassingen wordt daarom veelal $\underline{s}^2$ als schatting van σ^2 gebruikt.

Opmerking 1;1

Het argument, dat $\underline{s}^2$ een zuivere schatting van σ^2 is en $\underline{s}'^2$ niet, verliest zijn waarde, indien men niet σ^2 maar σ wenst te schatten, hetgeen in de praktijk gewoonlijk het geval is. Immers beschouwt men $\underline{s}$ als schattint van σ, dan is $\underline{s}$ weer onzuiver. Men kan dit inzien door in (4,26) uit hoofdstuk III voor $\underline{x}$ te substitueren $\underline{s}$, waardoor men vindt

$$\mathcal{E}\underline{s}^2 > (\mathcal{E}\underline{s})^2,$$

dus

$$\sigma^2 > (\mathcal{E}\underline{s})^2 \quad \text{of} \quad \sigma > \mathcal{E}\underline{s},$$

zodat de verwachting van $\underline{s}$ kleiner is dan de te schatten parameter σ.

Het gebruik van $\underline{s}^2$ in plaats van $\underline{s}'^2$ berust dan ook in feite op traditie. Het bovenstaande geldt nog steeds zonder de veronderstelling van normaliteit. Voor normaal verdeelde $\underline{x}$ kan de factor

$$c_n \overset{\text{def}}{=} \frac{\sigma}{\mathcal{E}\underline{s}}$$

berekend worden. De schatting $c_n \underline{s}$ is dan een zuivere schatting voor σ.

Zowel $\underline{s}$ als $\underline{s}'$ zijn asymptotisch raak; voor $n \to \infty$ zijn zij aequivalent, daar dan $\frac{n-1}{n} \to 1$. Beschouw nu

$$\underline{s}'^2 = \frac{1}{n} \sum_{i=1}^{n} (\underline{x}_i - \mu)^2 - (\overline{\underline{x}} - \mu)^2.$$

Volgens stelling 1;1 uit hoofdstuk IV convergeert $\overline{\underline{x}}$ stochastisch naar μ, dus $|\overline{\underline{x}} - \mu|$ naar 0, dus ook $(\overline{\underline{x}} - \mu)^2$ naar 0. Deze term kan dus verwaarloosd worden: behoudens een willekeurig kleine kans wordt hij zelf willekeurig klein. Noemen wij verder

$$\left.\begin{aligned} \underline{y}_i &\overset{\text{def}}{=} (\underline{x}_i - \mu)^2, \\ \mathcal{E}\underline{y}_i &= \sigma^2 \end{aligned}\right\} \qquad (i=1,2,\ldots,n)$$

en, wederom volgens de reeds aangehaalde stelling convergeert nu

$$\frac{1}{n}\sum_{i=1}^{n} \underline{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\underline{x}_i - \mu)^2$$

stochastisch naar de verwachting, dus naar σ^2.[1] Daarmede is echter het gestelde (zij het niet volledig exact) bewezen.

Wij zijn er dus in geslaagd de boven gestelde opgave, schattingen te vinden voor μ en σ^2, die zowel zuiver zijn als asymptotisch raak, op te lossen. Terloops merken wij op dat het in veel gevallen niet mogelijk is, schattingen voor onbekende parameters van kansverdelingen te vinden, die èn zuiver èn asymptotisch raak zijn.

Voorbeeld 1,2

Van 10 lampen van een bepaald type, aselect gekozen uit de produktie, is de levensduur (in uren) gemeten. De uitkomsten zijn vermeld in de eerste kolom van tabel 1;I.[2]

Tabel 1;I

Berekening van gemiddelde en variantie van een reeks waarnemingen

x_i	$x_i - 1000$	$(x_i - 1000)^2$
982	- 18	324
996	- 4	16
1047	47	2209
980	- 20	400
1057	57	3249
1025	25	625
990	- 10	100
1058	58	3364
994	- 6	36
1036	36	1296
totaal	165	11619

Om de berekening te vereenvoudigen kiezen wij eerst een z.g. "voorlopig gemiddelde", d.i. een getal x_o (het doet er niet veel

---

1) Volgens de genoemde stelling moet de voorwaarde vervuld zijn dat $\sigma^2(\underline{y}_i)$ eindig is; deze voorwaarde kan echter vermeden worden.

2) Deze gegevens zijn niet aan de praktijk ontleend.

toe welk precies), dat vermoedelijk in de buurt van het werkelijke gemiddelde $\bar{x}$ ligt. In dit geval ligt het voor de hand x_o=1000 te nemen. In de tweede kolom staat x_i-1000 berekend, in de derde $(x_i-1000)^2$. Nu is

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (x_i - x_o) + n\, x_o,$$

dus

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

In ons geval

$$\bar{x} = \frac{1}{10} \cdot 165 + 1000 = 1016{,}5.$$

Verder is (vgl. blz. 3)

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

In ons geval is dus

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 11619 - 10(16{,}5)^2 = 8896{,}5 .$$

Derhalve is

$$s^2 = \frac{1}{9} \cdot 8896{,}5 = 988{,}5, \quad s = 31{,}4 .$$

Willen wij nu bijv. schatten, hoe groot het percentage lampen is, dat een levensduur < 950 bezit, dan berekenen wij

$$\frac{1016{,}5 - 950}{31{,}4} = 2{,}12$$

en zoeken dit getal op in de tabel van de N(0,1)-verdeling. De uitkomst is 0,017, dus 1,7%. Van de steekproef is dan ook geen enkele uitkomst zo laag.

Opmerking 1;2

De hier aangegeven methode om het gemiddelde te bepalen vormt alleen een vereenvoudiging, wanneer men de berekeningen met potlood en papier uitvoert. Wanneer men een tafelrekenmachine gebruikt, kan men het beste x_o=0 kiezen en dus uitgaan van de

formule

$$(1;8) \qquad \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \quad .$$

Men berekent dan de som van kwadraten $\sum_{i=1}^{n} x_i^2$ en kan er profijt van trekken, dat de rekenmachine de som $\sum_{i=1}^{n} x_i$ tegelijkertijd noteert.

Voorbeeld 1,3. Kurken tellen

Een aardig historisch voorbeeld van een eenvoudig schattingsprobleem is het volgende.[1] In een kurkenfabriek, die millioenen kurken in voorraad had, waarvan dagelijks duizenden in kleinere en grotere partijen moesten worden afgeleverd, waren voortdurend een aantal arbeiders bezig met de hand kurken te tellen. Dit was een vermoeiende en zenuwslopende bezigheid, waarmee het bovendien nog niet mogelijk was aan het einde van het jaar de voorraad op te nemen. Men wenste daarom een kurkentel-apparaat te ontwikkelen

Aangezien het bij aflevering van partijen kurken niet op een paar meer of minder aankomt, kon Ir VAN ETTINGER de oplossing in statistische richting zoeken. Zijn methode komt hierop neer, dat uit een zak kurken van één maat een steekproef genomen wordt, waarvan het totale gewicht (g) en het aantal (n) bepaald wordt. Het gemiddelde gewicht van één kurk is dan dus $\frac{g}{n}$. Is nu het gewicht van de gehele zak kurken gelijk aan G, dan is

$$\frac{nG}{g}$$

een schatting van het aantal kurken uit de zak.

Deze methode werd bovendien als volgt gemechaniseerd. Een weegschaal werd vervaardigd, die in evenwicht is als het gewicht aan kurken op de éne schaal 99 maal zo groot is als dat op de andere. Men zet nu de zak kurken op de ene schaal en legt daaruit zoveel kurken op de andere, dat de weegschaal in evenwicht is. Het totale aantal kurken is dan, bij benadering, gelijk aan 100 maal het aantal op de tweede schaal overgebrachte kurken. De tijds-

---

1) J. VAN ETTINGER, Statistiek en produktiviteit, Statistica 6 (1952), pp. 19-28, pag. 20.

besparing is enorm, de investering gering en de arbeiders zijn trots op de eenvoudige wijze waarop zij kurken tellen. Het systeem is, in die fabriek, reeds bijna 30 jaar in gebruik.

Voorbeeld 1;4. Enquête

Wij beschouwen vervolgens een enquête, die erop gericht is de kennis van een bepaalde groep mensen te onderzoeken. Voor de eenvoud vatten wij één vraag in het oog, die juist of onjuist beantwoord kan worden. Deze vraag wordt voorgelegd aan een steekproef van n personen, aselect genomen uit de te onderzoeken groep. In statistisch jargon wordt deze te onderzoeken groep personen de populatie genoemd. Als model nemen wij aan, dat een fractie p van de personen uit deze groep de vraag goed zou beantwoorden en de overige (fractie q=1-p) fout. Het gaat er nu om deze fracties te schatten op grond van de antwoorden van de n personen van de steekproef.

Kiezen wij aselect één persoon uit de populatie, dan is de kans, dat wij een juist antwoord verkrijgen gelijk aan p. Bestaat de populatie uit N personen, dan bevinden zich daaronder dus Np, die een juist antwoord zouden geven en Nq, die de vraag onjuist zouden beantwoorden. Hieruit zijn er nu n aselect en zonder teruglegging genomen en wij bevinden ons dus in de situatie van voorbeeld 5;7 van hoofdstuk I. Het aantal juiste antwoorden in de steekproef van omvang n, dat wij met $\underline{k}$ aangeven, heeft dus een hypergeometrische verdeling.

$$(1;9) \qquad P[\underline{x}=k] = \frac{\binom{Np}{k}\binom{Nq}{n-k}}{\binom{N}{n}} \quad .$$

Volgens tabel 4;I uit hoofdstuk III is de verwachting van $\underline{x}$, wanneer wij de hier gebruikte symbolen substitueren

$$(1;10) \qquad \mathcal{E}k = np.$$

Dus

$$(1;11) \qquad \mathcal{E}\frac{k}{n} = p.$$

De fractie juiste antwoorden in de steekproef is dus een zuivere schatting voor de onbekende fractie p. De vraag naar het

asymptotisch gedrag doet zich hier niet op natuurlijke wijze voor, daar $n > N$ niet op kan treden. Voor $n=N$ vindt men p reeds exact.

Wat wel kan gebeuren is, dat N niet bekend is. Volgens (1;11) kan $\frac{x}{n}$ dan toch als schatting gebruikt worden. Is $N >> n$, dan is volgens (5;14) uit hoofdstuk I de kansverdeling van $\underline{x}$ bij benadering binomiaal:

$$(1;12) \qquad P[\underline{x}=k] \approx \binom{n}{k} p^k q^{n-k} ,$$

waarbij dan (1;10) geldig blijft. Aan de schatting verandert dus niets. Volgens stelling 1,2 uit hoofdstuk IV (de theoretische wet van de grote aantallen) is de schatting in het geval van een binomiale verdeling asymptotisch raak. (In het hier beschouwde geval impliceert $n \to \infty$ dan ook $N \to \infty$; is p bijv. de onbekende kans op succes bij een experiment, dat willekeurig vaak herhaald kan worden, dan doet deze complicatie zich niet voor.)

Hiermee besluiten wij de voorbeelden, waarin goede schattingen van de onbekende parameters alle min of meer voor de hand lagen. Naast heuristische redeneringen om schattingen te vinden, bestaan er ook verschillende algemene theorieën om schattingen af te leiden. De belangrijkste is de theorie van de meest aannemelijke schattingen, waarvan de grondslag gelegd is door R.A. FISHER; enkele elementen van deze theorie en voorbeelden behandelen wij in de volgende paragraaf. Op een andere methode wordt in paragraaf 4 in het kort ingegaan.

2. Meest aannemelijke schattingen

Het principe van de theorie zetten wij uiteen voor continue verdelingen. Voor discrete is het analoog, met kansen in plaats van verdelingsdichtheden.

Laat

$$(2;1) \qquad \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$$

n (al of niet onafhankelijke) waarnemingen voorstellen, die een simultane verdeling bezitten, met verdelingsdichtheid

$$(2;2) \qquad f(x_1, x_2, \ldots, x_n | \theta_1, \theta_2, \ldots, \theta_h),$$

waarin $\theta_1,\theta_2,\ldots,\theta_h$ onbekende parameters voorstellen, die uit de waarnemingen geschat moeten worden.

Stellen nu $x_1,x_2,\ldots,x_n$ de waarden voor, die in werkelijkheid gevonden zijn, zodat het dus getallen zijn, geen variabelen meer, dan is (2;2) een functie van de onbekende parameters $\theta_j(j=1,2,\ldots,h)$.

Deze functie wordt de aannemelijkheidsfunctie of kortweg aannemelijkheid genoemd en aangegeven als

$$(2;3) \qquad L(\theta_1,\theta_2,\ldots,\theta_h) \stackrel{\text{def}}{=} f(x_1,x_2,\ldots,x_n|\,\theta_1,\theta_2,\ldots,\theta_h).$$

Gewoonlijk (doch niet voor alle verdelingsdichtheden f) bezit L een maximum; het punt, waarin dit wordt aangenomen, wordt aangegeven met $\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_h$ en deze waarden - die functies van de $x_i(i=1,2,\ldots,n)$ zijn - worden de meest aannemelijke schattingen van $\theta_1,\theta_2,\ldots,\theta_h$ genoemd.

Toelichting aan voorbeelden

Voorbeeld 2;1. Normale verdeling (zie ook voorbeeld 1;1)

Laat $\underline{x}$ een normale verdeling bezitten met onbekend gemiddelde μ en onbekende variantie σ^2 en laten $\underline{x}_1,\ldots,\underline{x}_n$ n onderling onafhankelijke waarnemingen van deze grootheid zijn. Wegens de onafhankelijkheid kunnen wij stelling 3;3 uit hoofdstuk III op (2;3) toepassen, zodat wij vinden met $\theta_1 = \mu$ en $\theta_2 = \sigma^2$

$$(2;4) \qquad L(\mu,\sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} =$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}}.$$

In deze vorm beschouwen wij σ^2 (en niet σ) als onbekende parameter, daar dit de berekeningen vereenvoudigt. Om het maximum van L te vinden, bij gegeven waarden der x_i, differentiëren wij, ter verdere vereenvoudiging, niet L zelf, maar de logarithme van L naar μ en naar σ^2.

$$(2;5) \qquad \log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \tfrac{1}{2}\,\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\,,$$

$$(2;6) \qquad \frac{\partial \log L}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}\,,$$

$$(2;7) \qquad \frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \tfrac{1}{2}\,\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^4}\,.$$

Stellen wij deze afgeleiden gelijk aan 0, dan verkrijgen wij dus 2 vergelijkingen voor $\hat{\mu}$ en $\hat{\sigma}^2$, en wel

$$(2;8) \qquad \sum_{i=1}^{n}(x_i-\hat{\mu}) = 0$$

en

$$(2;9) \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2.$$

Uit de eerste volgt

$$(2;10) \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x} \qquad \text{(vgl. (1;4))}$$

en vervolgens uit de tweede

$$(2;11) \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = s'^2 \quad \text{(vgl. (1;5))}.$$

De meest aannemelijke schattingen komen dus in dit geval overeen met de boven op intuïtieve wijze gekozene.

Opmerkingen

2;1. Indien wij niet σ^2, maar σ zelf als onbekende parameter gekozen hadden, zou dezelfde uitkomst verkregen zijn. Dit is een speciaal geval van een algemene stelling, inhoudende dat de meest aannemelijke schatting van een continue een-eenduidige functie van een onbekende parameter verkregen wordt door dezelfde functie te nemen van de meest aannemelijke schatting van die parameter. Dus: is $\hat{\theta}$ de meest aannemelijke schatting van θ, dan

is $Q(\hat{\theta})$ de meest aannemelijke schatting van $Q(\theta)$ als Q een continue en ondubbelzinnig omkeerbare functie is.

2;2. Wij zien tevens, dat meest aannemelijke schattingen niet zuiver behoeven te zijn. Dit volgt trouwens reeds uit opmerking 2,1, daar zuiverheid, zoals wij in opmerking 1;1 gezien hebben, bij transformatie van de parameter verloren kan gaan.

Voorbeeld 2;2. Binomiale verdeling

Is $\underline{x}$ binomiaal verdeeld met onbekende p, maar bekende n, dan geldt dus

$$L(p) = P[\underline{x}=k|p] = \binom{n}{k}p^k q^{n-k} \quad . \tag{2;12}$$

Dus

$$\log L(p) = \log \binom{n}{k} + k \log p + (n-k) \log (1-p). \tag{2;13}$$

Differentiëren naar p geeft

$$\frac{d \log L}{dp} = \frac{k}{p} - \frac{n-k}{1-p} \quad , \tag{2;14}$$

zodat wij, om $\hat{p}$ te vinden, de vergelijking

$$\frac{k}{\hat{p}} = \frac{n-k}{1-\hat{p}} \tag{2;15}$$

moeten oplossen. Dit geeft

$$\hat{p} = \frac{k}{n} \quad . \tag{2;16}$$

Opmerking 2;3

Ook hier vinden wij dus de vroeger reeds vermelde schatting. Voor de hypergeometrische verdeling is dit niet het geval; daarbij treden complicaties op, o.a. doordat Np in dat geval geheel moet zijn, zodat p niet alle waarden tussen 0 en 1 kan bezitten. Wij zullen dit geval hier buiten beschouwing laten.

Voorbeeld 2;3. Hoogwaterstanden; exponentiële verdeling

Indien men de frequentieverdeling van hoogwaterstanden waargenomen in Hoek van Holland, beschouwt, blijkt dat deze, voor zoverre het niet te kleine waarden betreft, bij goede benadering exponentieel verdeeld zijn. Dit betekent, dat voor iedere niet te lage waarde x_o geldt

$$(2;17)\quad P\left[\underline{x} \geqq x \,\middle|\, \underline{x} \geqq x_o\right] = \frac{P\left[\underline{x} \geqq x \wedge \underline{x} \geqq x_o\right]}{P\left[\underline{x} \geqq x_o\right]} = \frac{P\left[\underline{x} \geqq x\right]}{P\left[\underline{x} \geqq x_o\right]} = e^{-\lambda(x-x_o)} \quad (x > x_o)$$ [1)]

Geldt dit voor één waarde x_o, dan ook voor iedere waarde $x_o' > x_o$. Stellen wij namelijk

$$(2;18)\qquad P\left[\underline{x} \geqq x_o\right] = P_o,$$

dan is

$$(2;19)\qquad P\left[\underline{x} \geqq x\right] = P_o\, e^{-\lambda(x-x_o)},$$

zodat voor een willekeurige waarde $x_o' > x_o$ geldt

$$(2;20)\qquad P\left[\underline{x} \geqq x_o'\right] = P_o\, e^{-\lambda(x_o'-x_o)}.$$

Voor $x \geqq x_o'$ is nu

$$P\left[\underline{x} \geqq x \,\middle|\, \underline{x} \geqq x_o'\right] = \frac{P\left[\underline{x} \geqq x \wedge \underline{x} \geqq x_o'\right]}{P\left[\underline{x} \geqq x_o'\right]}.$$

Hierin is de teller weer gelijk aan $P\left[\underline{x} \geqq x\right]$ en voor de noemer geldt (2;20). Dus is

$$(2;21)\quad P\left[\underline{x} \geqq x \,\middle|\, \underline{x} \geqq x_o'\right] = \frac{P_o . e^{-\lambda(x-x_o)}}{P_o . e^{-\lambda(x_o'-x_o)}} = e^{-\lambda(x-x_o')} \quad (x \geqq x_o').$$

Dit betekent, dat de keuze van het "beginpunt" x_o (of x_o') de vorm van de verdeling - en in het bijzonder de waarde van λ - niet beïnvloedt, hetgeen het mogelijk maakt de lage waarden van x, waar de verdeling niet bekend is, uit te schakelen zonder dat de betrekkelijk willekeurige keuze van het beginpunt veel invloed op de uitkomst uitoefent.

Deze eigenschap van de exponentiële verdeling, dat de voor-

---

1) P.J. WEMELSFELDER, Wetmatigheden in het optreden van stormvloeden, "De Ingenieur", 3 maart 1939.

waardelijke verdeling onder de voorwaarde $\underline{x} \geqq x_o$ niet van de waarde van x_o afhangt is daarom voor het statistische onderzoek van de hoogwaterstanden aan de Nederlandse kust - en daarmede voor de bepaling van dijkhoogten, die een voldoende bescherming tegen overstromingen bieden - van veel belang. Een verklaring voor het experimenteel vastgestelde feit, dat de verdeling der hoogwaterstanden exponentieel is, is (nog?) niet bekend. In voorbeeld 4;1 van hoofdstuk VI zal een methode worden aangegeven waarmee kan worden vastgesteld of de verdeling inderdaad bij goede benadering exponentieel is of niet.

De enige onbekende parameter in de verdelingsfunctie is λ en daarvan berekenen wij nu de meest aannemelijke schatting.

De verdelingsdichtheid, steeds onder de voorwaarde $\underline{x} \geqq x_o$, is

$$f(x) = F'(x) = \frac{d(1-P)}{dx} = \lambda e^{-\lambda(x-x_o)} ,$$

dus als $x_1, x_2, \ldots, x_n$ de waarnemingen zijn, is

$$(2;22) \qquad L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda(x_i-x_o)} = \lambda^n \, e^{-\lambda \sum_{i=1}^{n}(x_i-x_o)} .$$

Verder is

$$(2;23) \qquad \log L = n \log \lambda - \lambda \sum_{i=1}^{n}(x_i-x_o)$$

en

$$(2;24) \qquad \frac{d \log L}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n}(x_i-x_o).$$

Derhalve is

$$(2;25) \qquad \hat{\lambda} = \frac{n}{\sum_{i=1}^{n}(x_i-x_o)} = \frac{1}{\bar{x}-x_o} ,$$

dus $\hat{\lambda}^{-1}$ is het gemiddelde der waarnemingen boven x_o verminderd met x_o.

Men kan laten zien, dat voor deze schatting geldt

$$(2;26) \qquad \mathcal{E}\left(\frac{1}{\underline{\bar{x}}-x_o} \,\middle|\, \underline{x} \geqq x_o\right) = \frac{n}{n-1} \lambda .$$

De meest aannemelijke schatting is dus niet zuiver, terwijl

$$(2;27) \qquad \frac{n-1}{n} \cdot \frac{1}{\underline{\bar{x}}-x_o} = \frac{n-1}{\sum_{i=1}^{n}(\underline{x}_i-x_o)}$$

wel een zuivere schatting van λ is.

Zowel de schatting (2;25) als de schatting (2;27) is asymptotisch raak. Uit stelling 1;1 van hoofdstuk IV volgt namelijk, dat $\underline{\bar{x}}-x_o$ stochastisch convergeert naar $\mathcal{E}(\underline{x}-x_o|\underline{x}\geq x_o) = \frac{1}{\lambda}$

en dus convergeert $\underline{\hat{\lambda}} = \frac{1}{\underline{\bar{x}}-x_o}$ stochastisch naar λ.

Opmerking 2;4

Bij het werkelijke onderzoek van deze waterstanden werden niet alle hoogwaterstanden gebruikt, doch slechts een in overleg met het K.N.M.I. op grond van meteorologische overwegingen uitgekozen groep, behorende bij bepaalde depressies van een potentieel gevaarlijk karakter.

Voorbeeld 2;4. Regressie: theorie der kleinste kwadraten

Wij beschouwen, als laatste voorbeeld, een stochastische grootheid $\underline{y}$, waarvan de verdeling afhangt van een andere, niet stochastische grootheid x. Zo kan $\underline{y}$ bijv. de in een week door een koe geleverde hoeveelheid melk voorstellen (die van week tot week en van koe tot koe verschilt) en x de leeftijd (in jaren) van de koe. Of $\underline{y}$ kan de lengte van de remweg van een auto voorstellen en x de snelheid bij het begin van het remmen. In deze beide gevallen kan men x zelf kiezen - in het eerste geval door een koe van de gewenste leeftijd uit te zoeken, in het tweede door bij de gewenste snelheid te gaan remmen -, terwijl $\underline{y}$ waargenomen wordt.

In beide gevallen gaat het erom te onderzoeken hoe de kansverdeling van $\underline{y}$ van x afhangt.

Wij beschouwen hier alleen het eenvoudigste geval, nl. dat alleen de verwachting van $\underline{y}$ van x afhankelijk is en wel lineair:

$$(2;28) \qquad \mathcal{E}(\underline{y}|x) = \alpha + \beta x,$$

waarin α en β onbekende parameters zijn, die geschat moeten worden. Verder veronderstellen wij, dat de verdeling van $\underline{y}$, bij gegeven x, normaal verdeeld is met een onbekende spreiding σ, die niet van x afhangt.

Op dit probleem gaan wij nu de methode der meest aannemelijke schattingen toepassen. De waarnemingen, die verondersteld worden stochastisch onafhankelijk te zijn, zijn nu van de vorm

$$(2;29) \qquad (x_1,\underline{y}_1),(x_2,\underline{y}_2),\ldots,(x_n,\underline{y}_n),$$

daar iedere waar te nemen waarde $\underline{y}_i$ bij een bepaalde waarde x_i van x behoort. De waarnemingsuitkomsten $(x_i,y_i)(i=1,2,\ldots,n)$ kunnen als punten in een vlak worden uitgezet, zoals in fig. 2,1 gedaan is.

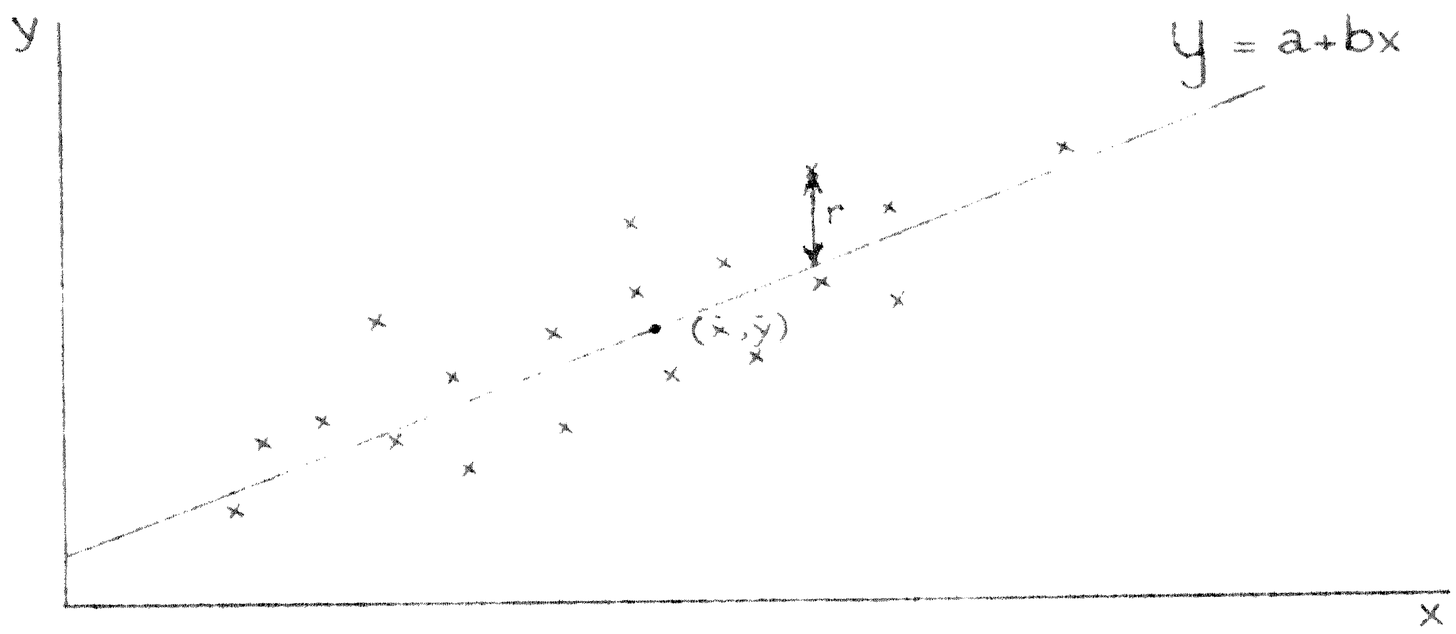

Fig. 2;1

Lineaire regressie van y op x

De verdelingsdichtheid van $\underline{y}_i$ is

$$(2;30) \qquad f(y_i|x_i) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{\{y_i - \mathcal{E}(\underline{y}_i|x_i)\}^2}{\sigma^2}}$$

Wegens de onafhankelijkheid der $\underline{y}_i$ is

$$(2;31) \qquad L(\alpha,\beta,\sigma^2) = \prod_{i=1}^{n} f(y_i|x_i)$$

en dus (vgl. (2;28))

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2. \tag{2;32}$$

Volgens de theorie van de meest aannemelijke schattingen moet log L gemaximaliseerd worden als functie van α, β en σ^2. Wij voeren dit in twee stappen uit: eerst maximaliseren wij log L als functie van α en β en daarna als functie van σ^2. Hiertoe geven wij (onder weglaten van de factor $-\frac{1}{2\sigma^2}$) de laatste term aan met Q, dus

$$Q = \sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2 . \tag{2;33}$$

Daar log L gemaximaliseerd moet worden, dienen wij van Q het minimum te bepalen als functie van α en β. Laat dit bereikt worden voor $\alpha = a$ en $\beta = b$, dan moeten (vergelijk stelling 6;7 van hoofdstuk II) a en b voldoen aan

$$Q_\alpha(a,b) = Q_\beta(a,b) = 0, \tag{2;34}$$

of

$$\sum_{i=1}^{n}(y_i-a-bx_i) = 0 \tag{2;35}$$

en

$$\sum_{i=1}^{n}(y_i-a-bx_i)x_i = 0. \tag{2;36}$$

Uit (2;35) volgt

$$\sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

of

$$a = \bar{y} - b\,\bar{x}\ , \tag{2,37}$$

waarin $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ en $\bar{x} = \frac{1}{n}\sum x_i$.

Substitutie van (2;37) in (2;36) leidt tot

$$\sum_{i=1}^{n}(y_i-\bar{y}+b\bar{x}-bx_i)x_i = \sum\{(y_i-\bar{y}) - b(x_i-\bar{x})\}x_i = 0$$

of

$$(2;38) \qquad b = \frac{\sum_{i=1}^{n}(y_i-\bar{y})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i} \quad .$$

Daar zowel $\sum_{i=1}^{n}(y_i-\bar{y})\,\bar{x}$ als $\sum_{i=1}^{n}(x_i-\bar{x})\,\bar{y}$ nul is, kan men voor (2;38) ook schrijven:

$$(2.39) \qquad b = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})} = \frac{\sum_{i=1}^{n}y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad .$$

Door de tweede partiële afgeleiden van Q naar α en β te bepalen, kan men laten zien dat Q inderdaad minimaal wordt, wanneer men de in (2;37) en (2.39) gevonden waarden substitueert (vgl. 6;7 van hoofdstuk II).

Nu wij voor a en b meest aannemelijke schattingen kennen, bezitten wij ook een meest aannemelijke schatting Y voor $\mathcal{E}(\underline{y}|x)$. Deze is

$$(2;40) \qquad Y = a + bx.$$

Y wordt de bij x behorende <u>regressiewaarde</u> genoemd; (2;40) stelt een rechte lijn voor, die door het punt $(\bar{x},\bar{y})$ gaat (het "zwaartepunt" van de "puntenwolk" in fig. 2;1) en die de (geschatte) <u>regressielijn</u> van y op x genoemd wordt. De "werkelijke" regressielijn wordt door (2;28) gegeven.

Uit (2;32) valt nu ook de meest aannemelijke schatting van σ^2 te berekenen, na in deze vorm voor α en β de reeds gevonden schattingen a en b ingevuld te hebben.

Noemen wij

$$(2;41) \qquad Q_o = \sum_{i=1}^{n}(y_i-a-bx_i)^2 \quad ,$$

dan gaat log L dus over in

$$- \frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{Q_o}{2\sigma^2}$$

en de afgeleide hiervan naar σ^2 is

$$- \frac{n}{2\sigma^2} + \frac{Q_o}{2\sigma^4} .$$

Stellen wij deze vorm gelijk aan 0, dan volgt daaruit de schatting $\hat{\sigma}^2$ van σ^2:

(2;42) $$\hat{\sigma}^2 = \frac{Q_o}{n} .$$

Wij onderzoeken nu of de gevonden schattingen zuiver zijn en beschouwen daartoe de $\underline{y}_i$ weer als stochastisch. Eerst de schatting b. In (2;39) is de noemer niet stochastisch, dus

(2;43) $$\mathcal{E}\underline{b} = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \cdot \sum_{i=1}^{n}(x_i-\bar{x})\,\mathcal{E}\underline{y}_i = \frac{\sum_{i=1}^{n}(x_i-\bar{x}).(\alpha+\beta x_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} =$$

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

Voor $\underline{a}$ keren wij terug tot (2;37) en berekenen eerst $\mathcal{E}\bar{\underline{y}}$.

(2;44) $$\mathcal{E}\bar{\underline{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(\underline{y}_i|x_i) = \frac{1}{n}\sum_{i=1}^{n}(\alpha+\beta x_i) = \alpha+\beta\bar{x} .$$

Substitutie van (2;37), (2;43) en (2;44) in $\mathcal{E}\underline{a}$ geeft

(2;45) $$\mathcal{E}\underline{a} = \mathcal{E}\bar{\underline{y}} - \mathcal{E}\underline{b}\,\bar{x} = \alpha+\beta\bar{x}-\beta\bar{x} = \alpha .$$

Wij zien dus dat $\underline{a}$ en $\underline{b}$ zuivere schattingen zijn van respectievelijk α en β . Verder is $\underline{\hat{\sigma}}^2$ geen zuivere schatting van σ^2. Een iets ingewikkelder berekening, die wij hier niet geven, leert dat

(2;46) $$\mathcal{E}\underline{\hat{\sigma}}^2 = \frac{n-2}{n}\sigma^2 ,$$

zodat $\underline{\hat{\sigma}}^2$ geen zuivere schatting is, maar

$$(2;47) \qquad \underline{s}^2 \overset{\text{def}}{=\!=} \frac{\underline{Q}_o}{n-2} = \frac{\sum_{i=1}^{n} (\underline{y}_i - \underline{a} - \underline{b}x_i)^2}{n-2}$$

wel.

De schattingen zijn ook <u>asymptotisch raak</u>, doch het bewijs daarvoor laten wij achterwege.

<u>Opmerkingen</u>

2;5. De hier toegepaste methode komt, meetkundig, daarop neer dat die lijn $Y = a+bx$ in fig. 2;1 gekozen wordt, waarvoor de som van de kwadraten van de afstanden der waargenomen punten tot die lijn, in verticale richting gemeten, minimaal is. Deze afstanden worden de <u>residuen</u> genoemd. Eén residu is in fig. 2;1 door de letter r aangegeven. Het bij (x_i, y_i) behorende residu is

$$(2;48) \qquad r_i \overset{\text{def}}{=\!=} y_i - a - bx_i \ ,$$

dus

$$(2;49) \qquad Q_o = \sum_{i=1}^{n} r_i^2 \ .$$

De methode ontleent daaraan de naam: <u>methode der kleinste kwadraten</u>. Onder deze naam is zij reeds zeer lang bekend als de basis van de z.g. <u>foutentheorie</u>, die veel toepassing vindt op het gebied van de natuur- en sterrekunde, ook voor veel gecompliceerde problemen dan de hier behandelde.

2;6. Ten onrechte wordt veelal ondersteld, dat de methode der kleinste kwadraten alleen voor normaal verdeelde $\underline{y}_i$ bruikbaar zou zijn. Uit het bovenstaande blijkt, dat bij de bewijzen der zuiverheid geen gebruik is gemaakt van die veronderstelling, zodat zij algemener gelden. Wel geldt speciaal in het normale geval, dat de methode der kleinste kwadraten ook de meest aannemelijke schattingen geeft. In andere gevallen kunnen de volgens de beide methoden verkregen schattingen verschillend zijn.

3. <u>Nauwkeurigheid van schattingen</u>

Een schatting wordt uit waarnemingen berekend. Deze waarnemingen worden als stochastische grootheden beschouwd, zodat niet alleen zij, maar ook de schatting een kansverdeling bezit. De uit-

komst, die verkregen wordt door bepaalde waargenomen waarden in de schattingsformule te substitueren - een concrete waarde van de schatting - kan zelf als één waarneming van die schatting beschouwd worden.

De kansverdeling van een schatting is een volledige beschrijving van de eigenschappen van die schatting, waarvan bijzondere eigenschappen, zoals zuiverheid en asymptotische raakheid speciale aspecten zijn. Een andere, in het voorafgaande nog niet beschouwde eigenschap is de nauwkeurigheid van een schatting, die grofweg door zijn spreiding gegeven wordt.

Voorbeeld 3;1. Spreiding van het gemiddelde van een steekproef

Wij beschouwen weer een steekproef van n waarnemingen

(3;1) $$\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$$

uit een normale verdeling met verwachting μ en variantie σ^2. Als schatting voor μ vonden wij in voorbeeld 1;1

(3;2) $$\overline{\underline{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i \ ,$$

hetgeen ook (vergelijk (2;10)) de meest aannemelijke schatting is. Volgens stelling 2;5 uit hoofdstuk IV is $\overline{\underline{x}}$ zelf ook normaal verdeeld met

(3;3) $$\mathcal{E}\overline{\underline{x}} = \mu \qquad \text{en} \qquad \sigma^2(\overline{\underline{x}}) = \frac{\sigma^2}{n} \ .$$

Voor de spreiding van $\overline{\underline{x}}$ geldt dus

(3;4) $$\sigma(\overline{\underline{x}}) = \frac{\sigma}{\sqrt{n}}$$

en men kan bewijzen (hetgeen wij hier niet zullen doen), dat er in het geval van een steekproef uit een normale verdeling geen enkele zuivere schatting van μ bestaat, die een kleinere spreiding bezit. Men noemt daarom $\overline{\underline{x}}$ de meest doeltreffende zuivere schatting van μ.

De grootheid

(3;5) $$\overline{\underline{x}}^* = \frac{\overline{\underline{x}} - \mu}{\sigma} \sqrt{n} \ ,$$

de gestandaardiseerde van $\underline{\bar{x}}$, is dus N(0,1)-verdeeld, zodat volgens tabel 4;II uit hoofdstuk III bijv. geldt:

$$P\left[|\underline{x}^*| \leq 1{,}96\right] = 0{,}95$$

of, na omrekenen van de ongelijkheid tussen de vierkante haken:

(3;6) $$P\left[\underline{\bar{x}} - 1{,}96\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \underline{\bar{x}} + 1{,}96\,\frac{\sigma}{\sqrt{n}}\right] = 0{,}95\ .$$

Bij voorbeeld 1;2 vonden wij $\underline{\bar{x}} = 1016{,}5$. De ongelijkheid tussen de vierkante haken van (3;6) gaat, bij substitutie van deze waarde over in

$$1016{,}5 - 1{,}96\,\frac{\sigma}{\sqrt{10}} \leq \mu \leq 1016{,}5 + 1{,}96\,\frac{\sigma}{\sqrt{10}}\ ,$$

waarmee wij nog niet veel opschieten als σ onbekend is. Nemen wij even aan, dat σ op grond van vroegere waarnemingen bekend is en bijv. 33 bedraagt, dan vinden wij na invullen van deze waarde

(3;7) $$996 \leq \mu \leq 1037\ .$$

Deze uitspraak, waarbij de onbekende parameter μ tussen twee grenzen ingesloten wordt, kan juist of onjuist zijn. Op grond van (3;6) zou men de neiging kunnen hebben te zeggen, dat er een kans 0,95 is, dat hij juist is. Strikt genomen is dit niet waar: (3;7) is òfwel juist (en dan is de kans, dat hij juist is gelijk aan 1) òfwèl onjuist (en dan is die kans gelijk aan 0). Maar wèl volgt uit (3;6), dat de gevolgde <u>methode</u> een kans 0,95 bezit óp een juiste uitkomst. Beperkt men dus zijn beschouwingen niet tot één geval (zoals de uitkomst (3;7)), maar beschouwt men de methode als zodanig, of een lange reeks toepassingen daarvan, dan kan men zeggen, dat de kans op een juiste uitspraak 0,95 bedraagt.

Men noemt een interval, zoals (3;6) of (3;7) een <u>betrouwbaarheidsinterval</u> voor μ en de kans 1-0,95 = 0,05 <u>de onbetrouwbaarheid</u>. Wij komen hierop terug in paragraaf 3 van hoofdstuk VI.

Meestal is σ niet bekend. In eerste instantie substitueert men dan veelal een schatting van σ, bijv. $s = \sqrt{s^2}$ (zie (1;7)) voor σ. Zoals wij hebben gezien is deze schatting niet zuiver: gemiddeld wordt σ door <u>s</u> onderschat. Dit doet vermoeden, dat de grenzen, waartussen μ op deze wijze ingesloten wordt wel eens te dicht bij elkaar kunnen komen, of ook dat de kans op een verkeerde

uitspraak bij toepassing van deze methode groter dan α zal zijn. Dit is inderdaad juist en men kan de substitutie van s voor σ dan ook slechts als een benadering beschouwen. Voor grote steekproeven ($n \to \infty$) bestaat dit bezwaar niet. Dan convergeert nl. s stochastisch naar σ. Voor kleine steekproeven bestaat - in het normale geval - een exacte methode, die op de z.g. verdeling van STUDENT berust. Wij gaan daar niet op in.

Uit (3;6) is (nogmaals) te zien, dat μ in principe met willekeurige nauwkeurigheid bepaald kan worden door n voldoende groot te nemen. Immers voor $n \to \infty$ naderen de beide grenzen tot elkaar. Dit is aequivalent met het vroeger bewezene, dat $\bar{x}$ voor $n \to \infty$ stochastisch tot μ convergeert.

In de praktijk dient deze stelling echter met voorzichtigheid gehanteerd te worden. Indien men bijv. de lengte van een voorwerp een aantal malen opmeet met een meetinstrument dat niet nauwkeuriger dan in cm afleesbaar is en de werkelijke lengte van het voorwerp is gelijk aan 10,9 cm, dan zullen alle waarnemingen (of vrijwel alle) de waarde 11 cm aangeven en men behoeft er niet op te hopen, dat het gemiddelde tot 10,9 cm zal convergeren. De kansverdeling der waarnemingen is dan ook niet normaal, zelfs niet continu, daar alleen gehele waarden aangenomen kunnen worden. In de praktijk zijn echter waarnemingen altijd van discreet karakter. De moraal daarvan is, dat het in de regel zinloos is het gemiddelde van een aantal metingen in meer decimalen op te geven dan op het meetinstrument afgelezen kunnen worden. Hiertegen wordt vaak gezondigd.

Voorbeeld 3;2. Binomiale verdeling

Bij een binomiale verdeling met parameters n en p vonden wij in voorbeeld 2;2 als meest aannemelijke schatting voor p de grootheid $\frac{k}{n}$. De variantie hiervan is volgens formule (4;29) en stelling 4;5 uit hoofdstuk III

$$(3;8) \qquad \sigma^2\left(\frac{k}{n}\right) = \frac{1}{n^2} \cdot \sigma^2(k) = \frac{1}{n^2} \cdot npq = \frac{pq}{n} .$$

De variantie, die weer de kleinst mogelijke voor een zuivere schatting van p is, is weer in de regel onbekend, doch kan uit de waarnemingen geschat worden. Een zuivere schatting ervan is

(3;9) $$\frac{\underline{k}(n-\underline{k})}{n^2(n-1)} ,$$

hetgeen volgt uit

(3;10) $$\mathcal{E}\underline{k}(n-\underline{k}) = n\,\mathcal{E}\underline{k} - \mathcal{E}\underline{k}^2 = n\,\mathcal{E}\underline{k} - \sigma^2(\underline{k}) - (\mathcal{E}\underline{k})^2 =$$
$$= n^2p - npq - (np)^2 = n(n-1)\,pq .$$

Voorbeeld 3;3. Lineaire regressie

Als laatste voorbeeld berekenen wij nog de varianties van de regressiecoëfficiënten $\underline{a}$ en $\underline{b}$ uit (2;37) en (2;39). Daar alle $\underline{y}_i$ dezelfde variantie σ^2 bezitten en onafhankelijk verdeeld zijn, volgt uit (2;37) onmiddellijk

(3;11) $$\sigma^2(\underline{a}) = \sigma^2(\overline{\underline{y}}) = \frac{\sigma^2}{n} .$$

Volgens (2;39) is $\underline{b}$ een lineaire combinatie der $\underline{y}_i$ van de vorm

$$\underline{b} = \sum_{i=1}^{n} c_i \underline{y}_i ,$$

met

$$c_i = \frac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} ,$$

waarin de x_i bekende (niet stochastische) getallen zijn. Derhalve geldt

$$\sigma^2(\underline{b}) = \sum_{i=1}^{n} c_i^2\, \sigma^2(\underline{y}_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 ,$$

dus

(3;12) $$\sigma^2(\underline{b}) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} .$$

Daar verder zowel $\underline{a}$ als $\underline{b}$ lineaire combinaties van de normaal verdeelde grootheden $\underline{y}_i$ zijn, zijn zij zelf ook normaal verdeeld. Bovendien geldt voor de schattingen $\underline{a}$ en $\underline{b}$ dat zij - althans in het normale geval - de kleinst mogelijke spreiding bezitten.

Tenslotte valt op te merken, dat de optimale eigenschap van kleinst mogelijke spreiding, die in het bovenstaande herhaaldelijk is vermeld, niet toevallig ontstaat. Onder vrij algemene voorwaarden bezitten nl. de meest aannemelijke schattingen, althans asymptotisch voor $n \to \infty$, de eigenschap minimale spreidingen te bezitten en normaal verdeeld te zijn. In vele gevallen is de eerstgenoemde eigenschap ook reeds voor eindige n vervuld, terwijl de laatstgenoemde vaak voorkomt als de waarnemingen normaal verdeeld zijn. Het zijn o.a. deze eigenschappen, die de theorie der meest aannemelijke schattingen haar aantrekkelijkheid verlenen.

4. De momentenmethode

De oudste methode om onbekende parameters van een verdeling te schatten is de zogenaamde momentenmethode, welke afkomstig is van K. PEARSON.

PEARSON stelde voor de onbekende parameters zodanig te kiezen dat zoveel mogelijk momenten in de steekproef dezelfde waarde bezitten als in de verdeling. Men zorgt er dan eerst voor dat de eerste momenten (verwachtingen) gelijk zijn, vervolgens indien mogelijk de tweede gereduceerde momenten (varianties), dan indien mogelijk de derde momenten, enzovoorts. Bij een normale verdeling, die twee onbekende parameters heeft, μ en σ^2 worden dus twee momenten gelijk gemaakt: de verwachting en de variantie. Men eist

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \mu \qquad (4;1)$$

en

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sigma^2 \qquad (4;2)$$

en geeft dus als schatting voor μ op $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ en als schatting voor σ^2: $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$. Hier leidt deze methode dus tot dezelfde schattingen als de methode van de meest aannemelijke schattingen (vgl. (2;10) en (2;11)).

In het algemeen behoeft dit niet het geval te zijn. Een belangrijk voordeel van deze schattingsmethode is dat de schattingen vaak veel eenvoudiger berekend kunnen worden dan de meest aanneme-

lijke schattingen. Juist om deze reden wordt ze meestal toegepast bij gamma-verdelingen. Stellen wij de gamma-verdeling weer voor door

$$(4;3) \qquad f(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} \qquad \text{(vgl. (1;25) van hfdst.III)},$$

dan is het eerste moment

$$(4;4) \qquad \mathcal{E}\underline{x} = \frac{\alpha}{\lambda} \qquad \text{(vgl. (4;9) van hfdst.III)}$$

en de variantie

$$(4;5) \qquad \sigma^2(\underline{x}) = \frac{\alpha}{\lambda^2} \qquad \text{(vgl. (4;25) van hfdst.III)}.$$

Wanneer de steekproefuitkomst $x_1,\ldots,x_n$ gegeven is en dus het rekenkundig gemiddelde $\frac{1}{n}\sum_{i=1}^{n} x_i$ en de variantie $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ in de steekproef bekend zijn, dan worden op eenvoudige wijze schattingen voor α en λ gevonden door deze waarden gelijk te stellen aan (4;4) resp. (4;5). De schattingen luiden dan

$$(4;6) \qquad \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \text{voor } \lambda$$

en

$$(4;7) \qquad \frac{(\sum_{i=1}^{n} x_i)^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \text{voor } \alpha .$$

De meest aannemelijke schattingen voor λ en α kunnen in dit geval niet expliciet opgeschreven worden. Wij zullen daarom later (onder andere in paragraaf 4 van hoofdstuk VI) gebruik maken van de schattingen (4;6) en (4;7).

Tegenover het genoemde voordeel van de hier behandelde schattingsmethode staat het nadeel dat de schattingen in het algemeen bij gelijke steekproefomvang een grotere spreiding bezitten dan meest aannemelijke schattingen en dus onnauwkeuriger zijn; dit blijft vaak gelden wanneer de steekproefomvang willekeurig toeneemt.