WFk Y.bM

Onderafdeling der Wiskunde

Mathematische Statistiek

SYLLABUS NAAR HET COLLEGE VAN PROF. DR. R. DOORNBOS

NAJAARSSEMESTER 1970

TECHNISCHE HOGESCHOOL EINDHOVEN

DICT.NR. 2.240
PRIJS f 1,50

TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

MATHEMATISCHE STATISTIEK

Syllabus naar het college van

Prof. Dr. R. Doornbos

Najaarssemester 1970

Inhoudsbeschrijving

MATHEMATISCHE STATISTIEK

R. Doornbos

Najaarssemester 1970

JdG, 22 Juli 2005

Onderafdeling der Wiskunde

Mathematische Statistiek

Syllabus naar het college van Prof.dr. R. Doornbos

Najaarssemester 1970

Technische Hogeschool Eindhoven

College Mathematische Statistiek, najaar 1970

Hoofdstuk I, Momenten en voortbrengende functies

(1.1) Definitie: Zij $a_0, a_1, a_2, \dots$ een rij van reële getallen. Als

$$A(s) = a_0 + a_1 s + a_2 s^2 + \dots$$

convergeert in een interval $-s_0 < s < s_0$, dan heet de functie $A(s)$ de voortbrengende functie van de rij $\{a_j\}$.

Als de rij $\{a_j\}$ begrensd is, dan leert vergelijking met de meetkundige reeks dat $A(s)$ zeker convergeert voor $|s| < 1$.

Voorbeelden

1) $a_j = 1$ voor alle j; $A(s) = \frac{1}{1-s}$.

2) $\{a_j\} = 0,0,1,1,\dots$; $A(s) = \frac{s^2}{1-s}$.

3) $a_j = \binom{n}{j}$ (n geheel, dus $\binom{n}{j} = 0$ voor $j > n$); $A(s) = (1+s)^n$.

Beschouw nu een stochastische variabele $\underline{x}$ die de waarden 0,1,2,... kan aannemen, met kansen

$$P(\underline{x} = j) = p_j \ .$$

Bij deze kansen hoort de voortbrengende functie

(1.2) $$A(s) = p_0 + p_1 s + p_2 s^2 + \dots \ .$$

$A(1) = 1$, dus $A(s)$ convergeert voor $|s| \leq 1$.

Voorbeeld:

Als $\underline{x}$ het aantal ogen is dat wordt geworpen met een zuivere dobbelsteen, dan is de voortbrengende functie

$$\frac{1}{6}(s + s^2 + s^3 + s^4 + s^5 + s^6) = \frac{s(1-s^6)}{6(1-s)} \ .$$

(1.3) Stelling: De verwachting $E\underline{x}$ kan uit (1.2) worden gevonden:

$$E\underline{x} = A'(1) \ .$$

Bewijs: Als $E\underline{x}$ eindig is volgt het uit de definitie. Als $E\underline{x} = +\infty$, dan neemt $A'(s)$ onbeperkt toe voor $s \to 1$.

(1.4) Stelling: Als $E\underline{x}^2$ eindig is, dan geldt

$$E\underline{x}^2 = A''(1) + A'(1) ,$$

en dus

$$\mathrm{var}(\underline{x}) = A''(1) + A'(1) - \{A'(1)\}^2 .$$

Bewijs: $E\underline{x}^2 = E\underline{x}(\underline{x}-1) + E(\underline{x})$.

De voortbrengende functie (1.2) kan ook geschreven worden als een verwachting:

(1.5) $$A(s) = E(s^{\underline{x}}) .$$

In het algemeen geldt

$$\frac{d^r}{ds^r} A(1) = E\underline{x}(\underline{x}-1)\ldots(\underline{x}-r+1) ,$$

het zogenaamde r^e factoriële moment van de verdeling van $\underline{x}$. Daarom heet (1.2) ook wel de voortbrengende functie van de factoriële momenten.
Met behulp van deze functie kunnen momenten van discrete verdelingen vaak gemakkelijk worden gevonden.

Voorbeeld:
De Poissonverdeling met verwachting μ heeft de voortbrengende functie

$$A(s) = Es^{\underline{x}} = \sum_{x=0}^{\infty} \frac{e^{-\mu}\mu^x s^x}{x!} = e^{\mu(s-1)} .$$

Volgens stelling (1.3) is

$$E\underline{x} = A'(1) = \mu$$

en volgens (1.4) is

$$\mathrm{var}(\underline{x}) = A''(1) + A'(1) - \{A'(1)\}^2 = \mu^2 + \mu - \mu^2 = \mu .$$

Stel $\underline{x}$ en $\underline{y}$ zijn onafhankelijk verdeelde stochastische variabelen met verdelingen $P(\underline{x}=j) = a_j$ en $P(\underline{y}=j) = b_j$ $(j = 0,1,2,\ldots)$.
De verdeling van de som $\underline{z} = \underline{x}+\underline{y}$ is

(1.6) $$c_r = P(\underline{z}=r) = a_0b_r + a_1b_{r-1} + a_2b_{r-2} + \ldots + a_rb_0 .$$

Definitie: De rij $\{c_r\}$, gedefinieerd door (1.6) heet de convolutie van $\{a_j\}$ en $\{b_j\}$ en wordt geschreven als

(1.7) $$\{c_k\} = \{a_k\} * \{b_k\} .$$

Opmerking:

1) De rijen $\{a_j\}$ en $\{b_j\}$ hoeven niet beslist kansverdelingen te zijn.
2) Het analogon voor continue functies werd besproken in het hoofdstuk over Laplace-transformaties in Wiskunde III.

Als nu $A(s) = \Sigma\, a_k s^k$ en $B(s) = \Sigma\, b_k s^k$ de voortbrengende functies van $\{a_k\}$ en $\{b_k\}$ zijn, dan zien we dat de coëfficiënt van s^r in het product $A(s)\cdot B(s)$ gelijk is aan c_r uit (1.6). Dus geldt de volgende stelling:

Stelling: Als $\{a_k\}$ en $\{b_k\}$ rijen zijn met voortbrengende functies $A(s)$ en $B(s)$ en $\{c_k\}$ is de convolutie van deze rijen, dan is de voortbrengende functie $C(s) = \Sigma\, c_k s^k$ het product

(1.8) $$C(s) = A(s)\cdot B(s) .$$

Als $\underline{x}$ en $\underline{y}$ onafhankelijke stochastische variabelen zijn die de waarden 0,1,2,... kunnen aannemen en met voortbrengende functies $A(s)$ en $B(s)$, dan heeft de som $\underline{x}+\underline{y}$ de voortbrengende functie $A(s)\cdot B(s)$.

Deze stelling is door inductie gemakkelijk uit te breiden tot meer dan twee rijen. Vaak hebben wij te maken met de som van een aantal onafhankelijke stochastische variabelen met dezelfde verdeling: $\underline{s}_n = \underline{x}_1 + \dots + \underline{x}_n$. Als elk van de $\underline{x}_i$ de voortbrengende functie $A(s)$ heeft, dan is de voortbrengende functie van $\underline{s}_n$: $A^n(s)$.

Voorbeelden:

1) Wat is de kans om met 5 dobbelstenen 15 ogen te gooien?
De voortbrengende functie van de verdeling van het aantal ogen met één dobbelsteen is:

$$\frac{1}{6}\,(s+s^2+s^3+s^4+s^5+s^6) = \frac{s(1-s^6)}{6(1-s)} .$$

Dus van 5 dobbelstenen:

$$\frac{s^5(1-s^6)^5}{6^5(1-s)^5} = \frac{s^5(1-s^6)^5}{6^5} \sum_{k=0}^{\infty} \binom{-5}{k}(-s)^k = \frac{s^5(1-s^6)^5}{6^5} \sum_{k=0}^{\infty} \binom{4+k}{k} s^k .$$

De coëfficiënt van s^{15} hierin is

$$\frac{1}{6^5}\left\{\binom{14}{10} - 5\binom{8}{4}\right\}$$

en dit is de gevraagde kans.

2) Als $\underline{y}$ binomiaal verdeeld is met parameters n (aantal trekkingen) en p (kans op succes), dan kan $\underline{y}$ beschouwd worden als de som van n variabelen $\underline{x}_i$ die elk een Bernoulli verdeling hebben.
De voortbrengende functie van $\underline{x}_i$ is:

$$A(s) = (q + ps) .$$

Dus de voortbrengende functie van $\underline{y}$ is:

$$B(s) = A^n(s) = (q+ps)^n = \sum_{k=0}^{n} \binom{n}{k} q^{n-k}(ps)^k .$$

De bekende waarden voor $E\underline{x}$ en $var(\underline{x})$ volgen uit de stellingen (1.3) en (1.4):

$$E\underline{x} = B'(1) = np$$

$$var(\underline{x}) = B''(1) + B'(1) - \{B'(1)\}^2 = n(n-1)p^2 + np - n^2p^2 = npq .$$

3) <u>De negatief binomiale verdeling</u>
Bij trekkingen uit een alternatief met kans p op succes telt men het aantal mislukkingen $\underline{x}$ voordat r successen zijn bereikt.

(1.9) $$P(\underline{x}=x) = \binom{x+r-1}{x} p^{r-1} q^x p = \binom{x+r-1}{x} p^r q^x = \binom{-r}{x} p^r (-q)^x .$$

De voortbrengende functie is

$$A(s) = \sum_{x=0}^{\infty} \binom{-r}{x} p^r (-qs)^x = p^r(1-sq)^{-r} .$$

We zien dat $\underline{x}$ weer opgevat kan worden als de som van r variabelen, nl. het aantal mislukkingen voordat één succes is bereikt.
De verwachting van $\underline{x}$ is:

$$E\underline{x} = A'(1) = qr\, p^r(1-q)^{-r-1} = \frac{rq}{p} .$$

4) Niet van elke discrete verdeling kunnen de momenten met de voortbrengende functie worden gevonden. Bij de hypergeometrische verdeling bijv. moet dit rechtstreeks gebeuren.
Als uit een vaas met m zwarte en n witte ballen zonder teruglegging k stuks worden getrokken, dan heeft het aantal zwarte ballen $\underline{x}$ van deze k een hypergeometrische verdeling

(1.10) $$P(\underline{x}=x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} .$$

Daar de som van deze kansen 1 is, geldt

(1.11) $$\sum_{x=0}^{k} \binom{m}{x}\binom{n}{k-x} = \binom{m+n}{k} .$$

De verwachting is

$$E\underline{x} = \sum_{x=0}^{k} \frac{x\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} = \sum_{x=1}^{k} \frac{m\binom{m-1}{x-1}\binom{n}{(k-1)-(x-1)}}{\binom{m+n}{k}} =$$

$$= (\text{volgens } (1.11)) = \frac{m\binom{m+n-1}{k-1}}{\binom{m+n}{k}} = \frac{mk}{m+n} .$$

Op dezelfde manier vindt men

$$E\underline{x}(\underline{x}-1) = \frac{m(m-1)k(k-1)}{(m+n)(m+n-1)} ,$$

waaruit volgt

$$\operatorname{var}(\underline{x}) = \frac{mnk(m+n-k)}{(m+n)^2(m+n-1)} .$$

De momenten $\mu_r' = E\underline{x}^r$ van een verdeling kunnen direct worden gevonden uit de momenten voortbrengende functie, gedefinieerd door

(1.12) $$m(t) = Ee^{t\underline{x}} = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

als $\underline{x}$ een continue verdeling heeft en

(1.13) $$m(t) = Ee^{t\underline{x}} = \sum_{x} e^{tx} f(x)dx$$

als $\underline{x}$ discreet is.

Als $m(t)$ bestaat voor $-h^2 < t < h^2$, dan is $m(t)$ in de buurt van $t = 0$ continu differentieerbaar naar t en er geldt:

$$\frac{d^r}{dt^r} m(t) = E\underline{x}^r e^{t\underline{x}} ,$$

dus

$$\frac{d^r}{dt^r} m(0) = E\underline{x}^r = \mu_r' .$$

Eigenlijk is de naam momenten voortbrengende functie niet geheel juist, want $m(t)$ brengt niet de rij van de momenten voort, maar de rij $\frac{\mu_r'}{r!}$, immers

$$m(t) = E\left(1 + \underline{x}t + \frac{\underline{x}^2 t^2}{2!} + \frac{\underline{x}^3 t^3}{3!} + \ldots\right) .$$

<u>Voorbeeld</u>:

De normale verdeling

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

heeft als momenten voortbrengende functie

$$m(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{xt - \frac{(x-\mu)^2}{2\sigma^2}} dx =$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu-t\sigma^2)^2}{2\sigma^2}} e^{\mu t + \frac{1}{2}t^2\sigma^2} dx = e^{\mu t + \frac{1}{2}t^2\sigma^2} .$$

Hieruit is gemakkelijk af te leiden dat $E\underline{x} = \mu$ en $var(\underline{x}) = \sigma^2$.

De momenten om het gemiddelde

$$\mu_r = E(\underline{x} - \mu_1')^2$$

kunnen direct door differentiatie worden gevonden uit de functie

(1.14) $\quad n(t) = Ee^{t(\underline{x}-\mu)}$,

als $\mu = E\underline{x}$.

In het geval van de normale verdeling is

$$n(t) = e^{\frac{1}{2}t^2\sigma^2} = 1 + \frac{t^2\sigma^2}{2} + \frac{t^4\sigma^4}{2^2 2!} + \frac{t^6\sigma^6}{2^3 3!} + \dots .$$

Hieruit volgt dus direct dat

$$\mu_{2n+1} = 0$$

$$\mu_{2n} = \frac{(2n)!\,\sigma^{2n}}{2^n\, n!} .$$

Opmerking:
De momenten voortbrengende functie $m(t)$ is nauw verwant met de Laplace-transformatie. Voor continue stochastische variabelen, die alleen positieve waarden aan kunnen nemen, is $m(t) = F(-t)$, als $F(t)$ de Laplace-getransformeerde is:

$$F(t) = \int_0^\infty e^{-tx} f(x)dx .$$

In beschouwingen over Laplace-transformaties wordt $m(t)$ daarom wel de tweezijdige Laplace-getransformeerde genoemd. In de theoretische statistiek wordt veelal de karakteristieke functie $Ee^{it\underline{x}}$ gebruikt met zuiver imaginaire exponent, die voor elke reële waarde van t convergeert. Afgezien van een factor 2π in de exponent is dit hetzelfde als de Fourier-getransformeerde, die we tegenkwamen in Wiskunde III.

(1.15) Stelling: Stel $\underline{x}$ en $\underline{y}$ zijn stochastische variabelen met kansdichtheden $f(x)$ en $g(y)$ (geheel discreet of geheel continu). Als de momenten voortbrengende functies van $\underline{x}$ en $\underline{y}$ beide bestaan en gelijk zijn voor alle t in een interval $-h^2 \leq t \leq h^2$, dan zijn de dichtheden f en g gelijk.

Hieruit volgt dus, dat, onder de in de stelling genoemde voorwaarden de kansdichtheid door de momenten wordt bepaald.

Stel $\underline{x}$ en $\underline{y}$ zijn onderling onafhankelijk verdeeld met kansdichtheden $f(x)$ en $g(y)$. De simultane verdeling van $\underline{x}$ en $\underline{y}$ is dan $f(x)g(y)$.
De verdeling van de som $\underline{z} = \underline{x}+\underline{y}$ is

(1.16) $$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx = \int_{-\infty}^{\infty} f(z-y)g(y)dy \; .$$

Als $\underline{x}$ en $\underline{y}$ alleen positieve waarden kunnen aannemen gaat (1.16) over in

(1.17) $$h(z) = \int_{0}^{z} f(x)g(z-x)dx = \int_{0}^{z} f(z-y)g(y)dy \; ,$$

de uit Wiskunde III bekende <u>convolutie</u> $f * g$. Daar werd ook bewezen dat voor de Laplace-getransformeerden geldt:

$$L(f * g) = L(f)L(g) \; .$$

Wij hebben nu het convolutiebegrip eerst uitgebreid tot het discrete geval ((1.6) en (1.7)) en nu ook voor functies die van $-\infty$ tot $+\infty$ zijn gedefinieerd.
In het algemeen geldt, voor onafhankelijke variabelen:

(1.18) $$Ee^{t(\underline{x}+\underline{y})} = Ee^{t\underline{x}}e^{t\underline{y}} \; ,$$

of: de momenten voortbrengende functie van de som is het product van de momenten voortbrengende functies.

Voor meerdimensionale verdelingen worden momenten op voor de hand liggende manier gedefinieerd, bv.

(1.19) $$E\underline{x}^p\underline{y}^q\underline{z}^r = \iiint_{-\infty}^{\infty} x^p y^q z^r \, f(x,y,z)dx\,dy\,dz \; ,$$

als $f(x,y,z)$ de simultane verdeling is van $\underline{x}$, $\underline{y}$ en $\underline{z}$.
Het belangrijkste speciale geval is de <u>covariantie</u>

(1.20) $$\sigma_{x,y} = E(\underline{x}-E\underline{x})(\underline{y}-E\underline{y}) = E\underline{x}\underline{y} - E\underline{x}E\underline{y} \; .$$

De <u>correlatiecoëfficiënt</u> $\rho_{x,y}$ van $\underline{x}$ en $\underline{y}$ wordt gedefinieerd door

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x\sigma_y} \; ,$$

als σ_x en σ_y de standaardafwijkingen van $\underline{x}$ en $\underline{y}$ voorstellen.

Een simultane momenten voortbrengende functie is bv.

(1.22) $\quad m(t_1,t_2,t_3) = Ee^{t_1\underline{x}+t_2\underline{y}+t_3\underline{z}}$.

Het moment (1.19) vindt men door $m(t_1,t_2,t_3)$ px naar t_1, qx naar t_2 en rx naar t_3 te differentiëren en dan de drie t's nul te stellen.

(1.23) Definitie: Stel $\underline{x}$ en $\underline{y}$ hebben de simultane kansdichtheid $f(x,y)$. De marginale dichtheid van $\underline{x}$ is $h(x)$. De voorwaardelijke dichtheid van $\underline{y}$, gegeven $\underline{x}=x$ wordt aangegeven door

$$g(y \mid x) = \frac{f(x,y)}{h(x)} .$$

De voorwaardelijke verwachting van $\underline{y}$, gegeven $\underline{x}=x$ wordt aangeduid met $E(\underline{y} \mid x)$ en gedefinieerd door

$$E(\underline{y} \mid x) = \int_{-\infty}^{\infty} yg(y \mid x)dy .$$

(1.24) Stelling: $E[E(\underline{y} \mid \underline{x})] = E\underline{y}$.

Bewijs: $$E[E(\underline{y} \mid \underline{x})] = \int_{-\infty}^{\infty} h(x) \int_{-\infty}^{\infty} yg(y \mid x)dy\,dx =$$

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

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)dx\,dy = E\underline{y} .$$

Vraagstukken:

1) Stel $\underline{x}$ is een stochastische variabele die de waarden 0,1,2,... kan aannemen. De voortbrengende functie van de kansverdeling is $P(s)$.
Wat zijn de voortbrengende functies die horen bij $\underline{x}+1$ en $2\underline{x}$?

2) $P(s)$ is de voortbrengende functie van de rij $\{P(\underline{x}=j)\}$ $(j = 0,1,2,\ldots)$.
Wat is de voortbrengende functie van
a) $\{P(\underline{x} \leq j)\}$
b) $\{P(\underline{x} < j)\}$
c) $\{P(\underline{x} \geq j)\}$
d) $\{P(\underline{x} > j+1)\}$
e) $\{P(\underline{x} = 2j)\}$.

3) In een reeks trekkingen uit een alternatief met kans p op succes (S) en q op mislukking (F) is u_n de kans dat de eerste combinatie SF voorkomt in de $(n-1)^e$ en n^e trekking.

Wat is de voortbrengende functie van de rij $\{u_n\}$, wat het gemiddelde en de variantie van de bijbehorende stochastische variabele?

Hoofdstuk II, Enige continue kansverdelingen

De eenvoudigste continue verdeling is de homogene (of uniforme of rechthoekige) verdeling

$$(2.1) \qquad f(x) = \frac{1}{\beta - \alpha} \quad , \quad \alpha \leqslant x \leqslant \beta$$

$$= 0 \quad , \quad \text{overal elders.}$$

Iedere continue stochastische variabele $\underline{x}$ kan door de transformatie

$$\underline{y} = G(\underline{x}) \; ,$$

waarbij $G(x)$ de cumulatieve verdeling van $\underline{x}$ is, in een homogeen tussen 0 en 1 verdeelde variabele $\underline{y}$ worden overgevoerd. Omdat $G(x)$ een monotoon niet dalende functie is geldt immers:

$$P\{G(\underline{x}) \leqslant G(x)\} = P(\underline{x} \leqslant x) = G(x) \; ,$$

of m.a.w.

$$P(\underline{y} \leqslant y) = y \; , \quad 0 \leqslant y \leqslant 1$$

dus

$$f(y) = 1 \; , \quad 0 \leqslant y \leqslant 1 \; .$$

De normale verdeling

$$(2.2) \qquad n(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ligt ten grondslag aan veel methoden die in de toegepaste statistiek worden gebruikt. De momenten voortbrengende functie werd op pag. 6 reeds behandeld.

De gamma verdeling heeft de volgende gedaante:

$$(2.3) \qquad f(x;\alpha,\beta) = \frac{1}{\alpha!\beta^{\alpha+1}} x^{\alpha} e^{-x/\beta} \; , \quad 0 < x < \infty \; .$$

β is een schaalparameter, die > 0 moet zijn, terwijl $\alpha > -1$ is. In de meeste praktische toepassingen is α geheel en positief of een geheel getal plus $\frac{1}{2}$. Als α niet geheel is, wordt $\alpha!$ gedefinieerd door

$$\alpha! = \Gamma(\alpha+1) = \int_0^\infty t^{\alpha} e^{-t} dt \; .$$

Voor gehele positieve α komt dit met de gewone definitie van de faculteit overeen, want

$$0! = \int_0^\infty e^{-t} dt = 1$$

en

$$\alpha! = \int_0^\infty t^\alpha e^{-t} dt = -t^\alpha e^{-t} \Big|_0^\infty + \alpha \int_0^\infty t^{\alpha-1} e^{-t} dt = \alpha(\alpha-1)! \; .$$

Dat (2.3) een kansverdeling is volgt nu uit

$$\frac{1}{\alpha!\beta^{\alpha+1}} \int_0^\infty x^\alpha e^{-x/\beta} dx = (\text{stel } x/\beta = y) = \frac{1}{\alpha!} \int_0^\infty y^\alpha e^{-y} dy = 1 \; .$$

Als we nog de waarde van $(-\frac{1}{2})!$ bepalen, dan kunnen we $\alpha!$ berekenen voor alle waarden van α die ons interesseren.

$$(-\tfrac{1}{2})! = \int_0^\infty t^{-\frac{1}{2}} e^{-t} dt = (\text{stel } t = \tfrac{1}{2} z^2) = \sqrt{2} \int_0^\infty e^{-\frac{1}{2} z^2} dz = \tfrac{1}{2}\sqrt{2}\sqrt{2\pi} = \sqrt{\pi} \; ,$$

omdat $\frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-\frac{1}{2} z^2} dz$ de helft is van het oppervlak onder de normale kansdichtheid.

De momenten voortbrengende functie is:

$$(2.4) \qquad m(t) = \frac{1}{\alpha!\beta^{\alpha+1}} \int_0^\infty x^\alpha e^{(tx - x/\beta)} dx = (\text{stel } x = y\beta)$$

$$= \frac{1}{\alpha!} \int_0^\infty y^\alpha e^{-y(1-\beta t)} dy = (1-\beta t)^{-\alpha-1} \; .$$

Hieruit volgt:

$$(2.5) \qquad \begin{cases} \mu = m'(0) = \beta(\alpha+1) \; , \\ \sigma^2 = m''(0) - \{m'(0)\}^2 = \beta^2(\alpha+1) \; . \end{cases}$$

Uit (2.4) volgt dat de som van twee onafhankelijke stochastische variabelen met gammaverdelingen met parameters α_1, β en α_2, β als momenten voortbrengende functie heeft: $(1-\beta t)^{-\alpha_1-\alpha_2-2}$, en dus een gamma verdeling volgt met parameters $\alpha_1 + \alpha_2 + 1$ en β.

Uit (2.4) volgt ook dat we een variabele met gamma verdeling $f(x;\alpha,\beta)$ kunnen opvatten als de som van $\alpha+1$ onafhankelijke variabelen met verdeling

$$f(x;0,\beta) = \frac{1}{\beta} e^{-x/\beta} ,$$

de <u>negatief exponentiële</u> verdeling met gemiddelde β en variantie β^2. Wij merken op dat hieruit ook meteen de formules (2.5) voor gemiddelde en variantie van de gamma verdeling volgen.

Er bestaat een nauw verband tussen de gamma verdeling en de Poisson verdeling. Als bij een proces het aantal gebeurtenissen per tijdseenheid een Poisson verdeling heeft met gemiddelde $1/\beta$ (radioactiviteit, draadbreuken, aanvragen voor telefoongesprekken), dan heeft de tijd $\underline{t}$ tussen twee gebeurtenissen een negatief exponentiële verdeling met gemiddelde β. Hetzelfde treedt op bij Poisson verdelingen in de ruimte (aantallen bacteriën e.d.). Daar kan een negatief exponentiële verdeling worden verbonden met een volume. Dat de bovengenoemde relatie bestaat is eenvoudig als volgt in te zien. Het aantal gebeurtenissen $\underline{k}$ in een tijd t heeft een Poisson verdeling met gemiddelde t/β:

$$P(\underline{k} = k) = \frac{e^{-t/\beta}(t/\beta)^k}{k!} .$$

De kans dat $\underline{k}$ gelijk is aan 0 is hetzelfde als de kans dat $\underline{t}$, de tijd tussen twee gebeurtenissen, groter is dan t:

$$P(\underline{t} > t) = 1 - F(t) = P(\underline{k} = 0) = e^{-t/\beta} .$$

$$F(t) = 1 - e^{-t/\beta} ,$$

$$f(t) = \frac{dF(t)}{dt} = 1/\beta \; e^{-t/\beta} .$$

Met partiële integratie kan men afleiden:

$$\int_0^x f(t;\alpha,\beta)dt = 1 - \sum_{k=0}^{\alpha} \frac{e^{-x/\beta}(x/\beta)^k}{k!} .$$

De <u>beta verdeling</u> is

$$(2.6) \qquad f(x;\alpha_1,\alpha_2) = \frac{(\alpha_1+\alpha_2+1)!}{\alpha_1!\alpha_2!} x^{\alpha_1}(1-x)^{\alpha_2} , \quad 0 \leq x \leq 1, \; \alpha_1 > -1, \; \alpha_2 > -1.$$

Het bewijs dat het oppervlak onder deze verdeling 1 is, is het gemakkelijkst

te geven met behulp van de gamma verdeling.
Stel $\underline{x}$ en $\underline{y}$ zijn onafhankelijk verdeeld volgens gamma verdelingen met parameters (α_1,β) en (α_2,β). De som $\underline{z}$ heeft dan, zoals werd aangetoond, opnieuw een gamma verdeling met parameters $\alpha_1 + \alpha_2 + 1$ en β. Dus er geldt:

$$\frac{1}{(\alpha_1+\alpha_2+1)!\beta^{\alpha_1+\alpha_2+2}} z^{\alpha_1+\alpha_2+1} e^{-z/\beta} =$$

$$= \frac{1}{\alpha_1!\alpha_2!\beta^{\alpha_1+\alpha_2+2}} \int_0^z (z-y)^{\alpha_1} y^{\alpha_2} e^{-z/\beta} dy .$$

Of:

$$\int_0^z (z-y)^{\alpha_1} y^{\alpha_2} dy = \frac{\alpha_1!\alpha_2!}{(\alpha_1+\alpha_2+1)!} z^{\alpha_1+\alpha_2+1} .$$

Stellen we nu $u = y/z$, dan volgt

$$\int_0^1 (1-u)^{\alpha_1} u^{\alpha_2} du = \frac{\alpha_1!\alpha_2!}{(\alpha_1+\alpha_2+1)!} ,$$

hetgeen te bewijzen was.

De momenten kunnen direct worden gevonden:

$$\mu_r' = \frac{(\alpha_1+\alpha_2+1)!}{\alpha_1!\alpha_2!} \int_0^1 x^{\alpha_1+r}(1-x)^{\alpha_2} dx = \frac{(\alpha_1+\alpha_2+1)!(\alpha_1+r)!}{(\alpha_1+\alpha_2+r+1)!\,\alpha_1!} .$$

In het bijzonder

$$\mu = \mu_1' = \frac{\alpha_1+1}{\alpha_1+\alpha_2+2} \qquad \sigma^2 = \mu_2' - (\mu_1')^2 = \frac{(\alpha_1+1)(\alpha_2+1)}{(\alpha_1+\alpha_2+2)^2(\alpha_1+\alpha_2+3)} .$$

De <u>verdeling van Cauchy</u>

$$(2.7) \qquad f(x) = \frac{1}{\pi}\frac{1}{1+(x-\mu)^2} , \quad -\infty \leq x \leq \infty$$

wordt vaak gebruikt ter illustratie van bepaalde anomalieën.
De cumulatieve verdeling is

$$F(x) = \frac{1}{\pi}\int_{-\infty}^{x} \frac{dt}{1+(t-\mu)^2} = \tfrac{1}{2} + \frac{1}{\pi}\,\mathrm{arctg}(x-\mu) .$$

Het gemiddelde bestaat alleen in beperkte zin, nl. als

$$\lim_{A\to\infty} \frac{1}{\pi} \int_{\mu-A}^{\mu+A} \frac{x\,dx}{1+(x-\mu)^2} = \mu .$$

Hogere momenten bestaan niet.

We zeggen dat $\underline{x}$ de <u>log normale verdeling</u> heeft als $\underline{y} = \ln \underline{x}$ normaal verdeeld is.

Dus

$$f(y) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} ,$$

(2.8) $$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} , \quad 0 \leqslant x \leqslant \infty .$$

De verwachting en de variantie zijn als volgt te vinden:

$$\mu = \frac{1}{\sqrt{2\pi}\,\sigma} \int_0^\infty e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} dx = (x = e^y).$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^\infty e^{y-\frac{(y-\mu)^2}{2\sigma^2}} dy = (\text{vgl. p.6, } m(t))$$

$$= m(1) = e^{\mu+\frac{1}{2}\sigma^2} .$$

$$E\underline{x}^2 = m(2) = e^{2\mu+2\sigma^2} .$$

$$\sigma^2 = E\underline{x}^2 - \mu^2 = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = e^{2\mu+\sigma^2}(e^{\sigma^2}-1) .$$

<u>Vraagstukken</u>:

1) De momenten van $\underline{x}$ zijn $\mu_r' = r!$
   Wat is de momenten voortbrengende functie en wat is de verdeling?

2) $\underline{x}$ is homogeen verdeeld tussen 0 en 1. Welke functie van $\underline{x}$ heeft gamma verdeling met $\alpha = 0$ en $\beta = 1$?

3) $\underline{x}$ heeft beta verdeling met $\alpha_1 = 0$, $\alpha_2 = 1$. Welke transformatie geeft gamma verdeling met $\alpha = 0$, $\beta = 1$?

4) $\mu_r' = \frac{r!}{(r/2)!}$ voor r even en $\mu_r' = 0$ voor r oneven. Wat is de momenten voortbrengende functie en wat de verdeling?

Wij bespreken nu een eigenschap van kansverdelingen die later bij de schattings- en de toetsingstheorie kan worden toegepast. Dit is het begrip <u>volledigheid</u>. Wij beschouwen een kansdichtheid $f(x;\theta)$, waarbij θ een parameter is. De verdeling mag eventueel meerdimensionaal zijn zodat x wordt vervangen door $x_1,\dots,x_n$ en er mogen meer dan één parameters θ_i zijn.

(2.9) <u>Definitie</u>: De klasse van verdelingen $f(x;\theta)$ heet <u>volledig</u> als aan de volgende voorwaarden is voldaan:

(1) $f(x;\theta) > 0$ voor $a < x < b$
$= 0$ voor $x \leqslant a$ of $x \geqslant b$.

(2) $\alpha_0 < \theta < \alpha_1$.

(3) a en b onafhankelijk van θ.

(4) Voor iedere functie $u(x)$ die continu is in het interval $a < x < b$ volgt uit: $E\{u(\underline{x})\} = 0$ voor alle θ in het interval (2) dat $u(x) \equiv 0$ voor $a < x < b$.

Als $f(x;\theta)$ een discrete kansverdeling is wordt met (1) bedoeld: $f(x;\theta) > 0$ voor alle gehele waarden van x tussen a en b. Evenzo wordt in dat geval in (4) met $u(x) \equiv 0$ bedoeld: $u(x) = 0$ voor alle gehele waarden van x tussen a en b.

Volledigheid is vaak moeilijk te bewijzen. Wanneer een klasse van verdelingen niet volledig is kan dit soms door een voorbeeld worden aangetoond. Immers het is voldoende één functie $u(x)$ te vinden die niet identiek gelijk is aan nul en waarvoor geldt: $E\{u(\underline{x})\} = 0$.

<u>Voorbeelden</u>:

1) $f(x;\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$, $\quad -\infty < x < \infty$
$0 < \sigma^2 < \infty$

is niet volledig, want $E\underline{x} = 0$ voor elke σ^2.

2) $f(x;p) = p^x(1-p)^{1-x}$, $\quad x = 0,1$
$\alpha_0 \leqslant p \leqslant \alpha_1,\ 0 \leqslant \alpha_0 < \alpha_1 \leqslant 1$

is wel volledig. Stel nl. $E\{u(\underline{x})\} = 0$. Dus $u(0)(1-p) + u(1)p = 0$, voor elke p. Hieruit volgt: $u(0) = 0$ en $u(1) = 0$.

Nu enkele verdelingen waarvan we zonder bewijs vermelden dat ze volledig zijn:

3) $f(x;\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$, $\quad - \infty < u < \infty$

$\alpha_0 < \mu < \alpha_1$.

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

$\alpha_0 < \mu < \alpha_1$

$0 < \beta_0 < \sigma^2 < \beta_1$.

5) $f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$, $\quad x = 0,1,2,\ldots$

$0 < \alpha_0 < \lambda < \alpha_1$.

Niet volledig is:

6) $P(\underline{x} = -1) = \theta$

$P(\underline{x} = x) = (1-\theta)^2\theta^x$, $\quad x = 0,1,2,\ldots$

$0 < \theta < 1$.

Dit kan worden aangetoond door te laten zien dat $E\underline{x} = 0$ voor elke θ.

Hoofdstuk III, Steekproeven

We noemen hier in het kort enige begrippen en stellingen die grotendeels reeds bij het college IVb ter sprake zijn gekomen.

(3.1) Definitie: De stochastische variabelen $\underline{x}_1,\dots,\underline{x}_n$ vormen een aselecte steekproef uit de populatie met verdeling $f(x)$ als de simultane verdeling van $\underline{x}_1,\dots,\underline{x}_n$ gelijk is aan

$$g(x_1,\dots,x_n) = f(x_1)f(x_2)\dots f(x_n) .$$

(3.2) Definitie: Een statistische grootheid (Engels: statistic) is een functie van de elementen van een steekproef, die geen onbekende parameters bevat.

Voorbeelden:

Als $\underline{x}$ normaal verdeeld is met onbekend gemiddelde μ en onbekende standaardafwijking σ, dan zijn $\underline{x}-\mu$ en $\underline{x}/\sigma$ geen statistische grootheden, maar $\underline{x}$, $\underline{x}+3$ en $\log \underline{x}$ wel. Ook het steekproefgemiddelde $\overline{\underline{x}} = \frac{1}{n}\Sigma\, \underline{x}_i$ is een statistische grootheid.

(3.3) Definitie: Als $\underline{x}_1,\dots,\underline{x}_n$ een aselecte steekproef is uit de verdeling $f(x)$ dan is

$$\underline{m}_r' = \frac{1}{n}\sum_1^n \underline{x}_i^r$$

het r^e steekproefmoment.

(3.4) Stelling: $E\underline{m}_r' = \mu_r'$.

Bewijs dit zelf.

(3.5) Stelling: $\operatorname{var} \overline{\underline{x}} = \sigma^2/n$,

als $\overline{\underline{x}} = \frac{1}{n}\sum_1^n \underline{x}_i$ het steekproefgemiddelde is en σ^2 de variantie van de verdeling.

(Wiskunde IVb).

(3.6) Ongelijkheid van Bienaymé-Cebysev (Wiskunde IVb)

$$P(|\underline{x} - \mu| \geq a) \leq \frac{\sigma^2}{a^2} \quad ,$$

als μ en σ^2 het gemiddelde en de variantie van $\underline{x}$ voorstellen.

(3.7) Zwakke wet van de grote aantallen

Bij elke $\varepsilon > 0$ en $\delta > 0$ is er een n aan te wijzen zodat

$$P(|\bar{\underline{x}}_m - \mu| < \varepsilon) > 1 - \delta$$

voor alle $m \geq n$, als $\bar{\underline{x}}_m$ het steekproefgemiddelde is van een steekproef ter grootte m.
Bewijs dit zelf.

(3.8) Sterke wet van de grote aantallen

$$P(\bar{\underline{x}}_n \to \mu) = 1 \; .$$

Dit wil zeggen dat bij $\varepsilon > 0$ en $\delta > 0$ een m kan worden gevonden zodat de kans, dat de ν ongelijkheden

$$|\bar{\underline{x}}_{m+k} - \mu| < \varepsilon \,, \qquad \text{voor } k = 1,2,\ldots,\nu$$

simultaan gelden, voor elke ν groter is dan $1 - \delta$.
Wij bewijzen deze stelling hier niet.

(3.9) Centrale limietstelling

Onder vrij zwakke voorwaarden geldt:
Als $\underline{x}_1,\ldots,\underline{x}_n$ onderling onafhankelijke variabelen zijn met gemiddelden $\mu_1,\ldots,\mu_n$ en standaardafwijkingen $\sigma_1,\ldots,\sigma_n$, dan nadert de verdeling van

$$\underline{y} = \frac{\Sigma\, \underline{x}_i - \Sigma\, \mu_i}{\sqrt{\Sigma\, \sigma_i^2}} \quad ,$$

voor $n \to \infty$ tot de normale verdeling met gemiddelde 0 en standaardafwijking 1.

Wij schetsen hier alleen een bewijs voor het geval $\underline{x}_1,\ldots,\underline{x}_n$ een steekproef vormen uit één verdeling met gemiddelde μ en standaardafwijking σ en waarvan de momenten voortbrengende functie $m_1(t)$ bestaat.

Noem de momenten voortbrengende functie van $\dfrac{\underline{x}_i - \mu}{\sigma}$ $m_2(t)$. De momenten voortbrengende functie van $\dfrac{\Sigma(\underline{x}_i - \mu)}{\sigma}$ is dan

$$Ee^{t\frac{\Sigma(\underline{x}_i-\mu)}{\sigma}} = \{m_2(t)\}^n .$$

De momenten voortbrengende functie van $\underline{y} = \dfrac{\Sigma(\underline{x}_i - \mu)}{\sigma\sqrt{n}}$ is dan

$$m_3(t) = \left\{m_2\left(\frac{t}{\sqrt{n}}\right)\right\}^n = \left\{1 + \frac{\mu_1}{1!\sigma}\frac{t}{\sqrt{n}} + \frac{\mu_2}{2!\sigma^2}\left(\frac{t}{\sqrt{n}}\right)^2 + \frac{\mu_3}{3!\sigma^3}\left(\frac{t}{\sqrt{n}}\right)^3 + \ldots\right\}^n .$$

Maar $\mu_1 = 0$ en $\mu_2 = \sigma^2$, dus:

$$m_3(t) = \left\{1 + \frac{1}{n}\left(\tfrac{1}{2}t^2 + \frac{\mu_3}{3!\sigma^3}\frac{t^3}{\sqrt{n}} + \ldots\right)\right\}^n$$

Dus

$$\lim_{n\to\infty} m_3(t) = \lim_{n\to\infty}\left(1 + \frac{\frac{1}{2}t^2}{n}\right)^n = e^{\frac{1}{2}t^2} .$$

Dit is de momenten voortbrengende functie van de normaal verdeelde stochastische variabele met gemiddelde 0 en standaardafwijking 1, hetgeen te bewijzen was.

Uit de centrale limietstelling volgt bv. dat een grootheid $\underline{k}$ die een binomiale verdeling heeft met parameters n en p voor grote n bij benadering normaal is verdeeld. Immers $\underline{k}$ kan worden opgevat als de som van n onderling onafhankelijke stochastische grootheden met een alternatieve verdeling:
$f(x) = p^x(1-p)^{1-x}$ $(x = 0,1)$. De som

$$P(a \leq \underline{k} \leq b) = \sum_{k=a}^{b} \binom{n}{k} p^k(1-p)^{n-k}$$

wordt dan benaderd door de integraal

$$\frac{1}{\sqrt{2\pi}}\int_{a'}^{b'} e^{-\frac{1}{2}t^2}\,dt ,$$

waarin

$$a' = \frac{a - \frac{1}{2} - np}{\sqrt{npq}} \quad \text{en} \quad b' = \frac{b + \frac{1}{2} - np}{\sqrt{npq}} \; .$$

De zogenaamde continuïteitscorrecties $\pm\frac{1}{2}$ treden op omdat $\underline{k}$ alleen gehele waarden aan kan nemen waardoor de kans op de waarde a het best wordt benaderd door de integraal van $a - \frac{1}{2}$ tot $a + \frac{1}{2}$.

Vraagstukken:

1) Een bureau voor opinieonderzoek wil een steekproef van kiezers nemen die zo groot is dat de kans, om voor een zekere partij minder dan 50% van de stemmen te vinden als dit percentage in werkelijkheid 52% is, slechts 0,01 bedraagt. Wat is de minimale steekproefgrootte?

2) $\underline{\bar{x}}_1$ en $\underline{\bar{x}}_2$ zijn gemiddelden van twee steekproeven ter grootte n uit een populatie met variantie σ^2. Bepaal n zo dat de kans dat de beide gemiddelden meer dan σ verschillen kleiner is dan 0,01.

Hoofdstuk IV, Puntschattingen

Alvorens de theorie van de puntschattingen te behandelen bespreken we eerst het algemene beslissingsprobleem. Stel dat een beslissing moet worden genomen uit een aantal mogelijke beslissingen. De juiste beslissing hangt af van een onbekende parameter θ die de kansverdeling $f(x;\theta)$ bepaalt. Als θ bekend zou zijn zou de juiste beslissing ook bekend zijn. Stel Ω is de verzameling van mogelijke parameterwaarden θ en A is de verzameling van mogelijke beslissingen a. Om op grond van een aselecte steekproef $\underline{x}_1,\dots,\underline{x}_n$ van $f(x;\theta)$ een beslissing a uit A te kunnen kiezen moeten we beschikken over een beslissingsfunctie of strategie

(4.1) $\quad \underline{a} = d(\underline{x}_1,\dots,\underline{x}_n)$,

die aan iedere realisatie $x_1,\dots,x_n$ een $a \in A$ toevoegt.

Voorbeelden:

1) Stel $f(x;\theta)$ is de normale verdeling met gemiddelde μ en standaardafwijking 1. Dus θ is de verwachting μ. De parameterruimte Ω kan bv. de reële rechte zijn: $\Omega = \{\mu \mid -\infty < \mu < \infty\}$. Stel verder dat bij iedere μ een andere beslissing hoort. Dan kunnen we een beslissing voorstellen door $\hat{\mu}$, een schatting van μ en $A = \{\hat{\mu} \mid -\infty < \hat{\mu} < \infty\}$. Als beslissingsfunctie kan genomen worden:

$$\underline{\hat{\mu}} = d(\underline{x}_1,\dots,\underline{x}_n) = \frac{1}{n}\sum_1^n \underline{x}_i .$$

Dit is een voorbeeld van een puntschatting.

2) We gaan uit van dezelfde verdeling als in het vorige voorbeeld, maar nu zijn er maar 2 beslissingen, nl. $a_1: \mu < 0$ en $a_2: \mu \geq 0$. Dus $A = \{a_1,a_2\}$. De beslissingsfunctie is nu bv.

$$\begin{aligned} d(\underline{x}_1,\dots,\underline{x}_n) &= a_1 \text{ als } \underline{\bar{x}} < 0 \\ &= a_2 \text{ als } \underline{\bar{x}} \geq 0 . \end{aligned}$$

Dit is een voorbeeld van het toetsen van een hypothese.

In beide voorbeelden kunnen meerdere beslissingsfuncties worden gekozen. Hoe een "goede" beslissingsfunctie kan worden gekozen wordt later besproken.

Om de gevolgen van het kiezen van een bepaalde beslissingsfunctie te kunnen waarderen voeren wij een verliesfunctie in.

(4.2) Definitie: Een verliesfunctie $\ell(a;\theta)$ is een reële functie van a en θ die voldoet aan:

1) $\ell(a;\theta) \geq 0$ voor alle $a \in A$ en $\theta \in \Omega$.
2) Bij iedere $\theta \in \Omega$ is er minstens één $a \in A$ waarvoor $\ell(a;\theta) = 0$.

De waarden van a waarvoor $\ell(a;\theta) = 0$ heten de juiste beslissingen bij de parameterwaarde θ. Omdat $\underline{a}$ van de steekproefwaarden $\underline{x}_1,\dots,\underline{x}_n$ afhangt, heeft de verliesfunctie ook een stochastische uitkomst.
Om toch bij iedere θ één getal aan een beslissingsfunctie toe te kunnen kennen, beschouwen wij de risicofunctie, dit is de verwachting van de verliesfunctie.

(4.3) Definitie: De risicofunctie $R(d;\theta)$ wordt gedefinieerd door

$$R(d;\theta) = E[\ell(\underline{a}\,;\theta)] \ .$$

Voorbeeld:
Stel $f(x;\theta) = N(x;\theta,1)$, de normale verdeling met gemiddelde θ en standaardafwijking 1. Stel verder $\ell(a;\theta) = (a-\theta)^2$. Als we als schatter $\underline{a}$ nemen het steekproefgemiddelde $\underline{\bar{x}}$, dan is

$$R(d;\theta) = E(\underline{\bar{x}}-\theta)^2 = \frac{1}{n} \ .$$

Bij puntschattingen, waar $\hat{\theta}$ een schatting is voor θ, is $\ell(\hat{\theta};\theta) = 0$ dan en slechts dan als $\hat{\theta} = \theta$. De beslissingsfunctie

$$\underline{\hat{\theta}} = d(\underline{x}_1,\dots,\underline{x}_n)$$

heet dan een schatter voor θ. De verliesfunctie is dan meestal van de vorm

$$\ell(\hat{\theta};\theta) = c(\theta)(\hat{\theta}-\theta)^2 \ , \qquad (4.4)$$

waarbij $c(\theta) > 0$ voor alle θ. Voor het vergelijken van twee risicofuncties behorend bij verschillende schatters doet de vorm van $c(\theta)$ niet ter zake, omdat gemakkelijk is in te zien dat volgens (4.4)

$$R(d_1;\theta) > R(d_2;\theta)$$

equivalent is met

$$E(\underline{\hat{\theta}}_1-\theta)^2 > E(\underline{\hat{\theta}}_2-\theta)^2 \ ,$$

als $\underline{\hat{\theta}}_1 = d_1(\underline{x}_1,\dots,\underline{x}_n)$ en $\underline{\hat{\theta}}_2 = d_2(\underline{x}_1,\dots,\underline{x}_n)$ twee verschillende schatters voorstellen.

De <u>relatieve doeltreffendheid</u> (relative efficiency) van d_1 ten opzichte van d_2 wordt gedefinieerd door

$$(4.5) \qquad r(d_1,d_2;\theta) = \frac{R(d_2;\theta)}{R(d_1;\theta)} \ .$$

De grootheid $E(\underline{\hat{\theta}}-\theta)^2$ heet in het Engels: <u>mean-squared error</u>. In de Nederlandse terminologie heet de wortel hieruit de <u>onnauwkeurigheid</u>. Als $E\underline{\hat{\theta}} = \theta'$, dan geldt:

$$(4.6) \qquad E(\underline{\hat{\theta}}-\theta)^2 = E(\underline{\hat{\theta}}-\theta')^2 + (\theta'-\theta)^2 \ .$$

De afwijking $(\theta'-\theta)$ van de verwachting van $\underline{\hat{\theta}}$ van de parameterwaarde heet de <u>onzuiverheid</u> (bias). In woorden luidt (4.6) dus: Het kwadraat van de onnauwkeurigheid van een schatter is de som van de variantie en van het kwadraat van de onzuiverheid van de schatter.
In het algemeen zal men zoeken naar een schatter met de kleinste onnauwkeurigheid. Behalve in triviale gevallen is er echter geen schatter te vinden die de onnauwkeurigheid minimaliseert voor elke $\theta \in \Omega$.

(4.7) <u>Definitie</u>: Stel $\underline{x}_1,\ldots,\underline{x}_n$ is een aselecte steekproef uit de verdeling $f(x;\theta)$ en $\underline{\hat{\theta}} = d(\underline{x}_1,\ldots,\underline{x}_n)$ is een statistische grootheid. Stel $\underline{\theta}^* = d^*(\underline{x}_1,\ldots,\underline{x}_n)$ is een willekeurige andere statistische grootheid. Als de verdeling van $\underline{\theta}^*$, onder de voorwaarde $\underline{\hat{\theta}} = \hat{\theta}$ de parameter θ niet bevat, dan heet $\underline{\hat{\theta}}$ <u>voldoende</u> voor θ.

Een eenvoudig criterium om na te gaan of een statistische grootheid $\underline{\hat{\theta}}$ voldoende is voor θ wordt gegeven door de volgende stelling:

(4.8) <u>Stelling</u>: Stel $\underline{x}_1,\ldots,\underline{x}_n$ vormen een aselecte steekproef uit de verdeling $f(x;\theta)$, $a < x < b$, waarin a en b niet afhangen van θ. Dan en slechts dan als de simultane verdeling van $\underline{x}_1,\ldots,\underline{x}_n$ kan worden ontbonden als

$$(4.9) \qquad g(x_1,\ldots,x_n;\theta) = h(\hat{\theta};\theta)k(x_1,\ldots,x_n) \ ,$$

waarbij $k(x_1,\ldots,x_n)$ de parameter θ niet bevat, is $\hat{\theta}$ voldoende voor θ.

<u>Bewijs</u>: Wij bewijzen eerst dat uit (4.9) volgt dat $\hat{\theta}$ voldoende is. Daartoe voeren wij de volgende transformatie uit:

$$\begin{cases} \hat{\theta} = \hat{\theta}(x_1,\ldots,x_n) \\ \theta^* = \theta^*(x_1,\ldots,x_n) \\ x_3 = x_3 \\ \cdot \\ \cdot \\ x_n = x_n \,. \end{cases}$$

De determinant van Jacobi bevat de parameter θ niet, omdat $\hat{\theta}$ en θ^* statistische grootheden zijn, die dus alleen afhangen van $(x_1,\ldots,x_n)$. De simultane verdeling van $\hat{\theta}$, θ^*, $x_3,\ldots,x_n$ wordt dus gevonden uit het rechterlid van (4.9) en heeft de gedaante:

$$h(\hat{\theta};\theta)\ell(\hat{\theta},\theta^*,x_3,\ldots,x_n) \,.$$

De simultane verdeling van $\hat{\theta}$ en θ^* wordt gevonden door over $x_3,\ldots,x_n$ te integreren. Omdat de integratiegrenzen niet van θ afhangen (zie gegeven) wordt hierdoor geen θ in de functie $\ell(\hat{\theta},\theta^*,x_3,\ldots,x_n)$ ingevoerd. Wij vinden dus voor de simultane verdeling van $\hat{\theta}$ en θ^*:

$$f(\hat{\theta},\theta^*;\theta) = h(\hat{\theta};\theta)p(\theta^*,\hat{\theta}) \,,$$

waarbij p geen θ bevat. Afgezien van constante factoren is nu echter $h(\hat{\theta};\theta)$ de verdeling van $\underline{\hat{\theta}}$ en $p(\theta^*,\hat{\theta})$ de voorwaardelijke verdeling van $\underline{\theta}^*$, onder de voorwaarde $\underline{\hat{\theta}} = \hat{\theta}$. Deze laatste verdeling bevat geen θ en dus is $\underline{\hat{\theta}}$ voldoende.

Om te bewijzen dat uit het feit dat $\underline{\hat{\theta}}$ voldoende is ook de ontbinding (4.9) volgt, bedenken wij eerst het volgende. De verdeling van $\underline{\theta}^*$, onder de voorwaarde $\underline{\hat{\theta}} = \hat{\theta}$ is onafhankelijk van θ. Dit geldt voor iedere statistische grootheid $\underline{\theta}^*$, dus ook voor de statistische grootheden $\underline{x}_2,\ldots,\underline{x}_n$ simultaan. Dus de simultane verdeling van $\underline{\hat{\theta}}$, $\underline{x}_2,\ldots,\underline{x}_n$ is als volgt te schrijven:

$$f(\hat{\theta},x_2,\ldots,x_n;\theta) = h(\hat{\theta};\theta)p(x_2,\ldots,x_n \mid \underline{\hat{\theta}} = \hat{\theta}) \,,$$

waarbij in de functie p geen θ voorkomt. Voer nu de transformatie uit:

$$\begin{cases} x_1 = x_1(\hat{\theta},x_2,\ldots,x_n) \\ x_2 = x_2 \\ \cdot \\ \cdot \\ x_n = x_n \,, \end{cases}$$

waarna onmiddellijk (4.9) volgt.

(4.10) Stelling: Als $\hat{\underline{\theta}}$ voldoende is voor θ, dan is $\tilde{\underline{\theta}} = u(\hat{\underline{\theta}})$ ook voldoende voor θ, als de inverse functie v van u éénwaardig is en bovendien is $\hat{\underline{\theta}}$ voldoende voor $u(\theta)$.

Bewijs dit zelf.

Voorbeelden:

1) $f(x;\mu) = N(x;\mu,1)$, normale verdeling, verwachting μ, standaardafwijking 1. De simultane verdeling van een steekproef $\underline{x}_1,\dots,\underline{x}_n$ is:

$$g(x_1,\dots,x_n;\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\Sigma(x_i-\mu)^2} =$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{n}{2}(\bar{x}-\mu)^2} e^{-\frac{1}{2}\Sigma(x_i-\bar{x})^2} .$$

Dit is van de gedaante (4.9) en dus is $\bar{\underline{x}}$ voldoende voor μ.

2) $f(x;\sigma^2) = N(x;0,\sigma^2)$.

$$g(x_1,\dots,x_n;\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\Sigma x_i^2} .$$

Dit is weer van de gedaante (4.9) met $k(x_1,\dots,x_n) = 1$. Dus $\Sigma\, \underline{x}_i^2$ is voldoende voor σ^2 (en volgens stelling (4.10) ook voor σ.

(4.11) Definitie: Stel $\underline{x}_1,\dots,\underline{x}_n$ vormen een aselecte steekproef uit de verdeling $f(x;\theta_1,\dots,\theta_k)$. De m statistische grootheden

$$\hat{\underline{\theta}}_1 = d_1(\underline{x}_1,\dots,\underline{x}_n)$$
$$\hat{\underline{\theta}}_2 = d_2(\underline{x}_1,\dots,\underline{x}_n)$$
$$\cdot$$
$$\cdot$$
$$\hat{\underline{\theta}}_m = d_m(\underline{x}_1,\dots,\underline{x}_n)$$

vormen een verzameling van m voldoende statistische grootheden voor de parameters $\theta_1,\dots,\theta_k$ als de verdeling van $\underline{x}_1,\dots,\underline{x}_n$ onder de voorwaarde $\hat{\underline{\theta}}_i = \theta_i$ $(i = 1,\dots,m)$ niet afhangt van $\theta_1,\dots,\theta_k$.

Opmerking:

1) Als $k = 1$ is deze definitie gelijkwaardig met definitie (4.7), zoals volgt uit het bewijs van stelling (4.8).
2) Meestal zal $m = k$ zijn. Als $m > k$ mag zijn, zijn er veel verzamelingen van voldoende statistische grootheden, bijv. de steekproef $\underline{x}_1,\dots,\underline{x}_n$ zelf. Interessant zijn minimale verzamelingen, d.w.z. dat zo'n verzameling een functie is van elke andere verzameling van voldoende statistische grootheden.

(4.12) Stelling: $\underline{\hat\theta}_1,\dots,\underline{\hat\theta}_m$ zijn dan en slechts dan voldoende voor $\theta_1,\dots,\theta_k$, als de simultane verdeling van $\underline{x}_1,\dots,\underline{x}_n$ kan worden ontbonden in

$$g(x_1,\dots,x_n;\theta_1,\dots,\theta_k) = h(\hat\theta_1,\dots,\hat\theta_m;\theta_1,\dots,\theta_k)g(x_1,\dots,x_n).$$

Het bewijs verloopt precies zo als dat van stelling (4.8).

(4.13) Stelling: Stel $\underline{\hat\theta}_1,\dots,\underline{\hat\theta}_m$ zijn voldoende. Als een één-éénduidige transformatie deze statistische grootheden overvoert in $\hat\alpha_1,\dots,\hat\alpha_m$, dan is de verzameling $\hat\alpha_1,\dots,\hat\alpha_m$ ook voldoende.

Voorbeeld:

$f(x;\mu,\sigma^2) = N(x;\mu,\sigma^2)$, normale verdeling met gemiddelde μ en variantie σ^2.

$$g(x_1,\dots,x_n;\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2} =$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\bar x)^2}\, e^{-\frac{n}{2\sigma^2}(\bar x-\mu)^2}.$$

Dus $\Sigma(\underline{x}_i - \bar{\underline{x}})^2$ en $\bar{\underline{x}}$ zijn voldoende voor μ en σ^2. Maar ook (op grond van stelling (1.13)): $\Sigma\,\underline{x}_i$ en $\bar{\underline{x}}$ of: $\frac{1}{n-1}\Sigma(\underline{x}_i - \bar{\underline{x}})^2$ en $\bar{\underline{x}}$.

Wij gaan nu een aantal eigenschappen van schatters definiëren, en wel achtereenvolgens:

Zuiver (unbiased)

Asymptotisch raak (consistent)

Asymptotisch doeltreffend (efficient)

Nauwkeurigst zuiver (minimum variance unbiased).

(4.14) Definitie: Een schatter $\underline{\hat{\theta}}$ wordt een zuivere schatter voor θ als

$$E(\underline{\hat{\theta}}) = \theta, \quad \text{voor alle } \theta \in \Omega.$$

(4.15) Definitie: Een rij schatters $\{\underline{\hat{\theta}}_n\}$ voor θ heet asymptotisch raak, als

$$\lim_{n\to\infty} P(\theta-\varepsilon < \underline{\hat{\theta}}_n < \theta+\varepsilon) = 1 \quad \text{voor elke } \theta \in \Omega \text{ en voor elke } \varepsilon > 0.$$

Een sterkere eis is de volgende:

(4.16) Definitie: Een rij schatters $\{\underline{\hat{\theta}}_n\}$ voor θ heet asymptotisch nauwkeurig (squared-error consistent) als

$$\lim_{n\to\infty} E(\underline{\hat{\theta}}_n - \theta)^2 = 0 \quad \text{voor alle } \theta \in \Omega.$$

Dat (4.16) sterker is dan (4.15) is als volgt in te zien:

$$E(\underline{\hat{\theta}}_n - \theta)^2 = E(\underline{\hat{\theta}}_n - E(\underline{\hat{\theta}}_n))^2 + (E(\underline{\hat{\theta}}_n) - \theta)^2 \quad (\text{vgl. } (4.6)).$$

Uit $\lim_{n\to\infty} E(\underline{\hat{\theta}}_n - \theta)^2 = 0$ volgt dus dat:

$$\lim_{n\to\infty} \operatorname{var}(\underline{\hat{\theta}}_n) = 0 \quad \text{en} \quad \lim_{n\to\infty} E(\underline{\hat{\theta}}_n) = \theta.$$

Met behulp van de ongelijkheid van Bienaymé-Cebysev (3.6) kan dan het asymptotische raak zijn worden bewezen.

(4.17) Definitie: Een rij schatters $\{\underline{\hat{\theta}}_n\}$ heet asymptotisch doeltreffend (squared-error asymptotically efficient of efficient) als $\{\underline{\hat{\theta}}_n\}$ asymptotisch nauwkeurig is en er geen andere asymptotisch nauwkeurige rij $\{\underline{\theta}^*_n\}$ bestaat waarvoor

$$\limsup \frac{E(\underline{\hat{\theta}}_n - \theta)^2}{E(\underline{\theta}^*_n - \theta)^2} > 1 \quad \text{voor alle } \theta \text{ in enig open interval.}$$

(4.18) Definitie: Een rij schatters $\{\underline{\hat{\theta}}_n\}$ heet best asymptotisch normaal (BAN) als voldaan is aan

a) De verdeling van $\sqrt{n}(\underline{\hat{\theta}}_n - \theta)$ nadert voor $n \to \infty$ tot de normaal met gemiddelde 0 en variantie $\sigma^2(\theta)$.

b) $\{\underline{\hat{\theta}}_n\}$ is asymptotisch raak.

c) Er bestaat geen rij $\{\underline{\theta}^*_n\}$ die aan a) en b) voldoet met limietwaarde $\sigma^{*2}(\theta)$ voor de variantie, waarvoor

$$\frac{\sigma^2(\theta)}{\sigma^{*2}(\theta)} > 1 \quad \text{voor alle } \theta \text{ in enig open interval.}$$

(4.19) Definitie: Een schatter $\hat{\underline{\theta}}$ heet de nauwkeurigste zuivere schatter voor θ als

a) $E\hat{\underline{\theta}} = \theta$.

b) $\mathrm{var}(\hat{\underline{\theta}})$ is kleiner dan de variantie van elke andere zuivere schatter.

(4.20) Stelling van Rao en Blackwell: Stel $\underline{x}_1,\dots,\underline{x}_n$ is een aselecte steekproef uit $f(x;\theta)$ en $\hat{\underline{\theta}} = d(\underline{x}_1,\dots,\underline{x}_n)$ is voldoende voor θ. Stel $\underline{t} = t(\underline{x}_1,\dots,\underline{x}_n)$ is een zuivere schatter voor $u(\theta)$. Noem de voorwaardelijke verwachting van $\underline{t}$, onder de voorwaarde $\hat{\underline{\theta}}$, $v(\hat{\underline{\theta}})$, dus $E(\underline{t} \mid \hat{\underline{\theta}}) = v(\hat{\underline{\theta}})$. Dan geldt:

a) $E[v(\hat{\underline{\theta}})] = u(\theta)$.

b) $\mathrm{var}[v(\hat{\underline{\theta}})] < \mathrm{var}(\underline{t})$.

c) $v(\hat{\underline{\theta}})$ is een statistische grootheid.

Bewijs: We merken eerst op dat het geval moet worden uitgezonderd dat $\underline{t}$ een functie van $\hat{\underline{\theta}}$ is. Dan is nl. $E(\underline{t} \mid \hat{\underline{\theta}}) = v(\hat{\underline{\theta}}) = \underline{t}$ en in b) geldt dan het = teken.
Dat c) geldt volgt uit het feit dat $\hat{\underline{\theta}}$ voldoende is. Dat wil zeggen dat de verdeling van $\underline{t}$, onder de voorwaarde $\hat{\theta}$ niet van θ afhangt, dus $p(t \mid \hat{\theta})$ bevat θ niet en de verwachting $E(\underline{t} \mid \hat{\underline{\theta}})$ dus evenmin.
Vervolgens bewijzen we a). De simultane verdeling van $\underline{t}$ en $\hat{\underline{\theta}}$ is

$$g(t,\hat{\theta};\theta) = h(\hat{\theta};\theta)p(t \mid \hat{\theta}) .$$

Dus:

$$Ev(\hat{\underline{\theta}}) = \int\Big[\int tp(t \mid \hat{\theta})dt\Big]h(\hat{\theta};\theta)d\hat{\theta} = \iint tg(t,\hat{\theta};\theta)dtd\hat{\theta} = E\underline{t} = u(\theta) .$$

Om b) te bewijzen gaan we uit van $\mathrm{var}(\underline{t})$:

$$\mathrm{var}(\underline{t}) = E[\underline{t} - u(\theta)]^2 = E[\{\underline{t} - v(\hat{\underline{\theta}})\} + \{v(\hat{\underline{\theta}}) - u(\theta)\}]^2 =$$

$$= E[\underline{t} - v(\hat{\underline{\theta}})]^2 + 2E[(\underline{t} - v(\hat{\underline{\theta}})][v(\hat{\underline{\theta}}) - u(\theta)] + \mathrm{var}[v(\hat{\underline{\theta}})] .$$

De tweede term is gelijk aan:

$$\iint [t - v(\hat{\theta})][v(\hat{\theta}) - u(\theta)]h(\hat{\theta};\theta)p(t \mid \hat{\theta})dtd\hat{\theta} .$$

Maar $\int (t - v(\hat{\theta}))p(t \mid \hat{\theta})dt = 0$, dus de gehele term is 0.

Hieruit volgt dus:

$$\mathrm{var}(\underline{t}) > \mathrm{var}[v(\underline{\hat{\theta}})] ,$$

tenzij $\underline{t} = v(\underline{\hat{\theta}})$, hetgeen we in het begin hebben uitgesloten.

(4.21) Stelling: Stel $\underline{\hat{\theta}} = d(\underline{x}_1,\ldots,\underline{x}_n)$ is voldoende voor θ. Stel dat de verdeling van $\underline{\hat{\theta}}$ volledig is. Als er een functie $v(\underline{\hat{\theta}})$ bestaat met $E[v(\underline{\hat{\theta}})] = u(\theta)$, dan is $v(\underline{\hat{\theta}})$ de nauwkeurigste zuivere schatter voor $u(\theta)$.

Bewijs: Omdat de verdeling van $\underline{\hat{\theta}}$ volledig is, is er slechts één functie $v(\underline{\hat{\theta}})$ met $E[v(\underline{\hat{\theta}})] = u(\theta)$. Volgens stelling (4.20) hebben alle andere zuivere schatters een grotere variantie, $v(\underline{\hat{\theta}})$ moet dus de nauwkeurigste zuivere schatter zijn.

Wij gaan nu een aantal methoden bespreken waarmee schatters van parameters kunnen worden verkregen, namelijk:

1) De methode van de meest aannemelijke schattingen (maximum likelihood method).
2) De momentenmethode.
3) De Bayes-methode.

(4.22) Definitie: De aannemelijkheidsfunctie van een steekproef $\underline{x}_1,\ldots,\underline{x}_n$ uit de verdeling $f(x;\theta)$ is de simultane kansdichtheid van $\underline{x}_1,\ldots,\underline{x}_n$, waarin de variabelen $\underline{x}_1,\ldots,\underline{x}_n$ stochastisch zijn, opgevat als functie van θ:

$$\underline{L}(\theta) = g(\underline{x}_1,\ldots,\underline{x}_n;\theta) = \prod_{i=1}^{n} f(\underline{x}_i;\theta) .$$

(4.23) Definitie: Stel $\underline{L}(\theta)$ is de aannemelijkheidsfunctie van de variabelen $\underline{x}_1,\ldots,\underline{x}_n$. Als $\underline{\hat{\theta}} = d(\underline{x}_1,\ldots,\underline{x}_n)$ de waarde van θ is waarvoor $\underline{L}(\theta)$ maximaal is, dan heet $\underline{\hat{\theta}}$ de meest aannemelijke schatter voor θ.

Vaak is het eenvoudiger om in plaats van $L(\theta)$ de functie $\log L(\theta)$ te bekijken, die in hetzelfde punt zijn maximum bereikt.

De generalisatie voor meer dan één parameter ligt voor de hand.
Onder vrij zwakke voorwaarden hebben de meest aannemelijke schatters de volgende eigenschappen:

1) Asymptotisch nauwkeurig (en dus asymptotisch raak).
2) Asymptotisch doeltreffend en BAN.
3) Ze zijn een functie van de voldoende statistische grootheid.

Dat de laatste eigenschap geldt is eenvoudig in te zien, immers volgens stelling (4.8) geldt, als $\underline{t}$ voldoende is voor θ:

$$L(\theta) = g(x_1,\dots,x_n;\ ) = h(t;\theta)k(x_1,\dots,x_n)\ .$$

Het maximum wordt bereikt als $h(t;\theta)$ maximaal is, waaruit $\hat{\theta}$ als functie van t wordt gevonden.
Verder geldt dat de meest aannemelijke schatter <u>invariant</u> is. Dat betekent dat als $\underline{\hat{\theta}}$ de meest aannemelijke schatter voor θ is en $u(\theta)$ is een éénwaardige functie, $u(\underline{\hat{\theta}})$ de meest aannemelijke schatter voor $u(\theta)$ is. Hiermee samen hangt het feit dat de meest aannemelijke schatter in het algemeen niet zuiver is, want $\underline{\hat{\theta}}$ en $u(\underline{\hat{\theta}})$ zullen meestal niet beide zuiver zijn, omdat niet geldt dat

$$E[u(\underline{\hat{\theta}})] = u[E(\underline{\hat{\theta}})]\ .$$

<u>Voorbeelden</u>:

1) Stel $\underline{x}_1,\dots,\underline{x}_n$ vormen een steekproef uit de alternatieve verdeling

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}\ , \qquad x = 0,1;\ 0 \leqslant \theta \leqslant 1\ .$$

$$L(\theta) = \theta^{\Sigma x_i}(1-\theta)^{m-\Sigma x_i}$$

$$\log L(\theta) = \Sigma\, x_i \log\theta + (n-\Sigma\, x_i)\log(1-\theta)$$

$$\frac{d \log L(\theta)}{d\theta} = \frac{\Sigma\, x_i}{\theta} - \frac{(n-\Sigma\, x_i)}{1-\theta}\ .$$

Door deze uitdrukking nul te stellen vinden we

$$\hat{\theta} = \frac{\Sigma\, x_i}{n} \quad \text{en dus} \quad \underline{\hat{\theta}} = \frac{\Sigma\, \underline{x}_i}{n}\ .$$

2) Voor de normale verdeling $f(x;\mu,\sigma^2)$ wordt gevonden:

$$\underline{\hat{\mu}} = \frac{1}{n}\Sigma\, \underline{x}_i = \underline{\bar{x}}$$

$$\underline{\hat{\sigma}}^2 = \frac{1}{n}\Sigma(\underline{x}_i - \underline{\bar{x}})^2\ .$$

Zoals bekend is de laatste schatter niet zuiver, maar dit kan worden gecorrigeerd door vermenigvuldiging met $\frac{n}{n-1}$.

3) Stel $\underline{x}$ is homogeen verdeeld tussen α en β, dus

$$f(x) = \frac{1}{\beta - \alpha} \quad , \quad \alpha \leqslant x \leqslant \beta \, .$$

De aannemelijkheidsfunctie voor een steekproef $\underline{x}_1,\dots,\underline{x}_n$ is nu

$$L(\alpha,\beta) = \frac{1}{(\beta-\alpha)^n} \qquad \text{voor } \alpha \leqslant \underline{x}_i \leqslant \beta \ (i = 1,\dots,n) \ \text{ en 0 elders.}$$

Het maximum wordt bereikt als $(\beta - \alpha)$ minimaal is, dus als

$$\alpha = \underline{x}_{(1)} = \min_i \underline{x}_i$$

en

$$\beta = \underline{x}_{(n)} = \max_i \underline{x}_i \, .$$

De <u>momentenmethode</u> voor het schatten van de parameters $\theta_1,\dots,\theta_k$ van een verdeling $f(x;\theta_1,\dots,\theta_k)$ bestaat uit het oplossen van de k vergelijkingen:

(4.24) $\quad \mu'_t = m'_t \, , \quad (t = 1,\dots,k) \, ,$

waarin

$$\mu'_t = E\underline{x}^t$$

$$m'_t = \frac{1}{n}\sum_1^n x_i^t \, .$$

De eerste k momenten worden dus aan de eerste k steekproefmomenten gelijk gesteld.

Onder vrij algemene voorwaarden zijn de schatters verkregen volgens de momentenmethode:

1) Asymptotisch nauwkeurig (en dus asymptotisch raak).
2) Asymptotisch normaal.

Ze zijn in het algemeen niet asymptotisch doeltreffend of BAN

<u>Voorbeelden</u>:

1) Voor de normale verdeling wordt hetzelfde resultaat gevonden als met de methode van de meest aannemelijke schatters.

2) Voor de homogene verdeling tussen α en β vindt men:

$$E\underline{x} = \tfrac{1}{2}(\beta + \alpha) = \frac{1}{n}\sum x_i = \bar{x}$$

$$E\underline{x}^2 = \frac{1}{3}(\beta^2 + \beta\alpha + \alpha^2) = \frac{1}{n}\Sigma\, x_i^2 \ .$$

Hieruit volgt, na enige herleiding,

$$\alpha = \bar{x} - \sqrt{\frac{3}{n}\Sigma(x_i - \bar{x})^2} \ ,$$

$$\beta = \bar{x} + \sqrt{\frac{3}{n}\Sigma(x_i - \bar{x})^2} \ .$$

De methode van de Bayes-schatters gaat er van uit dat de te schatten parameter θ zelf een stochastische variabele is met kansdichtheid $p(\theta)$. In sommige gevallen, bv. als θ de fractie defecte exemplaren is dat wordt gefabriceerd door een of ander productieproces, is deze aanname redelijk. Men kan zich dan bv. voorstellen dat θ op één dag constant is, maar dat van dag tot dag een andere waarde van θ volgens een kansverdeling $p(\theta)$ wordt gerealiseerd. Als $\underline{\theta}$ stochastisch is, wordt de risicofunctie gedefinieerd door (4.3) ook stochastisch en we kunnen de verwachting ervan berekenen. Dit verwachte risico geven we aan met $B(d)$:

$$(4.25) \qquad B(d) = E[R(d;\underline{\theta})] = \int_{-\infty}^{\infty} R(d;\theta)p(\theta)d\theta \ .$$

De Bayes-schatter voor θ is die functie van de steekproef $\underline{x}_1,\dots,\underline{x}_n$, die $B(d)$ minimaliseert.

De simultane verdeling van $\underline{x}_1,\dots,\underline{x}_n$ en $\underline{\theta}$ kan op 2 manieren in factoren worden ontbonden:

$$(4.26) \qquad q(x_1,\dots,x_n,\theta) = g(x_1,\dots,x_n \mid \theta)p(\theta) = h(\theta \mid x_1,\dots,x_n)k(x_1,\dots,x_n) \ .$$

De dichtheid $p(\theta)$ heet de a priori verdeling van θ, h is de a posteriori verdeling. Hiervan gebruik makend kan $B(d)$ als volgt worden herleid:

$$B(d) = \int_{-\infty}^{\infty} R(d;\theta)p(\theta)d\theta =$$

$$= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \ell[d(x_1,\dots,x_n);\theta]g(x_1,\dots,x_n \mid \theta)dx_1\dots dx_n\right\}p(\theta)d\theta =$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} k(x_1,\dots,x_n)\left\{\int_{-\infty}^{\infty} \ell[d(x_1,\dots,x_n);\theta]h(\theta \mid x_1,\dots,x_n)\, d\theta\right\}dx_1\dots dx_n \ .$$

Als nu $\hat{\theta} = d(x_1,\dots,x_n)$ zo wordt gekozen dat het gedeelte tussen accolades

$$(4.27) \qquad v(\hat{\theta};x_1,\dots,x_n) = \int_{-\infty}^{\infty} \ell(\hat{\theta};\theta)h(\theta \mid x_1,\dots,x_n)d\theta$$

wordt geminimaliseerd voor elke gerealiseerde steekproef $x_1,\dots,x_n$, dan is ook $B(d)$ minimaal en dan is $\underline{\hat{\theta}} = d(\underline{x}_1,\dots,\underline{x}_n)$ de Bayes schatter voor θ. v heet het a posteriori risico. We hebben dus bewezen de volgende stelling:

(4.28) Stelling: De waarde van $\hat{\theta}$ als functie van $x_1,\dots,x_n$ die de functie $v(\hat{\theta};x_1,\dots,x_n)$, gedefinieerd door (4.27) minimaliseert, is de Bayes schatter voor θ voor de verliesfunctie $\ell(\hat{\theta};\theta)$.

Onder zwakke voorwaarden hebben de Bayes-schatters voor een willekeurige a priori verdeling dezelfde eigenschappen als de meest aannemelijke schatters.

Voorbeelden:

1) Stel $f(x \mid \theta) = \theta^x(1-\theta)^{1-x}$, $\quad x = 0,1;\ 0 \leq \theta \leq 1$

$\ell(\hat{\theta};\theta) = (\hat{\theta}-\theta)^2$

$p(\theta) = 1$, $\quad 0 \leq \theta \leq 1$.

De voorwaardelijke verdeling van een steekproef $\underline{x}_1,\dots,\underline{x}_n$, gegeven $\underline{\theta} = \theta$ is:

$$g(x_1,\dots,x_n \mid \theta) = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

en omdat $p(\theta) = 1$, is dit ook de verdeling $q(x_1,\dots,x_n;\theta)$ (vgl. 4.26). Dus

$$h(\theta \mid x_1,\dots,x_n) \propto \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i} .$$

Dit is dus blijkbaar een beta verdeling en de ontbrekende factor is daarmee bekend:

$$h(\theta \mid x_1,\dots,x_n) = \frac{(n+1)!}{(\Sigma x_i)!(n-\Sigma x_i)!}\,\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i} .$$

Dus $v(\hat{\theta};x_1,\dots,x_n)$ is in dit geval:

$$v(\hat{\theta};x_1,\dots,x_n) = \frac{(n+1)!}{(\Sigma x_i)!(n-\Sigma x_i)!}\int_0^1 (\hat{\theta}-\theta)^2\,\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}\,d\theta =$$

$$= \hat{\theta}^2 - 2\hat{\theta}E(\underline{\theta} \mid x_1,\dots,x_n) + E(\underline{\theta}^2 \mid x_1,\dots,x_n) .$$

Het minimum wordt bereikt voor

$$\hat{\theta} = E(\underline{\theta} \mid x_1,\ldots,x_n) = \frac{\Sigma\, x_i + 1}{n+2} \quad .$$

Opmerking: Als $\ell(\hat{\theta};\theta) = c(\hat{\theta}-\theta)^2$ volgt uit de hierboven gegeven afleiding dat de Bayes-schatter altijd gevonden wordt uit:

$$\hat{\theta} = E(\underline{\theta} \mid x_1,\ldots,x_n) \; .$$

2)

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

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

$$\ell(\hat{\mu};\mu) = (\hat{\mu}-\mu)^2 \; .$$

$$g(x_1,\ldots,x_n \mid \mu)p(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\Sigma(x_i-\mu)^2}\, \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\mu^2} \; .$$

Het gedeelte dat van μ afhangt is:

$$e^{-\frac{1}{2}\left[(n+1)\mu^2 - 2\,\Sigma\, x_i\, \mu\right]} \; .$$

Dus is

$$h(\mu \mid x_1,\ldots,x_n) \propto e^{-\frac{(n+1)}{2}\left(\mu - \frac{\Sigma\, x_i}{n+1}\right)^2} \; .$$

Dit is dus een normale verdeling met gemiddelde $\frac{\Sigma\, x_i}{n+1}$ en variantie $\frac{1}{n+1}$.
Zoals bij het vorige vraagstuk werd opgemerkt is de Bayes-schatting de voorwaardelijke verwachting:

$$\hat{\mu} = E(\underline{\mu} \mid x_1,\ldots,x_n) = \frac{\Sigma\, x_i}{n+1} = \frac{n}{n+1}\,\bar{x} \; .$$

Hoofdstuk V, Steekproefverdelingen

Om de eigenschappen van een schatter goed te kunnen bestuderen is het nodig de verdeling van deze schatter te kennen. Bovendien zullen we dergelijke verdelingen later moeten gebruiken bij het opstellen van betrouwbaarheidsintervallen en bij het toetsen van hypothesen.

We moeten dus een methode hebben om verdelingen te bepalen van statistische grootheden, van functies dus van stochastische variabelen met bekende verdelingen. Voor functies van één variabele, met continue verdelingsdichtheid, werd dit reeds besproken in Wiskunde IVb.

Voor een monotone differentieerbare functie $\underline{y} = u(\underline{x})$ was het resultaat dat de verdelingsdichtheid van $\underline{y}$ gelijk is aan

$$(5.1) \qquad g(y) = f[v(y)] \left| \frac{d[v(y)]}{dy} \right| ,$$

als $f(x)$ de verdelingsdichtheid van $\underline{x}$ is en $x = v(y)$ de inverse functie van $y = u(x)$.

Als $u(x)$ niet monotoon is en dus de inverse $v(y)$ niet eenwaardig, dan wordt het gebied van de x-as waar $f(x) > 0$ is in intervallen ingedeeld waar $u(x)$ wel monotoon is en $g(y)$ wordt gevonden door de formule (5.1) op de afzonderlijke intervallen toe te passen en daarna te sommeren.

Voorbeeld

Stel $\underline{x}$ heeft de verdeling

$$\begin{aligned} f(x) &= \tfrac{2}{9}(x+1) \qquad && -1 < x < 2 \\ &= 0 && \text{elders.} \end{aligned}$$

Gevraagd wordt de verdeling van $\underline{u} = \underline{x}^2$.

De inverse functie is

$$\begin{aligned} x &= -\sqrt{u} \qquad && -1 < x < 0, \quad 0 < u < 1 \\ x &= +\sqrt{u} && 0 < x < 2, \quad 0 < u < 4. \end{aligned}$$

In het interval $0 < u < 1$ is de inverse tweewaardig. Daar geldt dus

$$g(u) = f(-\sqrt{u})\frac{1}{2\sqrt{u}} + f(\sqrt{u})\frac{1}{2\sqrt{u}} = \frac{2}{9\sqrt{u}} \qquad 0 < u < 1 .$$

In het interval $1 < u < 4$ is

$$g(u) = f(\sqrt{u})\frac{1}{2\sqrt{u}} = \frac{\sqrt{u}+1}{9\sqrt{u}} \qquad 1 < u < 4 .$$

Als $\underline{x}$ een discrete verdeling heeft, dan wordt de verdeling van $\underline{y} = u(\underline{x})$ direct gevonden door toepassing van de wetten van de kansrekening:

$$(5.2) \qquad P[\underline{y} = y] = \sum_i P[\underline{x} = x_i \mid u(x_i) = y] \ .$$

Dit kan direct gegeneraliseerd worden voor een functie van meer variabelen $\underline{x}_1,\dots,\underline{x}_n$.

Voorbeeld:

Stel $\underline{x}$ kan de waarden i aannemen met kansen p_i $(i = 0,1,\dots,5)$. De verdeling van $\underline{u} = (\underline{x} - 2)^2$ wordt als volgt gevonden:

x	0	1	2	3	4	5
$P[\underline{x} = x]$	p_0	p_1	p_2	p_3	p_4	p_5
u	4	1	0	1	4	9

u	0	1	4	9
$P[\underline{u} = u]$	p_2	$p_1 + p_3$	$p_0 + p_4$	p_5

Als $\underline{u}$ een functie is van meerdere variabelen $\underline{x}_1,\dots,\underline{x}_n$, dan wordt in het continue geval de verdeling van $\underline{u}$ gevonden door de vergelijking

$$u = u(x_1,\dots,x_n)$$

naar één van de x_i's op te lossen, bv.:

$$x_1 = x_1(u,x_2,\dots,x_n) \ .$$

Op dezelfde wijze als (5.1) werd gevonden kan worden afgeleid dat de simultane verdelingsdichtheid van $\underline{u}, \underline{x}_2,\dots,\underline{x}_n$ is

$$(5.2) \qquad g(u,x_1,\dots,x_n) = f\{x_1(u,x_2,\dots,x_n),x_2,\dots,x_n\} \left|\frac{\partial x_1}{\partial u}\right| \ .$$

De kansdichtheid van $\underline{u}$ wordt gevonden door over $x_2,\dots,x_n$ te integreren. Is de oplossing $x_1(u,x_2,\dots,x_n)$ niet eenduidig, dan moet (5.2) over de verschillende oplossingen worden gesommeerd.

Willen we de simultane verdeling hebben van r $(r \leqslant n)$ functies

$$\begin{cases} \underline{u}_1 = u_1(\underline{x}_1,\dots,\underline{x}_n) \\ \cdot \\ \cdot \\ \cdot \\ \underline{u}_r = u_r(\underline{x}_1,\dots,\underline{x}_n) \ , \end{cases}$$

dan worden uit deze vergelijkingen r x_i's opgelost, bv.:

$$(5.3)\quad \begin{cases} x_1 = x_1(u_1,\dots,u_r,x_{r+1},\dots,x_n) \\ \vdots \\ x_r = x_r(u_1,\dots,u_r,x_{r+1},\dots,x_n). \end{cases}$$

De determinant van Jacobi van deze transformatie wordt bepaald:

$$\left|\frac{\partial x_i}{\partial u_j}\right| = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \dots & \frac{\partial x_1}{\partial u_r} \\ \vdots & & \\ \frac{\partial x_r}{\partial u_1} & \dots & \frac{\partial x_r}{\partial u_r} \end{vmatrix}$$

De simultane verdeling van $\underline{u}_1,\dots,\underline{u}_r$, $\underline{x}_{r+1},\dots,\underline{x}_n$ is nu

$$(5.4)\quad g(u_1,\dots,u_r,x_{r+1},\dots,x_n) = f\{x_1(u_1,\dots,u_r),\dots,x_r(u_1,\dots,u_r), x_{r+1},\dots,x_n\} \left|\frac{\partial x_i}{\partial u_j}\right|_+$$

waarin $\left|\frac{\partial x_i}{\partial u_j}\right|_+$ de absolute waarde van de determinant van Jacobi voorstelt.

Is de oplossing (5.3) niet éénduidig, dan wordt het rechterlid van (5.4) weer een som. Het bewijs volgt uit de stelling over de transformatie van meervoudige integralen besproken in Wiskunde IVa. De verdeling van $\underline{u}_1,\dots,\underline{u}_r$ wordt nu weer verkregen door te integreren over $x_{r+1},\dots,x_n$. Vaak kan deze integratie worden vermeden door het gebruik van momenten voortbrengende functies, zoals in een aantal nu volgende gevallen wordt gedemonstreerd.

De verdeling van het <u>gemiddelde van een steekproef uit een normale verdeling</u> wordt als volgt gevonden. De momenten voortbrengende functie van $\underline{\bar{x}} = \frac{1}{n}\Sigma\,\underline{x}_i$ is:

$$m(t) = Ee^{t\frac{\Sigma \underline{x}_i}{n}} = e^{n\mu\frac{t}{n}+\frac{1}{2}\frac{t^2}{n^2}\sigma^2} = e^{\mu t+\frac{1}{2}t^2\frac{\sigma^2}{n}}.$$

Dit is de momenten voortbrengende functie van een normale verdeling met gemiddelde μ en variantie $\frac{\sigma^2}{n}$, waarmee bewezen is dat dit de verdeling van $\underline{\bar{x}}$ is.

De chi-kwadraat (χ^2) verdeling ontstaat op de volgende wijze. Stel $\underline{y}_1,\ldots,\underline{y}_k$ zijn onderling onafhankelijk normaal verdeeld met gemiddelde 0 en variantie 1. Gevraagd wordt nu de verdeling van $\underline{u} = \Sigma\, \underline{y}_i^2$. De momenten voortbrengende functie is:

$$Ee^{t\underline{u}} = Ee^{t\Sigma \underline{y}_i^2} = \left(\frac{1}{2\pi}\right)^{k/2} \int\cdots\int e^{t\Sigma y_i^2 - \frac{1}{2}\Sigma y_i^2}\, dy_1\ldots dy_k =$$

$$= \left(\frac{1}{2\pi}\right)^{k/2} \int\cdots\int e^{-\frac{1-2t}{2}\Sigma y_i^2}\, dy_1\ldots dy_k\ .$$

Omdat $\sqrt{\frac{1-2t}{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1-2t}{2}y_i^2}\, dy_i = 1$ (normale verdeling met gemiddelde 0 en variantie $\frac{1}{1-2t}$) geldt:

$$Ee^{t\underline{u}} = \frac{1}{(1-2t)^{k/2}}\ .$$

Dit is, volgens (2.4), de momenten voortbrengende functie van een gamma verdeling met parameters $\alpha = \frac{k}{2} - 1$ en $\beta = 2$.
Dus de kansdichtheid van $\underline{u}$ is:

(5.5) $$f(u) = \frac{1}{(\frac{k}{2}-1)!\,2^{k/2}}\, u^{k/2-1}\, e^{-\frac{1}{2}u}\ .$$

Dit wordt genoemd een χ^2 verdeling met k vrijheidsgraden.

(5.6) Definitie: De simultane verdelingsdichtheid van de variabelen $\underline{y}_1,\ldots,\underline{y}_p$ heet p-dimensionaal normaal als deze gegeven wordt door

$$f(y_1,\ldots,y_p) = \frac{|R|^{\frac{1}{2}}}{(2\pi)^{p/2}}\, e^{-\frac{1}{2}\sum_i\sum_j (y_i-\mu_i)(y_j-\mu_j)r_{ij}}\ ,$$

waarin $R = (r_{ij})$ een positief definiete $p \times p$ matrix is.

We bewijzen hier niet dat dit een zinvolle definitie is, omdat de integraal 1 oplevert. Verder kan worden bewezen, dat de inverse van de matrix R: $V = R^{-1}$ een matrix is met elementen σ_{ij}, de zogenaamde variantie-covariantie matrix, dus

$$\begin{cases} \sigma_{ii} = \sigma_i^2 = E(\underline{y}_i - \mu_i)^2 \\ \sigma_{ij} = E(\underline{y}_i - \mu_i)(\underline{y}_j - \mu_j)\ . \end{cases}$$

Wij beschouwen nu opnieuw k onderling onafhankelijke normaal verdeelde grootheden $\underline{y}_1,\ldots,\underline{y}_k$ met gemiddelde 0 en variantie 1 en zoeken de simultane verdeling van $\underline{u}$ en $\underline{v}$ gedefinieerd door

$$\begin{cases} \underline{u} = n\overline{\underline{y}}^2 = \frac{1}{n}(\Sigma\, \underline{y}_i)^2 \\ \underline{v} = \Sigma\,(\underline{y}_i - \overline{\underline{y}})^2 \,. \end{cases}$$

De simultane momenten voortbrengende functie (1.22) van $\underline{u}$ en $\underline{v}$ is

$$(5.7) \qquad \mathcal{E}e^{t_1\underline{u}+t_2\underline{v}} = \frac{1}{(2\pi)^{k/2}} \int\cdots\int e^{t_1/n\,(\Sigma y_i)^2 + t_2\,\Sigma\,(y_i - \overline{y})^2 - \frac{1}{2}\Sigma y_i^2}\, dy_1\ldots dy_n \,.$$

De integrand is $e^{-\frac{1}{2}Q}$, waarin

$$Q = -\frac{2t_1}{n}(\Sigma\, y_i)^2 - 2t_2\,\Sigma\,(y_i - \overline{y})^2 + \Sigma\, y_i^2 = \sum_i \sum_j y_i y_j \cdot r_{ij} \,,$$

met

$$r_{ii} = 1 - 2t_2 - \frac{2(t_1 - t_2)}{n} \qquad (i = 1,\ldots,k)$$

$$r_{ij} = \quad - \frac{2(t_1 - t_2)}{n} \qquad (i,j = 1,\ldots,k;\ i \neq j)\,.$$

De determinant van een k × k matrix met op de hoofddiagonaal de waarde a en daarbuiten b is: $(a-b)^{k-1}\{a + (k-1)b\}$.
Dus $|R|$ is gelijk aan

$$|R| = (1-2t_2)^{k-1}[(1-2t_2) - 2(t_1 - t_2)] = (1-2t_2)^{k-1}(1-2t_1)\,.$$

Blijkbaar is, volgens de definitie van de meerdimensionale normale verdeling (5.6) de integraal (5.7) gelijk aan $|R|^{-\frac{1}{2}}$, dus:

$$\mathcal{E}e^{t_1\underline{u}+t_2\underline{v}} = \frac{1}{(1-2t_2)^{(k-1)/2}} \cdot \frac{1}{(1-2t_1)^{\frac{1}{2}}} \,.$$

Dit is het product van de momenten voortbrengende functies van twee χ^2-verdelingen met respectievelijk (1) en (k - 1) vrijheidsgraden. Dus de simultane momenten voortbrengende functie van $\underline{u}$ en $\underline{v}$ is de m.v.f. van twee <u>onderling onafhankelijke variabelen</u>, die χ^2-verdelingen hebben met 1 en (k - 1) vrijheidsgraden. Uit het meerdimensionale analogon van stelling (1.15) volgt nu dat $\underline{u}$ een χ^2-verdeling heeft met 1 vrijheidsgraad en $\underline{v}$ een χ^2-verdeling met (k - 1) vrijheidsgraden, en dat $\underline{u}$ en $\underline{v}$ onafhankelijk verdeeld zijn. De verdeling van $\underline{u}$ was uiteraard van tevoren bekend, omdat $\underline{u}$ het kwadraat is van één normaal (0,1) verdeelde variabele.

Uit het feit dat $\underline{u}$ en $\underline{v}$ onafhankelijk zijn volgt niet zonder meer dat $\bar{\underline{y}}$ en $\underline{v}$ ook onafhankelijk zijn, want $\bar{\underline{y}}$ is geen eenwaardige functie van $\underline{u}$. Het bewijs in het boek van Mood en Graybill van stelling 10.3 op pag. 230 is daarom onvolledig. Men kan echter op soortgelijke manier als hiervoor werd gedaan laten zien dat de momenten voortbrengende functie van $\bar{\underline{y}}$ en $\underline{v}$ gelijk is aan:

$$E e^{t_1 \bar{\underline{y}} + t_2 \underline{v}} = e^{\frac{1}{2}\frac{t_1^2}{n}} \times \frac{1}{(1-2t_2)^{(n-1)/2}} \, ,$$

waaruit volgt dat $\bar{\underline{y}}$ en $\underline{v}$ eveneens onafhankelijk zijn.

De F-verdeling wordt gedefinieerd als de verdeling van

$$(5.8) \qquad \underline{F} = \frac{\underline{u}/m}{\underline{v}/n} \, ,$$

waarin $\underline{u}$ en $\underline{v}$ onafhankelijke variabelen zijn die χ^2-verdelingen bezitten met respectievelijk m en n vrijheidsgraden. Men zegt nu dat F m en n vrijheidsgraden heeft. De simultane kansdichtheid van $\underline{u}$ en $\underline{v}$ is (volgens 5.5):

$$f(u,v) = \frac{1}{(\frac{m}{2}-1)!(\frac{n}{2}-1)!2^{(m+n)/2}} u^{m/2-1} v^{n/2-1} e^{-\frac{1}{2}(u+v)} \, .$$

De transformatie

$$F = \frac{n}{m}\frac{u}{v} \, , \quad u = \frac{mFv}{n} \, , \quad \frac{\partial u}{\partial F} = \frac{mv}{n} \, ,$$

geeft

$$g(F,v) = \frac{1}{(\frac{m-2}{2})!(\frac{n-2}{2})!2^{(m+n)/2}} \left(\frac{m}{n}\right)^{m/2} F^{m/2-1} v^{(m+n)/2-1} e^{-\frac{1}{2}v(1+\frac{mF}{n})} .$$

De integraal

$$\int_0^\infty v^{(m+n)/2-1} e^{-\frac{1}{2}v(1+\frac{mF}{n})} dv$$

is bekend van de gamma-verdeling (2.3). De waarde kan hieruit worden gevonden door te substitueren

$$\alpha = \frac{m+n}{2} - 1 \quad \text{en} \quad \beta = \frac{2}{1+\frac{mF}{n}} \, .$$

Wij vinden dan

$$(5.9) \qquad h(F) = \int_0^\infty g(F,v)dv = \frac{(\frac{m+n-2}{2})!}{(\frac{m-2}{2})!(\frac{n-2}{2})!}\left(\frac{m}{n}\right)^{m/2} \frac{F^{m/2-1}}{(1+\frac{mF}{n})^{(m+n)/2}} .$$

Een verdeling die zeer belangrijk is voor practische toepassingen is de t-verdeling van Student (pseudoniem voor W.S. Gosset). Deze wordt als volgt gedefinieerd: Als $\underline{y}$ een normale verdeling heeft met gemiddelde 0 en variantie 1 en $\underline{u}$ heeft een χ^2-verdeling met k vrijheidsgraden, terwijl $\underline{y}$ en $\underline{u}$ onafhankelijk zijn, dan heeft

$$(5.10) \qquad \underline{t} = \frac{\underline{y}}{\sqrt{\underline{u}/k}}$$

een t-verdeling met k vrijheidsgraden. Het is eenvoudig af te leiden dat de verdelingsdichtheid gelijk is aan

$$(5.11) \qquad h(t) = \frac{1}{\sqrt{k\pi}} \frac{[(k-1)/2]!}{[(k-2)/2]!} \frac{1}{(1+t^2/k)^{(k+1)/2}} .$$

Omdat $\underline{y}^2$ een χ^2-verdeling heeft met 1 vrijheidsgraad, is $\underline{t}^2 = \frac{\underline{y}^2}{\underline{u}/k}$ een F-verdeling met 1 en k vrijheidsgraden.

Tenslotte bespreken we nog de verdeling van geordende waarnemingen (order statistics).

Als $\underline{x}_1,\ldots,\underline{x}_n$ een aselecte steekproef is met continue verdelingsdichtheid f(x), dan zoeken we naar de simultane verdeling van de geordende waarnemingen $\underline{y}_1,\ldots,\underline{y}_n$. Dit wil zeggen dat elke $\underline{y}_i$ gelijk is aan een van de $\underline{x}_i$'s, terwijl $\underline{y}_1 < \underline{y}_2 < \ldots < \underline{y}_n$. Wij voeren nu de volgende transformatie uit:

$$y_1 = x_{i_1} ; \; y_2 = x_{i_2} ; \; \ldots ; \; y_n = x_{i_n} ,$$

als $x_{i_1} < x_{i_2} < \ldots < x_{i_n}$.

Op deze wijze wordt de x-ruimte in n! gebieden verdeeld die worden verkregen door voor $(i_1,\ldots,i_n)$ alle n! permutaties van $(1,\ldots,n)$ te nemen. Voor elk van deze gebieden is de simultane verdeling van $\underline{y}_1,\ldots,\underline{y}_n$:

$$h(y_1,\ldots,y_n) = f(y_1)\ldots f(y_n) , \qquad (y_1 < y_2 < \ldots < y_n) .$$

Dus voor de hele ruimte wordt gevonden:

$$(5.12) \qquad h(y_1,\ldots,y_n) = n!f(y_1)\ldots f(y_n) , \qquad (y_1 < y_2 < \ldots < y_n) .$$

Hieruit kunnen de marginale verdelingen van de $\underline{y}_i$ en ook verdelingen van functies van de $\underline{y}_i$ worden afgeleid.
De verdeling van $\underline{y}_i$ wordt als volgt gevonden:

$$(5.13) \qquad g(y_i) = n!f(y_i) \int_{-\infty}^{y_i} \cdots \int_{-\infty}^{y_2} f(y_1)\ldots f(y_{i-1})dy_1\ldots dy_{i-1} \times$$

$$\times \int_{y_i}^{\infty} \cdots \int_{y_i}^{y_{i+3}} \int_{y_i}^{y_{i+2}} f(y_{i+1})\ldots f(y_n)dy_{i+1}\ldots dy_n =$$

$$= n!f(y_i) \frac{1}{(i-1)!} \{F(y_i)\}^{i-1} \frac{1}{(n-i)!} \{1-F(y_i)\}^{n-i} =$$

$$= \frac{n!}{(i-1)!(n-i)!} \{F(y_i)\}^{i-1}\{1-F(y_i)\}^{n-i} f(y_i) \ .$$

Dit kan als volgt worden geïnterpreteerd: $\frac{n!}{(i-1)!(n-i)!}$ is het aantal manieren waarop uit n variabelen groepen van $(i-1)$, 1 en $(n-i)$ stuks kunnen worden gevormd.
$\{F(y_i)\}^{i-1}$ is de kans dat alle $(i-1)$ variabelen uit de eerste groep kleiner zijn dan y_i.
$\{1-F(y_i)\}^{n-i}$ is de kans dat alle $(n-i)$ variabelen van de laatste groep groter zijn dan y_i.
$f(y_i)$ is de kansdichtheid in het punt y_i.
Het product van deze factoren is de kansdichtheid $g(y_i)$ van de i^e variabele na rangschikking in opklimmende volgorde in y_i.

Opmerking: Door de transformatie $F(y_i) = t_i$ uit te voeren, vinden we de verdeling van de i^e order statistic $\underline{t}_i$ uit een homogene (0,1) verdeling:

$$(5.14) \qquad h(t_i) = \frac{n!}{(i-1)!(n-i)!} t_i^{i-1}(1-t_i)^{n-i} ,$$

een beta-verdeling met parameters $(i-1)$ en $(n-i)$.

De simultane verdeling van $\underline{y}_1$ en $\underline{y}_n$ vindt men door in (5.12) $y_2,\ldots,y_{n-1}$ uit te integreren:

$$p(y_1,y_n) = n!f(y_1)f(y_n) \int_{y_1}^{y_n} \cdots \int_{y_1}^{y_3} f(y_2)\ldots f(y_{n-1})dy_2\ldots dy_{n-1} =$$

$$= n(n-1)\{F(y_n) - F(y_1)\}^{n-2} f(y_1) f(y_n) \ .$$

De verdeling van de range: $\underline{r} = \underline{y}_n - \underline{y}_1$ krijgt men nu door de transformatie:

$$y_n = r + y_1 \ .$$

Dit levert op

$$k(r,y_1) = n(n-1)\{F(r+y_1) - F(y_1)\}^{n-2} f(y_1) f(r+y_1)$$

en vervolgens:

$$(5.15) \qquad k(r) = n(n-1) \int_{-\infty}^{\infty} \{F(r+y_1) - F(y_1)\}^{n-2} f(y_1) f(r+y_1) dy_1 \ .$$

Voorbeelden:

1) Voor de homogene verdeling tussen 0 en 1 wordt (5.15):

$$h(r) = n(n-1) \int_0^{1-r} r^{n-2} dy_1 = n(n-1) r^{n-2} (1-r) \ ,$$

een beta-verdeling met parameters $(n-2)$ en 1.

2) Voor de negatief exponentiële verdeling $f(x) = \frac{1}{\beta} e^{-x/\beta}$ is $F(x) = 1 - e^{-x/\beta}$.

In dat geval wordt

$$k(r,y_1) = \frac{n(n-1)}{\beta^2} e^{-r/\beta} \{1 - e^{-r/\beta}\}^{n-2} e^{-ny/\beta}$$

en

$$h(r) = \int_0^{\infty} h(r,y_1) dy_1 = \frac{n-1}{\beta} e^{-r/\beta} \{1 - e^{-r/\beta}\}^{n-2} \ .$$

Hoofdstuk VI, Betrouwbaarheidsintervallen

In hoofdstuk IV hebben wij methoden besproken die ons een schatting geven van de waarde van één of meer onbekende parameters. Daarbij was duidelijk dat deze schatting van de onbekende parameterwaarde kon afwijken en dus een zekere onnauwkeurigheid bezat.
In dit hoofdstuk bespreken wij een methode waarbij als schatting voor een parameter niet één waarde wordt gegeven maar een gebied van mogelijke waarden. Het eenvoudigste voorbeeld van zo'n methode is reeds in Wiskunde IVb besproken, namelijk dat van het gemiddelde van een normale verdeling met bekende variantie. Wij weten bv. dat voor het steekproefgemiddelde $\underline{\bar{x}}$ van een steekproef uit een normale verdeling met gemiddelde μ en variantie σ^2 geldt:

$$P[\mu - 1{,}96\sigma/\sqrt{n} < \underline{\bar{x}} < \mu + 1{,}96\sigma/\sqrt{n}] = 0{,}95 .$$

Deze vergelijking is gelijkwaardig met de volgende

$$P[\underline{\bar{x}} - 1{,}96\sigma/\sqrt{n} < \mu < \underline{\bar{x}} + 1{,}96\sigma/\sqrt{n}] = 0{,}95 .$$

Deze uitspraak heeft betrekking op een stochastisch interval dat met een bepaalde kans een vaste, onbekende parameterwaarde bevat. We zeggen nu dat de uitspraak

$$(6.1) \qquad \bar{x} - 1{,}96\sigma/\sqrt{n} < \mu < \bar{x} + 1{,}96\sigma/\sqrt{n} ,$$

waarin $\bar{x}$ een realisatie is van $\underline{\bar{x}}$, een betrouwbaarheid heeft van 0,95 of 95%. Dat wil dus zeggen dat de uitspraak gemiddeld eens op de 20 keer onjuist zal zijn bij het doen van telkens nieuwe waarnemingen.

Bij practische problemen is meestal niet alleen μ maar ook σ onbekend. In plaats van σ moet nu de schatter $\underline{s}$ worden gebruikt, waarbij

$$\underline{s} = \sqrt{\frac{\Sigma(\underline{x}_i - \underline{\bar{x}})^2}{n-1}} .$$

In plaats van op de variabele

$$\underline{u} = \frac{\underline{\bar{x}} - \mu}{\sigma/\sqrt{n}} ,$$

die normaal verdeeld is met gemiddelde 0 en variantie 1, baseren we ons nu op

$$\underline{t} = \frac{\underline{\bar{x}} - \mu}{\underline{s}/\sqrt{n}} = \frac{\underline{\bar{x}} - \mu}{\sigma/\sqrt{n}} \times \frac{1}{\sqrt{\dfrac{\Sigma(\underline{x}_i - \underline{\bar{x}})^2}{\sigma^2(n-1)}}} .$$

Dit is het quotiënt van een normaal (0,1) verdeelde variabele en de wortel uit een χ^2-verdeelde grootheid gedeeld door het aantal vrijheidsgraden. Het laatste is als volgt in te zien:

$$\underline{u} = \frac{\Sigma(\underline{x}_i - \overline{\underline{x}})^2}{\sigma^2} = \Sigma\left\{\frac{(\underline{x}_i - \mu)}{\sigma} - \frac{(\overline{\underline{x}} - \mu)}{\sigma}\right\}^2 = \Sigma(\underline{y}_i - \overline{\underline{y}})^2 ,$$

waarin $\underline{y}_i$ normaal (0,1) verdeeld is, heeft volgens hoofdstuk V een χ^2-verdeling. Bovendien zijn $\dfrac{\overline{\underline{x}} - \mu}{\sigma/\sqrt{n}}$ en $\dfrac{\Sigma(\underline{x}_i - \overline{\underline{x}})^2}{\sigma^2}$ onafhankelijk van elkaar. Dus heeft $\underline{t}$ een t-verdeling met $(n-1)$ vrijheidsgraden. Als nu $t_{0.025}$ zo wordt gekozen dat

$$P[-t_{0.025} < \underline{t} < t_{0.025}] = 0,95 ,$$

dan geldt dus

$$P[-t_{0.025} < \frac{\overline{\underline{x}} - \mu}{\underline{s}/\sqrt{n}} < t_{0.025}] = 0,95 ,$$

of:

$$P[\overline{\underline{x}} - t_{0.025}\,\underline{s}/\sqrt{n} < \mu < \overline{\underline{x}} + t_{0.025}\,\underline{s}/\sqrt{n}] = 0,95 ,$$

waaruit dus volgt dat de uitspraak

$$(6.2) \qquad \overline{x} - t_{0.025}\, s/\sqrt{n} < \mu < \overline{x} + t_{0.025}\, s/\sqrt{n}$$

een betrouwbaarheid heeft van 95%.

<u>Opmerking</u>: De betrouwbaarheidsintervallen (6.1) en (6.2) zijn beide symmetrisch t.o.v. $\overline{x}$. Bij een symmetrische verdeling, zoals de normale en de t-verdeling geeft dit, bij een vaste onbetrouwbaarheid het kortste interval. Bij een vaste intervallengte wordt namelijk op deze manier het grootste oppervlak ingesloten. Aan weerszijden blijft een even groot oppervlak van de verdeling over. Intervallen die deze laatste eigenschap hebben heten <u>centraal</u>. Bij een niet symmetrische verdeling zijn ze in het algemeen niet het kortst en nooit symmetrisch. Toch worden meestal centrale intervallen toegepast omdat ze het eenvoudigst zijn te bepalen.

Zoals hiervoor reeds werd opgemerkt heeft de stochastische variabele

$$\underline{u} = \frac{\Sigma(\underline{x}_i - \overline{\underline{x}})^2}{\sigma^2}$$

een χ^2-verdeling met $(n-1)$ vrijheidsgraden, als $\underline{x}_1,\ldots,\underline{x}_n$ een steekproef uit een normale verdeling met variantie σ^2 is. Als de waarden a en b nu zo worden gekozen dat

$$P[a \leqslant \underline{u} \leqslant b] = \gamma ,$$

of

$$P\left[\frac{\Sigma(\underline{x}_i - \underline{\bar{x}})^2}{b} \leqslant \sigma^2 \leqslant \frac{\Sigma(\underline{x}_i - \underline{\bar{x}})^2}{a}\right] = \gamma ,$$

dan is het interval

$$(6.3) \qquad \frac{\Sigma(x_i - \bar{x})^2}{b} \leqslant \sigma^2 \leqslant \frac{\Sigma(x_i - \bar{x})^2}{a}$$

een betrouwbaarheidsinterval voor σ^2 met betrouwbaarheid γ. Gewoonlijk kiest men a en b zodanig dat $P[\underline{u} \leqslant a] = P[\underline{u} \geqslant b] = \frac{1}{2}(1-\gamma)$. Als het aantal vrijheidsgraden niet al te klein is, is het interval dat op deze manier wordt verkregen niet veel langer dan het kortst mogelijke.

Een betrouwbaarheidsgebied dat simultaan voor μ en σ^2 geldt kan niet worden geconstrueerd door de intervallen (6.2) en (6.3) te combineren en de betrouwbaarheden te vermenigvuldigen. De stochastische variabelen $\underline{t}$ en $\underline{u}$ waarop de intervallen voor μ en σ^2 zijn gebaseerd zijn namelijk niet onafhankelijk.
Wel onafhankelijk zijn, zoals reeds opgemerkt,

$$\frac{\underline{\bar{x}} - \mu}{\sigma/\sqrt{n}} \quad \text{en} \quad \frac{\Sigma(\underline{x}_i - \underline{\bar{x}})^2}{\sigma^2} .$$

Dus als

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

en

$$P\left[a' < \frac{\Sigma(\underline{x}_i - \underline{\bar{x}})^2}{\sigma^2} < b'\right] = \sqrt{\gamma} ,$$

dan geldt:

$$P\left[-a \quad \frac{\underline{\bar{x}} - \mu}{\sigma/\sqrt{n}} < a \quad \text{en} \quad a' < \frac{\Sigma(\underline{x}_i - \underline{\bar{x}})^2}{\sigma^2} < b'\right] = \gamma .$$

Dus het gebied begrensd door de parabool

$$\sigma^2 = \frac{n}{a^2}(\mu - \overline{x})^2$$

en de beide rechten

$$\sigma^2 = \frac{\Sigma(x_i - \overline{x})^2}{b'} \quad \text{en} \quad \sigma^2 = \frac{\Sigma(x_i - \overline{x})^2}{a'}$$

is een betrouwbaarheidsgebied met betrouwbaarheid γ voor μ en σ^2.

De betrouwbaarheidsintervallen (6.1), (6.2) en (6.3) konden worden bepaald omdat de stochastische variabelen

$$\frac{\underline{\overline{x}} - \mu}{\sigma/\sqrt{n}}, \quad \frac{\underline{\overline{x}} - \mu}{\underline{s}/\sqrt{n}} \quad \text{en} \quad \frac{\Sigma(\underline{x}_i - \underline{\overline{x}})^2}{\sigma^2}$$

bekende verdelingen hebben, waarin de te schatten parameter, te weten respectievelijk μ, μ en σ^2, niet meer voorkomt.
Als we een dergelijke functie niet hebben kan een andere methode worden gevolgd om een betrouwbaarheidsinterval te construeren.

Stel we willen voor de parameter θ, die voorkomt in een kansdichtheid $f(x;\theta)$, een betrouwbaarheidsinterval construeren met betrouwbaarheid 0,95. Een schatter $\underline{\hat{\theta}}$, bv. de maximum-likelihood schatter, die een functie is van $\underline{x}_1,\ldots,\underline{x}_n$, heeft als verdeling $g(\hat{\theta};\theta)$.
We bepalen nu, bij elke θ, twee grenzen $h_1(\theta)$ en $h_2(\theta)$, zodanig dat

$$P[\underline{\hat{\theta}} < h_1(\theta)] = 0,025$$

en

$$P[\underline{\hat{\theta}} > h_2(\theta)] = 0,025 \; .$$

De betrouwbaarheidsgrenzen voor θ worden nu, bij een waargenomen $\hat{\theta}$, bepaald door oplossing van de vergelijkingen:

$$(6.4) \qquad \begin{cases} h_1(\theta_1) = \hat{\theta} \\ h_2(\theta_2) = \hat{\theta} \; . \end{cases}$$

Het interval is dan:

$$(6.5) \qquad \theta_2 < \theta < \theta_1 \; .$$

In vrijwel alle practisch voorkomende gevallen zijn de functies $h_1(\theta)$ en $h_2(\theta)$ monotoon toenemend, zodat de vergelijkingen (6.4) elk maar één oplossing hebben.

Dat de betrouwbaarheid van het interval (6.5) gelijk is aan 0,95 is als volgt in te zien:

$$P[\theta < \underline{\theta}_1] = P[h_1(\theta) < h_1(\underline{\theta}_1)] = P[h_1(\theta) < \underline{\hat{\theta}}] = 0{,}975$$

$$P[\theta > \underline{\theta}_2] = P[h_2(\theta) > h_2(\underline{\theta}_2)] = P[h_2(\theta) > \underline{\hat{\theta}}] = 0{,}975\ .$$

Voorbeeld:

Stel $\quad f(x;\alpha) = \frac{2}{\alpha^2}(\alpha - x)\ , \qquad 0 < x < \alpha\ .$

De meest aannemelijke schatter, op grond van één waarneming $\underline{x}$ is $\underline{\hat{\alpha}} = 2\underline{x}$.
De kansdichtheid van $\underline{\hat{\alpha}}$ is

$$g(\hat{\alpha}) = \frac{1}{2\alpha^2}(2\alpha - \hat{\alpha})\ .$$

$$\int_0^{h_1(\alpha)} \frac{1}{2\alpha^2}(2\alpha - \hat{\alpha})d\alpha = 1 - \frac{1}{4\alpha^2}(2\alpha - h_1)^2 = 0{,}025\ .$$

Hieruit volgt

$$h_1(\alpha) = 2\alpha(1 - \sqrt{0{,}975})\ .$$

Evenzo kan worden berekend dat

$$h_2(\alpha) = 2\alpha(1 - \sqrt{0{,}025})\ .$$

De grenzen voor α worden dus berekend uit:

$$\begin{cases} 2\alpha(1 - \sqrt{0{,}975}) = 2x \\ 2\alpha(1 - \sqrt{0{,}025}) = 2x\ . \end{cases}$$

Dus het interval is:

$$\frac{x}{1 - \sqrt{0{,}025}} < \alpha < \frac{x}{1 - \sqrt{0{,}975}}\ .$$

Nu gaan we de hiervoor besproken algemene methode toepassen voor het geval van de binomiale verdeling:

$$P[\underline{x} = x \mid p] = \binom{n}{k} p^x (1-p)^{n-x} .$$

Een schatter van p is $\frac{\underline{x}}{n}$. We zouden dus nu bij iedere p twee waarden $h_1(p)$ en $h_2(p)$ moeten zoeken, zodanig dat

$$(6.6) \qquad \begin{aligned} &P[\underline{x}/n \leq h_1(p) \mid p] = 0{,}025 \\ &P[\underline{x}/n \geq h_2(p) \mid p] = 0{,}025 \end{aligned}$$

als de gewenste betrouwbaarheid 0,95 is.
Omdat $\underline{x}$ echter alleen de waarden 0,1,...,n kan aannemen kunnen zulke waarden in het algemeen niet worden gevonden.
Wat alleen nodig is zijn echter de oplossingen van de vergelijkingen (6.4).
Dat komt dus in ons geval neer op het bepalen van p_1 en p_2 zodanig dat

$$\begin{cases} P[\underline{x}/n \leq x/n \mid p_1] = 0{,}025 \\ P[\underline{x}/n \geq x/n \mid p_2] = 0{,}025 , \end{cases}$$

of

$$(6.7) \qquad \begin{cases} P[\underline{x} \leq x \mid p_1] = 0{,}025 \\ P[\underline{x} \geq x \mid p_2] = 0{,}025 , \end{cases}$$

waarin x het waargenomen steekproefresultaat is. De waarden p_1 en p_2 kunnen door interpolatie worden benaderd met een tabel van de binomiale verdeling, of er kan gebruik worden gemaakt van een relatie tussen de beta-verdeling en de binomiale verdeling, nl.

$$\sum_{y=0}^{k} \binom{n}{y} p^y (1-p)^{n-y} = 1 - F(p;k,n-k-1) ,$$

als $F(x;\alpha_1,\alpha_2)$ de cumulatieve beta-verdeling (2.6) met parameters α_1 en α_2 voorstelt. Voor n = 10 en een betrouwbaarheid van 0,95 zijn de betrouwbaarheidsintervallen bij verschillende x-waarden de volgende:

x	p_2	p_1
0	0	0,3085
1	0,0025	0,4450
2	0,0252	0,5561
3	0,0667	0,6525
4	0,1216	0,7376
5	0,1871	0,8129
6	0,2624	0,8784
7	0,3475	0,9333
8	0,4439	0,9748
9	0,5550	0,9975
10	0,6915	1

Algemeen geldt dat als $x = 0$, $p_2 = 0$ wordt gesteld en als $x = n$, $p_1 = 1$.

Opmerking: In het algemeen zal de onbetrouwbaarheid van de intervallen op deze wijze berekend, kleiner zijn dan 0,05. Als p gelijk is aan één van de p_2-waarden, dan is de éénzijdige onbetrouwbaarheid aan de onderkant precies 0,025, maar daartussen lager. Aan de bovenkant wordt de waarde 0,025 precies in de p_1-waarden aangenomen. De som van deze twee onbetrouwbaarheden is dus vrijwel steeds $< 0,05$. De waarde 0,05 heet daarom hier onbetrouwbaarheidsdrempel.

Hoofdstuk VII, Het toetsen van hypothesen

Zoals reeds op bladzijde 22 werd opgemerkt kan het toetsen van een hypothese worden geformuleerd als een beslissingsprobleem. De verzameling A van mogelijke beslissingen bestaat nu in de meeste gevallen uit slechts twee elementen a_1 en a_2. De verzameling van mogelijke parameterwaarden θ geven we weer aan met Ω. De verzameling Ω kan worden verdeeld in twee disjuncte deelverzamelingen ω_1 en ω_2, zodanig dat beslissing a_1 de beste is als $\theta \in \omega_1$ en dat a_2 de voorkeur verdient als $\theta \in \omega_2$. De verliesfunctie (verg. def. (4.2)) $\ell(a;\theta)$ heeft nu de volgende eigenschappen:

$$(7.1) \qquad \begin{cases} \ell(a;\theta) \geqslant 0 \\ \ell(a_1;\theta) = 0 \quad \text{als } \theta \in \omega_1 \\ \ell(a_2;\theta) = 0 \quad \text{als } \theta \in \omega_2 \, . \end{cases}$$

Als $\underline{s} = (\underline{x}_1,\ldots,\underline{x}_n)$ een aselecte steekproef is uit de verdeling $f(x;\theta)$, dan moet een beslissingsfunctie $d(\underline{x}_1,\ldots,\underline{x}_n)$ aan elke steekproef één van de beslissingen a_1 of a_2 toekennen. De genomen beslissing is

$$(7.2) \qquad \underline{a} = d(\underline{x}_1,\ldots,\underline{x}_n) \, .$$

Omdat er slechts twee mogelijke beslissingen zijn kan d worden gerepresenteerd door een verdeling van de n-dimensionale steekproefruimte S in twee disjuncte verzamelingen S_1 en S_2, zodanig dat a_1 wordt gekozen als $\underline{s} \in S_1$ en a_2 als $\underline{s} \in S_2$.
De risicofunctie $R(d;\theta)$ (def. (4.3)) wordt in dit geval

$$(7.3) \qquad R(d;\theta) = \ell(a_1;\theta)P(\underline{s} \in S_1 \mid \theta) + \ell(a_2;\theta)P(\underline{s} \in S_2 \mid \theta) \, .$$

Omdat volgens (7.1) altijd slechts één van de beide functies $\ell(a_1;\theta)$ of $\ell(a_2;\theta) > 0$ is, kan (7.3) worden gesplitst in

$$(7.4) \qquad R(d;\theta \in \omega_1) = \ell(a_2;\theta \in \omega_1)P(\underline{s} \in S_2 \mid \theta \in \omega_1)$$

en

$$(7.5) \qquad R(d;\theta \in \omega_2) = \ell(a_1;\theta \in \omega_2)P(\underline{s} \in S_1 \mid \theta \in \omega_2) \, .$$

De kansen die voorkomen in (7.4) en (7.5) zijn de kansen op het nemen van een verkeerde beslissing. Ze worden ook aangeduid als

$$(7.6) \qquad P(\underline{s} \in S_2 \mid \theta \in \omega_1) = P(I)$$

en

(7.7) $\quad P(\underline{s} \in S_1 \mid \theta \in \omega_2) = P(II)$.

Het <u>toetsen van een hypothese</u> is de volgende procedure:

(7.8) <u>Definitie</u>: Het toetsen van de hypothese $H_1 : \theta \in \omega_1$ tegen de alternatieve hypothese $H_2 : \theta \in \omega_2$ bestaat uit het kiezen van een beslissingsfunctie $\underline{a} = d(\underline{x}_1,\ldots,\underline{x}_n)$, die op grond van een steekproef de beslissing a_1 : accepteren (of: niet verwerpen) van H_1 dan wel de beslissing a_2 : verwerpen van H_1 ten gunste van H_2 aanwijst.

Ideaal zou een toets zijn die, voor elke $\theta \in \Omega$, $R(d;\theta)$ minimaliseert. Dit is in het algemeen niet mogelijk, bovendien is in de praktijk de verliesfunctie $\ell(a;\theta)$ niet precies bekend.
De traditionele methode is dan om een kleine kans α te kiezen, bv. $\alpha = 0{,}01$ of 0,05 en dan een klasse van toetsen te vinden waarvoor

(7.9) $\quad P(I) \leq \alpha$.

Daarna wordt een toets gezocht die, onder de voorwaarde (7.9) de kans (7.7):

(7.10) $\quad P(II) = \beta(\theta)$

minimaliseert. De kans P(I) heet de kans op een <u>fout van de 1^e soort</u>: het ten onrechte verwerpen van H_1. De kans P(II) heet de kans op een <u>fout van de 2^e soort</u>: het ten onrechte accepteren van H_1.
Het gebied S_2, dat tot verwerping van H_1 leidt heet het <u>kritieke gebied</u>.
Het minimaliseren van P(II) is hetzelfde als het maximaliseren van het <u>onderscheidingsvermogen</u> (<u>power</u>): $P(\underline{s} \in S_2)$ voor $\theta \in \omega_2$.

(7.11) <u>Definitie</u>: Het <u>onderscheidingsvermogen</u> van een toets is de kans H_1 te verwerpen. Dus:

$$\begin{aligned} P(\underline{s} \in S_2) &= P(I) \quad \text{als } \theta \in \omega_1 \,, \\ &= 1 - P(II) \quad \text{als } \theta \in \omega_2 \,. \end{aligned}$$

<u>Opmerking</u>: Deze begrippen kwamen ook al in Wiskunde IVb ter sprake.

<u>Voorbeeld</u>: $\underline{x}$ is normaal verdeeld met gemiddelde μ en variantie 1. De hypothesen zijn:

$H_1 : \mu = -1$

$H_2 : \mu = 0$.

Dus: $\Omega = \{\mu : \mu = -1,0\}$.

De verliesfunctie is:

$\ell(a_1; \mu = 0) = 1$

$\ell(a_2; \mu = -1) = 4$.

We nemen een steekproef bestaande uit één waarneming $\underline{x}$. Als nu S_1 en S_2 als volgt worden gekozen:

$S_1 = \{x : x < -1\}$,

$S_2 = \{x : x \geq -1\}$,

dan zijn de kansen op de acties a_1 en a_2 als volgt

	$\mu = -1$	$\mu = 0$
a_1	0,50	0,16
a_2	0,50	0,84

De risicofunctie heeft de waarden

$R(d;-1) = 4 \times 0,50 = 2$

$R(d;0) = 1 \times 0,16 = 0,16$.

(7.12) <u>Definitie</u>: Een enkelvoudige hypothese is een hypothese $H : \theta \in \omega$, waarbij ω uit één punt bestaat.

Als een enkelvoudige hypothese $\theta = \theta_1$ wordt getoetst tegen een enkelvoudige alternatieve hypothese $\theta = \theta_2$ dan neemt de risicofunctie maar twee waarden aan, nl.

$$R(d;\theta_1) = \ell(\theta_1)P(I) \quad \text{en} \quad R(d;\theta_2) = \ell(\theta_2)P(II) .$$

Iedere beslissingsfunctie d komt overeen met een punt in het $P(I)$, $P(II)$ vlak en met een punt in het $R(d;\theta_1)$, $R(d;\theta_2)$ vlak

<u>Voorbeeld</u>: De kans op het gooien van munt met een geldstuk is p. We toetsen de hypothese $H_1 : p = p_1 = 0,6$ tegen de alternatieve hypothese $H_2 : p = p_2 = 0,3$. De mogelijke acties zijn a_1: accepteer H_1 en a_2: accepteer H_2. Stel dat de verliesfunctie is:

$$\ell(a_1;p_1) = \ell(a_2;p_2) = 0 , \quad \ell(a_2;p_1) = \ell(p_1) = 1 ,$$

$$\ell(a_1;p_2) = \ell(p_2) = 2 .$$

Er wordt één waarneming gedaan. De bijbehorende stochastische variabele $\underline{x}$ wordt gedefinieerd door: $\underline{x} = 1$ als munt en $\underline{x} = 0$ als kruis wordt geworpen. De mogelijke beslissingsfuncties zijn:

$$d_1 : d_1(1) = a_1 \quad d_1(0) = a_1$$

$$d_2 : d_2(1) = a_1 \quad d_2(0) = a_2$$

$$d_3 : d_3(1) = a_2 \quad d_3(0) = a_1$$

$$d_4 : d_4(1) = a_2 \quad d_4(0) = a_2 .$$

De foutenkansen en de risico's zijn als volgt:

	P(I)	P(II)	$R(d;p_1)$	$R(d;p_2)$
d_1	0	1	0	2
d_2	0,4	0,3	0,4	0,6
d_3	0,6	0,7	0,6	1,4
d_4	1	0	1	0

Het is duidelijk dat, afhankelijk van de waarde van p één van de beslissingsfuncties (of <u>strategieën</u>) d_1 of d_4 de voorkeur heeft, terwijl d_2 een redelijke tussenweg is. Strategie d_3 daarentegen is voor beide mogelijke waarden van p slechter dan d_2. Men zegt dan dat d_2 d_3 <u>domineert</u> (beter is dan d_3). Verder laten wij nu nog <u>gemengde</u> (mixed of randomized) strategieën toe door te loten met kansen λ_1 en λ_2 ($\lambda_1 + \lambda_2 = 1$) tussen twee strategieën d_i en d_j. Alle mogelijke strategieën, de niet gemengde of <u>zuivere</u> en de gemengde vormen nu zowel in het P(I), P(II) vlak als in het $R(d;\theta_1)$, $R(d;\theta_2)$ vlak een convexe verzameling.

(7.13) <u>Definitie</u>: Een strategie (of beslissingsfunctie of toets) d is <u>toelaatbaar</u> (admissible) als er geen andere (zuivere of gemengde) strategie d* bestaat zodat

$$\forall_{\theta \in \Omega} R(d^*;\theta) \leq R(d;\theta)$$

$$\exists_{\theta \in \Omega} R(d^*;\theta) < R(d;\theta) .$$

De toelaatbare strategieën vormen in het $R(d;\theta_1)$, $R(d;\theta_2)$ vlak de "linkeronderrand" van het eerder genoemde convexe gebied.

(7.14) Definitie: Een Bayes strategie d met a priori kansen h_1 en $h_2 = 1 - h_1$ is een strategie die

$$B(d) = E[R(d;\underline{\theta})] = h_1 R(d;\theta_1) + h_2 R(d;\theta_2)$$

minimaal maakt.

(7.15) Stelling: Bij het toetsen van een enkelvoudige hypothese tegen een enkelvoudige alternatieve hypothese is elke toelaatbare strategie een Bayes strategie.

Bewijs: Door elk toelaatbaar punt van de rand van het convexe gebied van strategieën in het $R(d;\theta_1)$, $R(d;\theta_2)$ vlak kan een lijn $h_1 R(d;\theta_1) + h_2 R(d;\theta_2) = c$ worden getrokken, waarbij $h_1 + h_2 = 1$ en $h_i \geq 0$, die geen andere dan toelaatbare punten met het gebied gemeen heeft. Dit is de lijn die door parallelverschuiving wordt verkregen tot het gebied voor het eerst wordt geraakt. Dus correspondeert het eerstgenoemde punt met een Bayes strategie.

Opmerking: Het is eenvoudig in te zien dat omgekeerd ook elke Bayes strategie toelaatbaar is. Ga dit zelf na.

(7.16) Definitie: Een toets gebaseerd op een steekproef $\underline{x}_1,\ldots,\underline{x}_n$ uit de verdeling $f(x;\theta)$ voor de hypothese $H_1 : \theta = \theta_1$ tegen $H_2 : \theta = \theta_2$ is een likelihood-ratio toets als er een getal k bestaat zodat als volgt wordt beslist:

a_1 (accepteer H_1) als $\underline{\lambda} > k$

a_2 (accepteer H_2) als $\underline{\lambda} < k$

a_1 of a_2 als $\underline{\lambda} = k$,

waarin

$$\underline{\lambda} = \frac{f(\underline{x}_1;\theta_1)\ldots f(\underline{x}_n;\theta_1)}{f(\underline{x}_1;\theta_2)\ldots f(\underline{x}_n;\theta_2)} .$$

Voorbeeld: $\underline{x}$ is normaal verdeeld met gemiddelde μ en variantie 1. De hypothesen zijn:

$$H_1 : \mu = -1 \quad \text{en} \quad H_2 : \mu = 0 .$$

Er is één waarneming $\underline{x}$.

$$\underline{\lambda} = \frac{e^{-\frac{1}{2}(\underline{x}+1)^2}}{e^{-\frac{1}{2}\underline{x}^2}} = e^{-\frac{1}{2}(2\underline{x}+1)} .$$

Een likelihood-ratio toets is dus:
Accepteer H_1 als $\underline{x} < k$ en H_2 als $\underline{x} > k$.

(7.17) Stelling: Voor het toetsen van de hypothese $H_1 : \theta = \theta_1$ tegen $H_2 : \theta = \theta_2$ is elke Bayes strategie een likelihood-ratio toets.

Bewijs:

$$P(I) = 1 - \int\underset{S_1}{\cdots}\int f(x_1;\theta_1)\dots f(x_n;\theta_1)dx_1\dots dx_n$$

$$P(II) = \int\underset{S_1}{\cdots}\int f(x_1;\theta_2)\dots f(x_n;\theta_2)dx_1\dots dx_n$$

$$B(d) = h_1\ell(\theta_1)P(I) + h_2\ell(\theta_2)P(II) =$$

$$= h_1\ell(\theta_1) + \int\underset{S_1}{\cdots}\int[-h_1\ell(\theta_1)f(x_1;\theta_1)\dots f(x_n;\theta_1) +$$

$$+ h_2\ell(\theta_2)f(x_1;\theta_2)\dots f(x_n;\theta_2)]dx_1\dots dx_n .$$

Dit wordt geminimaliseerd door voor S_1 het gebied te kiezen waar de integrand negatief is, dus door H te accepteren als

$$\lambda = \frac{f(x_1;\theta_1)\dots f(x_1;\theta_1)}{f(x_1;\theta_2)\dots f(x_1;\theta_2)} > \frac{h_2\ell(\theta_2)}{h_1\ell(\theta_1)} .$$

Hiermee is het gestelde bewezen.

Als de verliesfunctie moeilijk kan worden bepaald, wordt meestal de volgende, meer traditionele weg gevolgd.
Om de hypothese $H_1 : \theta \in \omega_1$ te toetsen tegen $H_2 : \theta \in \omega_2$ wordt een toets gezocht, die bij vooraf gegeven α $(0 < \alpha < 1)$ bereikt dat

$$\forall_{\theta \in \omega_1} \quad P(I) \leq \alpha$$

$$\forall_{\theta \in \omega_2} \quad 1 - P(II) = \varphi(\theta) \text{ maximaal.}$$

Als ω_1 en ω_2 weer beide uit één punt bestaan, wordt deze toets weer met de likelihood-ratio methode gevonden, zoals volgt uit de volgende stelling. Dit staat bekend als het fundamentele lemma van Neyman en Pearson.

(7.18) <u>Stelling</u>: Het kritieke gebied R_α van de grootte α dat het onderscheidingsvermogen van de toets $H_1 : \theta = \theta_1$ tegen $H_2 : \theta = \theta_2$ op grond van een steekproef $\underline{x}_1,\dots,\underline{x}_n$ maximaliseert, wordt verkregen door het gebied te nemen waarin

$$\lambda = \frac{f(x_1;\theta_1)\dots f(x_n;\theta_1)}{f(x_1;\theta_2)\dots f(x_n;\theta_2)} < k_\alpha \, , \tag{7.19}$$

waarin k_α een constante is, zodanig dat

$$\int\dots\int_{R_\alpha} f(x_1;\theta_1)\dots f(x_n;\theta_1)dx_1\dots dx_n = \alpha \, . \tag{7.20}$$

<u>Bewijs</u>: We moeten maximaliseren

$$\varphi(\theta) = 1 - P(II) = \int\dots\int_{R_\alpha} f(x_1;\theta_2)\dots f(x_n;\theta_2)dx_1\dots dx_n$$

onder de voorwaarde (7.20).
Eerst delen we alle punten $(x_1,\dots,x_n)$ waarvoor

$$f(x_1;\theta_1)\dots f(x_n;\theta_1) = 0$$

$$f(x_1;\theta_2)\dots f(x_n;\theta_2) > 0$$

in bij R_α.
Dit heeft nl. geen invloed op (7.20) en vergroot $\varphi(\theta)$. Voor alle overige punten geldt:

$$\varphi(\theta) = \int_{R_\alpha}\!\!\!\dots\!\int f(x_1;\theta_2)\dots f(x_n;\theta_2)dx_1\dots dx_n =$$

$$= \int_{R_\alpha}\!\!\!\dots\!\int \frac{1}{\lambda} f(x_1;\theta_1)\dots f(x_n;\theta_1)dx_1\dots dx_n \ .$$

Dus $\varphi(\theta)$ is de verwachting, onder de hypothese H_1, van de functie die gelijk is aan $\frac{1}{\lambda}$ binnen R_α en gelijk aan 0 daarbuiten.
Dit wordt maximaal door R_α zo te kiezen dat erbinnen $\frac{1}{\lambda}$ groter is dan een bepaalde waarde, bv. c_α en daarbuiten kleiner.
Dus het kritieke gebied wordt gevormd door de punten waar $\frac{1}{\lambda} > c_\alpha$ of

$$\lambda < \frac{1}{c_\alpha} = k_\alpha \ .$$

Voorbeeld: Neem het geval van de normale verdeling met variantie 1. We toetsen $H_1 : \mu = \mu_1$ tegen $H_2 : \mu = \mu_2$ $(\mu_2 > \mu_1)$; $\underline{x}_1,\dots,\underline{x}_n$ is een steekproef.
Het likelihood-ratio criterium levert op:

$$\lambda = \frac{e^{-\frac{1}{2}\Sigma(x_i-\mu_1)^2}}{e^{-\frac{1}{2}\Sigma(x_i-\mu_2)^2}} = e^{-\frac{n}{2}[2\bar{x}(\mu_2-\mu_1)+\mu_1^2-\mu_2^2]} \ .$$

De kritieke zone bestaat uit kleine waarden van λ, hetgeen, omdat $\mu_2 > \mu_1$ is, overeenkomt met grote waarden van $\bar{x}$. De kritieke zone is dus:

$$\bar{\underline{x}} > c_\alpha \ ,$$

waarbij

$$P[\bar{\underline{x}} > c_\alpha \mid \mu = \mu_1] = \alpha \ .$$

Hieruit blijkt dat c_α niet van μ_2 afhangt, zolang $\mu_2 > \mu_1$.
Dit is een voorbeeld van een uniform meest onderscheidende toets (U.M.P.-test).

(7.21) Definitie: Een toets van de grootte (size) α voor de hypothese $H_1 : \theta \in \omega_1$ tegen $H_2 : \theta \in \omega_2$ heet uniform meest onderscheidend als het kritieke gebied R zodanig is dat

$$\forall_{\theta\in\omega_1} \ P(I) \leq \alpha$$

$$\forall_{\theta\in\omega_2} \ 1 - P(II) \text{ maximaal.}$$

Niet altijd zijn UMP-toetsen te vinden. Een mogelijkheid wordt soms geboden door zich te beperken tot de klasse van zuivere toetsen (unbiased tests).

(7.22) Definitie: Een toets voor $H_1 : \theta \in \omega_1$ tegen $H_2 : \theta \in \omega_2$ heet zuiver van de grootte α als

$$\forall_{\theta \in \omega_1} \quad P(I) \leq \alpha$$

$$\forall_{\theta \in \omega_2} \quad 1 - P(II) \geq \alpha .$$

Als een toets binnen deze klasse uniform meest onderscheidend is, dan heet deze toets uniform meest onderscheidend zuiver (UMPU).

Wanneer geen UMP of UMPU-toets kan worden gevonden kan men ook een geheel andere weg inslaan door een toets te nemen die voor grote steekproeven bepaalde gunstige eigenschappen en een bekende asymptotische verdeling heeft. Dit is de (gegeneraliseerde) likelihood-ratio toets. Conform de in dit verband gebruikelijke notatie schrijven we daarbij in plaats van H_1 en H_2 respectievelijk H_0 en H_a ("nul"-hypothese en alternatieve hypothese).
De nulhypothese luidt: $H_0 : \theta \in \omega$, $H_a : \theta \notin \omega$. Hierbij kan θ een vector $\theta_1, \ldots, \theta_k$ voorstellen.
De aannemelijkheidsfunctie is

$$L = \prod_{i=1}^{n} f(x_i;\theta) .$$

Het maximum over θ van L in ω noemen we $L(\hat{\omega})$ en het maximum over θ in Ω heet $L(\hat{\Omega})$.

(7.23) Definitie: De gegeneraliseerde likelihood-ratio toets voor $H_0 : \theta \in \omega$ tegen $H_a : \theta \notin \omega$ heeft als kritieke zone die steekproefwaarden waarvoor

$$0 < \lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} < A .$$

Voorbeeld: Neem de normale verdeling met gemiddelde μ en variantie σ^2.
H_0 is: $\mu = \mu_0$, $0 < \sigma^2 < \infty$, dus $H_a : \mu \neq \mu_0$, $0 < \sigma^2 < \infty$.

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-1/2\sigma^2 \, \Sigma(x_i-\mu)^2} .$$

Onder H_0 vinden we

$$\hat{\sigma}^2 = \frac{1}{n}\,\Sigma(x_i - \mu_o) \qquad \text{en} \qquad L(\hat{\omega}) = \left[\frac{n}{2\pi\,\Sigma\,(x_i - \mu_o)^2}\right]^{n/2} e^{-n/2}\ .$$

Onder H_a hebben we (vgl. pag. 31):

$$\hat{\mu} = \bar{x} \qquad \text{en} \qquad \hat{\sigma}^2 = \frac{1}{n}\,\Sigma(x_i - \bar{x})^2\ .$$

Dus
$$L(\hat{\Omega}) = \left[\frac{n}{2\pi\,\Sigma\,(x_i - \bar{x})^2}\right]^{n/2} e^{-n/2}\ .$$

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \left[\frac{\Sigma(x_i - \bar{x})^2}{\Sigma(x_i - \mu_o)^2}\right]^{n/2} = \left[\frac{1}{1 + \dfrac{n(\bar{x} - \mu_o)^2}{\Sigma(x_i - \bar{x})^2}}\right]^{n/2} = \left[\frac{1}{1 + \dfrac{t^2}{n-1}}\right]^{n/2},$$

waarin $\underline{t}$ de statistische grootheid is die we al tegenkwamen onderaan bladzijde 45.
Het kritieke gebied bestaat uit kleine waarden van λ, dus uit grote waarden van t^2, dus uitgedrukt in t bestaat het kritieke gebied uit twee gedeelten, symmetrisch t.o.v. $t = 0$.

Vraagstukken Mathematische Statistiek

1. $\underline{x}$ kan de waarden 0,1,2,.. aannemen. De voortbrengende functie van de kansverdeling is P(s).
   Wat zijn de voortbrengende functies die horen bij $\underline{x}+1$ en $2\underline{x}$?

2. P(s) is de voortbrengende functie van de rij $\{P(\underline{x}=j)\}$ $(j=0,1,2,\ldots)$.
   Wat is de voortbrengende functie van

   a) $\{P(\underline{x}\leq j)\}$
   b) $\{P(\underline{x}<j)\}$
   c) $\{P(\underline{x}\geq j)\}$
   d) $\{P(\underline{x}>j+1)\}$
   e) $\{P(\underline{x}=2j)\}$.

3. In een reeks trekkingen uit een alternatief met kans p op succes (S) en q op mislukking (F) is u_n de kans dat de eerste combinatie SF voorkomt in de $(n-1)^e$ en n^e trekking.
   Wat is de voortbrengende functie van de rij u_n, wat het gemiddelde en de variantie van de bijbehorende stochastische variabele?

4. Bereken de momenten voortbrengende functie van de verdelingsdichtheid

$$f(x)=\tfrac{1}{2}k\,e^{-x}(1+x)^2\,, \qquad -1\leq x\leq\infty$$

   en vind hieruit het gemiddelde, de variantie en het derde moment om het gemiddelde.

5. $\underline{u}_1$, $\underline{u}_2$, $\underline{u}_3$ en $\underline{u}_4$ zijn onafhankelijk verdeeld met verdelingsdichtheid

$$f(u)=e^{-u}\,, \qquad 0\leq u\leq\infty\,.$$

   Laat zien dat de simultane verdeling van

$$\underline{x}=\underline{u}_1+\underline{u}_2\,,\quad \underline{y}=\underline{u}_3+\underline{u}_2 \quad\text{en}\quad \underline{z}=\underline{u}_4+\underline{u}_2$$

   als momenten voortbrengende functie heeft:

$$\{(1-t_1-t_2-t_3)(1-t_1)(1-t_2)(1-t_3)\}^{-1}\,,$$

   waarbij t_1, t_2 en t_3 respectievelijk horen bij x, y en z.

6. Bepaal de momenten voortbrengende functie van de "dubbele exponentiële" verdeling

$$f(x) = \tfrac{1}{2}e^{-|x|} , \qquad (-\infty \leqslant x \leqslant \infty) .$$

Bewijs dat de verdeling van de som $\underline{y}$ van drie onafhankelijke trekkingen uit deze verdeling gegeven wordt door

$$g(y) = \frac{1}{16}.(3+3|y|+y^2)e^{-|y|} .$$

7. Bepaal de maximum likelihood schatter $\underline{\hat{\theta}}$ voor θ in de verdeling

$$f(t) = \frac{1}{\theta}\, e^{-t/\theta} \qquad (0 \leqslant t \leqslant \infty) ,$$

als functie van een steekproef $\underline{t}_1,\ldots,\underline{t}_n$.
Stel ook een vergelijking voor $\underline{\hat{\theta}}$ op, onder de voorwaarde dat $\underline{t} \leqslant T_0$.

8. Wat is de meest aannemelijke schatter voor θ in de verdeling

$$f(x;\theta) = \tfrac{1}{2}e^{-|x-\theta|} , \qquad -\infty \leqslant x \leqslant \infty ,$$

op grond van een steekproef $\underline{x}_1,\ldots,\underline{x}_n$?

9. Wat zijn de meest aannemelijke schatters van α en β in de verdeling

$$f(x) = \beta^{-1}(1-e^{-(\beta-\alpha)/\beta})^{-1}\, e^{-(x-\alpha)/\beta} , \qquad 0 \leqslant \alpha \leqslant x \leqslant \beta ,$$

op grond van een steekproef $\underline{x}_1,\ldots,\underline{x}_n$?

10. Laat zien dat voor de rechthoekige verdeling

$$f(x) = \frac{1}{\theta} , \qquad 0 \leqslant x \leqslant \theta ,$$

en een betrouwbaarheid $1-\alpha$, betrouwbaarheidsgrenzen voor θ zijn: $\underline{r}$ en $\underline{r}/\psi$, waarin $\underline{r}$ de steekproefrange is en ψ bepaald is door

$$\psi^{n-1}\{n-(n-1)\psi\} = \alpha .$$

11. De variabelen $\underline{x}_1$, $\underline{x}_2$ en $\underline{x}_3$ zijn onafhankelijk en ze hebben elk de verdeling

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

Wat is de simultane verdeling van de variabelen

$$\underline{y}_1 = (\underline{x}_1 - \underline{x}_2)/\sqrt{2}\ ,$$

$$\underline{y}_2 = (\underline{x}_1 + \underline{x}_2 - 2\underline{x}_3)/\sqrt{6}$$

$$\underline{y}_3 = (\underline{x}_1 + \underline{x}_2 + \underline{x}_3)/\sqrt{3}\ .$$

12. Voer in vraagstuk 11 de volgende transformatie uit:

$$\begin{aligned} x_1 &= r\cos\theta_1\cos\theta_2 \\ x_2 &= r\cos\theta_1\sin\theta_2 \\ x_3 &= r\sin\theta_1 \end{aligned} \qquad \left\{\begin{aligned} &0 \leq r \leq \infty \\ &0 \leq \theta_2 \leq 2\pi \\ &-\pi/2 \leq \theta_1 \leq \pi/2\ . \end{aligned}\right.$$

Wat zijn de kansverdelingen van $\underline{r}$, $\underline{\theta}_1$ en $\underline{\theta}_2$ en wat is de simultane verdeling?

13. Bewijs dat, als 3 punten onderling onafhankelijk homogeen verdeeld zijn over een cirkelomtrek, de kans dat ze op dezelfde halve cirkel liggen $\frac{3}{4}$ is.

14. Een steekproef van n onafhankelijke waarnemingen wordt getrokken uit een onbekende continue verdeling.
Bewijs dat de kans dat een $(n+1)^e$ waarneming groter is dan alle n voorafgaande gelijk is aan $1/(n+1)$.

15. $\underline{x}$ is normaal verdeeld met gemiddelde μ en variantie 1. Stel dat wij H_1 willen toetsen tegen H_2, waarbij

$$H_1 : \mu < 0 \qquad H_2 : \mu \geq 0\ .$$

De verliesfunctie is

$$\begin{aligned} \ell(a_1;\mu) &= 0 \ \text{ als } \mu < 0 \\ \ell(a_1;\mu) &= \mu^2 \ \text{ als } \mu \geq 0 \\ \ell(a_2;\mu) &= 0 \ \text{ als } \mu \geq 0 \\ \ell(a_2;\mu) &= \mu^2 \ \text{ als } \mu < 0\ , \end{aligned}$$

waarin a_i de beslissing "aanvaarden van H_i" betekent.
Bereken de risicofunctie voor de beide volgende beslissingsfuncties d_1 en d_2:

d_1 : aanvaard H_1 als $\underline{x} < 2$, verwerp H_1 als $\underline{x} \geq 2$

d_2 : aanvaard H_1 als $\underline{x} < 3$, verwerp H_1 als $\underline{x} \geq 3$.

Voor welke waarden van μ is d_1 beter dan d_2 ?

16. De verdelingsdichtheid van $\underline{x}$ is $f(x;\beta) = \frac{1}{\beta} e^{-x/\beta}$.
We willen de hypothese $H_1 : \beta = \beta_1$ toetsen tegen $H_2 : \beta = \beta_2$ op grond van een steekproef $\underline{x}_1,\ldots,\underline{x}_n$.
Wat is de vorm van een likelihood-ratio toets?

17. De verdelingsdichtheid van $\underline{x}$ is homogeen tussen 0 en θ. We willen $H_1 : \theta = 2$ toetsen tegen $H_2 : \theta = 2,5$, op grond van een steekproef $\underline{x}_1,\ldots,\underline{x}_n$.
De toets is

Kies H_1 als $\underline{x}_{(n)} \leq 2$.

Kies H_2 als $2 < \underline{x}_{(n)} \leq 2,5$.

Laat zien dat dit een likelihood-ratio toets is.

18. Laat zien dat de beste kritieke zone voor het toetsen van de hypothese $H_1 : \mu = \mu_1$ voor het gemiddelde van een Poissonverdeling tegen het alternatief $H_2 : \mu = \mu_2$ de volgende vorm heeft

$\bar{x} \leq a_\alpha$ als $\mu_1 > \mu_2$

$\bar{x} \geq b_\alpha$ als $\mu_1 < \mu_2$

als $\bar{x}$ het steekproefgemiddelde is en a_α en b_α constanten.

19. Laat ook zien dat voor de parameter p van een binomiale verdeling de beste kritieke zone de volgende gedaante heeft

$x \leq a_\alpha$ als $p_1 > p_2$

$x \geq b_\alpha$ als $p_1 < p_2$,

waarin x het waargenomen aantal "successen" in de steekproef is.