STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

TW 5

Probleem van der Vaart
Verdelingsfunctie van Aantal Inversies

Van der Vaart

Probleem - Van der Vaart

Notaties: Vf = Verdelingsfunctie (= kans op $\chi \leqq t$)
Wd = Waarschijnlijkheidsdichtheid
$\varepsilon, \eta, \vartheta$ zijn steeds positief
$q = 1-p$
$[x]$ = grootste gehele getal $\leqq x$.
Φ = geïntegreerde grenswaardefunctie

§1. Het aantal x-en tussen 0 en p.

Laat $x_1, \ldots, x_m$ gelijk verdeeld zijn tussen 0 en 1, dwz. met Wd 1 in de kubus $0<x_1<1, \ldots, 0<x_m<1$.

Zij p een getal tussen 0 en 1 ($0<p<1$) en stel

$A(p)$ = het aantal x-en tussen 0 en p.

Volgens E. Hopf bewijzen we nu eerst een verscherping van ~~Stelbin~~ de Wet van de Grote Getallen:

Stelling 1. De kans op $A(p) \geqq (p+\varepsilon)m$ is hoogstens $e^{-\vartheta m}$, waarin ϑ ~~xxx~~ alleen van ε afhangt.

Bewijs. Laat $z_\nu = 1$ of 0 zijn naarmate $x_\nu < p$ of niet ($\nu = 1, \ldots, m$). Beschouw de toevallige grootheid

$$e^{\lambda A(p)} = e^{\lambda(z_1+z_2+\ldots+z_m)} = e^{\lambda z_1} e^{\lambda z_2} \ldots e^{\lambda z_m}$$

De middelwaarde van dit product is, doordat de factoren onafhankelijk zijn, het product der middelwaarden

$$E\{e^{\lambda A(p)}\} = (p e^{\lambda} + q)^m$$

Gesteld nu dat de kans op $A(p) \geqq (p+\varepsilon)m$ groter was dan $e^{-\vartheta m}$, dan was het linkerlid groter dan

$$e^{-\vartheta m} e^{\lambda(p+\varepsilon)m}$$

dus was

$$e^{-\theta m} e^{\lambda(p+\varepsilon)m} < (pe^{\lambda}+q)^m$$

$$e^{-\theta} e^{\lambda(p+\varepsilon)} < pe^{\lambda}+q$$

$$e^{-\theta} e^{\lambda\varepsilon} < pe^{\lambda q}+qe^{-\lambda p}$$

Nu is voor $0<\lambda<1$

$$e^{\lambda q} < 1+\lambda q+\frac{\lambda^2 q^2}{2}\left(1+\tfrac{1}{2}+\tfrac{1}{4}+\ldots\right) = 1+\lambda q+\lambda^2 q^2$$

$$e^{-\lambda p} < 1-\lambda p+\tfrac{1}{2}\lambda^2 q^2 < 1-\lambda p+\lambda^2 q^2$$

dus

$$e^{-\theta} e^{\lambda\varepsilon} < p(1+\lambda q+\lambda^2 q^2)+q(1-\lambda p+\lambda^2 q^2)$$

$$\longleftarrow \quad = 1+\lambda^2 pq < e^{\lambda^2 pq}$$

dus

$$-\theta+\lambda\varepsilon < \lambda^2 pq$$

Kies nu bv.

(1) $\theta = \varepsilon^2$ en $\lambda = 2\varepsilon$

dan komt er

$$\varepsilon^2 < 4pq\varepsilon^2$$

wat onmogelijk is, daar $4pq$ hoogstens 1 bedraagt. Hiermee is Stelling 1 bewezen.

Zij P de kans dat voor alle p $(0<p<1)$ het aantal $A(p)$ tussen $(p-\varepsilon)m$ en $(p+\varepsilon)m$ ligt:

(2) $$(p-\varepsilon)m < A(p) < (p+\varepsilon)m$$

We bewijzen nu voor willekeurige $\eta>0$:

<u>Stelling 2. Voor voldoende grote m is $P>1-\eta$</u>

<u>Bewijs</u>. Als de ongelijkheid

$$A(p) < (p+\varepsilon)m$$

vervuld is voor de waarden $p = \frac{q}{m}-\varepsilon$

$(g = 1, 2, \ldots, m)$, waarvan $(p+\varepsilon)m$ een geheel getal g is, dan is de ongelijkheid vanzelf voor alle p vervuld. Immers als dit niet zo was, dan was $A(p) \geqq (p+\varepsilon)m$ voor een niet gehele waarde van $(p+\varepsilon)m$, maar dan was $A(p)$ ook minstens gelijk aan het naast grotere gehele getal $(p'+\varepsilon)m$, waarin $p' > p$ is. Maar $A(p')$ is weer minstens gelijk aan $A(p)$, dus was

$$A(p') \geqq (p'+\varepsilon)m \text{ voor gehele } (p'+\varepsilon)m$$

wat strijdt tegen de veronderstelling

$$A(p') < (p'+\varepsilon)m \text{ voor gehele } (p'+\varepsilon)m.$$

Hetzelfde geldt voor de ongelijkheid

$$A(p) > (p-\varepsilon)m$$

Als dus aan de ongelijkheden (2) niet is voldaan, dan is er òf een geheel getal $g = (p+\varepsilon)m$, zo dat $A(p) \geqq g$, òf een geheel getal $g = (p-\varepsilon)m$, zo dat $A(p) \leqq g$.

Nu is de kans op $A(p) \geqq g = (p+\varepsilon)m$ volgens Stelling 1 hoogstens $e^{-\theta m}$. Dit geldt voor elke waarde $g = 1, 2, \ldots, m$; de totale kans op een ongelijkheid $A(p) \geqq (p+\varepsilon)m$ is dus hoogstens $m e^{-\theta m}$. Hetzelfde geldt voor $A(p) \leqq (p-\varepsilon)m$. De totale kans, dat (2) niet vervuld is, is dus hoogstens

$$2 m e^{-\theta m} < \eta \qquad (3)$$

Hiermee is Stelling 2 bewezen.

Laat nu de x-en volgens de grootte gerangd en omgenummerd worden:

$$x_1 < x_2 < \dots < x_m$$

Dan volgt uit Stelling 2 rijpel onmiddellijk

Stelling 3. De kans dat voor alle g

(4) $$\frac{g}{m} - \varepsilon \leqq x_g \leqq \frac{g-1}{m} + \varepsilon$$

is groter dan $1-\eta$, als m groot genoeg is.

Bewijs. Als (4) niet vervuld is, dan is óf een $x_g < \frac{g}{m} - \varepsilon$ óf een $x_g > \frac{g-1}{m} + \varepsilon$. In het eerste geval is het aantal x-en kleiner dan $p = \frac{g}{m} - \varepsilon$ namelijk $g = (p+\varepsilon)m$. In het tweede geval is het aantal x-en kleiner dan $p = \frac{g-1}{m} + \varepsilon$ namelijk $g-1 = (p-\varepsilon)m$. In beide gevallen komen we in strijd met de ongelijkheid (2), en de kans daarop is $Q = 1 - P < \eta$.

§2. De verdelingsfunctie van V voor grote m.

Laat $x_1, \dots, x_m$ onafhankelijke zijn en een verdelingsfunctie $F(x)$ hebben, en laat $y_1, \dots, y_m$ weer onafhankelijk (ook van de x) zijn en elk een verdelingsfunctie $G(y) = F(y-\mu)$ hebben. Laat V het aantal non-inversies zijn, dwz. het aantal malen dat een y na een x komt. Gevraagd wordt de Vf van V. (aldus stijgende)

Neem aan dat F een continue, Vf

is. Door invoering van $X = F(x)$ als nieuwe veranderlijke in plaats van x kan men bereiken, dat de X_ν gelijk verdeeld zijn tussen 0 en 1. De Vf van $Y_\nu = F(y_\nu)$ wordt dan, als F^{-1} de omkeerfunctie van F is,

$$(5) \qquad H(Y) = G(F^{-1}Y)$$

De veranderlijken X en Y lopen nu slechts van 0 tot 1: voor $Y<0$ stellen we $H(Y)=0$, voor $Y \geqq 1$ stellen we $H(Y)=1$. Dat wil zeggen: de Y_ν die negatief zijn maken we 0, en die >1 zijn maken we 1. Voor het aantal non-inversies maakt dit niets uit, omdat de x en toch tussen 0 en 1 liggen.

Uit de definitie (5) volgt, dat $H(Y)$ voor $0<Y<1$ continu is.

Bijzondere gevallen:

1) Als $F = G$ is (dus $\mu = 0$), wordt $H(Y) = Y$.

2) Als de x-en van te voren al gelijk verdeeld waren, dan is $H(Y) = Y - \mu$ voor $0 \leqq Y < 1$.

3) Als de x-en een Gauss verdeling hebben

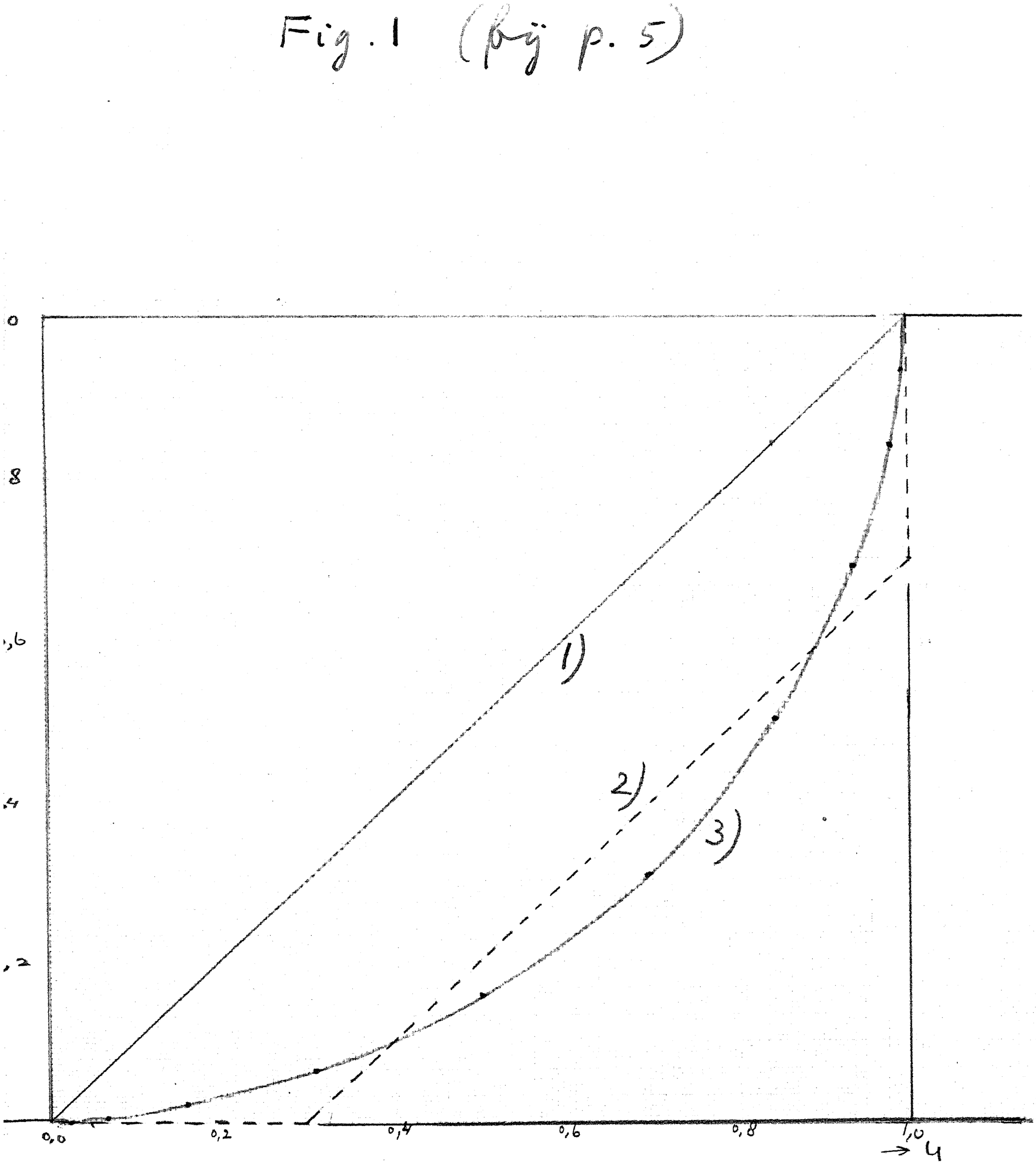
$$F(x) = \Phi(x) = \int_{-\infty}^{x} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}t^2}\, dt$$

dan wordt de kromme $Z = H(Y)$ voorgesteld door de parametervoorstelling

$$Y = \Phi(y), \quad Z = \Phi(y-\mu)$$

In Fig. 1 zijn de krommen $Z = H(Y)$ voor de 3 gevallen getekend. Het belangrijkste is geval 1), maar de volgende afleiding geldt voor willekeurige $H(Y)$,

Fig. 1 (Fig p. 5)



die voor $0<Y<1$ continu zijn.

Aan de uiteinden $Y=0$ en $Y=1$ behoeft $H(Y)$ niet continu te zijn, maar wel heeft $H(Y)$, als monotone Vf, steeds grenswaarden $H(+0)$ en $H(1-0)$. Neemt men deze erbij, dan wordt $H(Y)$ continu voor $0 \leqq Y \leqq 1$ en dus gelijkmatig continu:

$$(6) \quad |H(s)-H(t)|<\varepsilon \text{ voor } |s-t|<\delta$$

We bepalen nu de Vf van V eerst voor het geval dat de x-en worden vastgehouden. Laat het aantal x-en zijn, dat aan y_1 voorafgaat. Dan zijn $V_1, V_2, \ldots, V_n$ onafhankelijk. De Vf van V is dus de Vf van een som van n onafhankelijke termen, die elk dezelfde Vf hebben. We noemen deze Vf $K_1(v)$.

$K_1(v)$ is de kans dat $V_1 \leqq v$. Voor negatieve v is deze kans 0, voor $v \geqq m$ is de kans 1. Neem dus aan $0 \leqq v < m$. De ongelijkheid $V_1 \leqq v$ betekent dan dat er hoogstens $g=[v]$ van de x-en kleiner zijn dan y_1, dus dat x_{g+1} niet kleiner is dan y_1:

$$y_1 \leqq x_{g+1}$$

De kans hiervoor is $H(x_{g+1})$, dus is

$$(7) \quad K_1(v) = H(x_{g+1}), \quad g=[v]$$

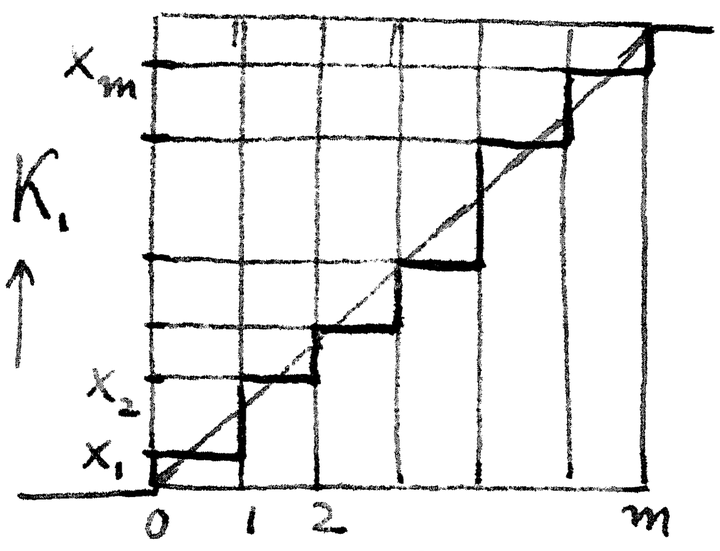


Fig. 2 De Vf van V

Voor het geval $H(t)=t$ is deze Vf in fig. 2 getekend.

Nu onderstellen we, dat de x-en aan de ongelijkheden (2) en dus ook aan (4) voldoen. We laten dus de genallen

met waarschijnlijkheid $Q < \eta$, die niet aan (2) voldoen, buiten beschouwing. Daardoor vervangen we in (2) en (4) de willekeurige ε door een andere constante δ, die ~~van~~ door (6) gedefinieerd wordt. We kiezen dus eerst ε en η willekeurig, bepalen ~~δ door (6) en θ door (4) en~~ δ volgens (6), stellen $\theta = \delta^2$ naar analogie van (1), en kiezen m zo groot dat aan (3) voldaan is.

Volgens (4) is van $g = 0, 1, \ldots, (m-1)$

$$\frac{g+1}{m} - \delta \leqq x_{g+1} \leqq \frac{g}{m} + \delta$$

Neemt men hierin $g = [v]$ en maakt het linkerlid iets kleiner, het rechterlid iets groter, dan volgt

$$\frac{v}{m} - \delta \leqq x_{g+1} \leqq \frac{v}{m} + \delta$$

Hiermit volgt volgens (6)

$$|H(x_{g+1}) - H\left(\tfrac{v}{m}\right)| < \varepsilon$$

dus wegens (7)

(8) $$|K_1(v) - H\left(\tfrac{v}{m}\right)| < \varepsilon$$

Dus: <u>Als $x_1, \ldots, x_m$ aan (2) voldoen met δ in plaats van ε, dan wordt de Vf van V_1 met een fout $< \varepsilon$ benaderd door de bekende functie $H\left(\tfrac{v}{m}\right)$.</u>

In geval 1) is $H\left(\tfrac{v}{m}\right) = \tfrac{v}{m}$; de grafiek van fig. 2 wordt dus gelijkmatig benaderd door de gebekende diagonaal.

Beschouw nu het geval $n=2$. De Vf van $V=V_1+V_2$ is

$$K_2(v) = \int K_1(v-u)\, d K_1(u) \tag{9}$$

Vervangen we hierin de eerste factor K_1 door de benadering $H\left(\frac{v-u}{m}\right)$, dan maken we een fout kleiner dan

$$\int \varepsilon\, d K_1(u) = \varepsilon$$

In het resultaat kunnen de functies H en K_1 ook verwisseld worden:

$$\int H\left(\frac{v-u}{m}\right) d K_1(u) = \int K_1(v-u)\, d H\left(\frac{u}{m}\right)$$

Vervangen we nu weer $K_1(v-u)$ door $H\left(\frac{v-u}{m}\right)$, dan maken we weer een fout $<\varepsilon$. Zo krijgen we voor $K_2(v)$ de benadering

$$H_2\left(\frac{v}{m}\right) = \int H\left(\frac{v-u}{m}\right) d H\left(\frac{u}{m}\right)$$

met een fout $<2\varepsilon$.

Op deze wijze kunnen we doorgaan. We definiëren een reeks functies $H_n(t)$ door

$$H_1(t) = H(t)$$

$$H_n(t) = \int_{-0}^{t} H_{n-1}(t-s)\, d H(s) \tag{10}$$

<u>Dan wordt voor willekeurige n de Vf van $V=V_1+\ldots+V_n$ benaderd door $H_n\left(\frac{v}{m}\right)$ met een fout $<n\varepsilon$.</u> Zolang n vastgehouden wordt, is deze fout willekeurig klein:

$$n\varepsilon = \eta \quad \text{voor} \quad \varepsilon = \frac{\eta}{n} \tag{11}$$

Tot dusver werden de x-en vastgehouden. De overeenkomstige verdelingsfunctie $K_n(x|v)$ hangt

dus van $x_1, \dots, x_n$ en v af, maar wordt door een van de x onafhankelijke functie H_n benaderd. De benadering geldt voor alle x, die aan (2) met δ voldoen.

Nu gaan we over alle waarden van de x-en integreren. De uiteindelijke ~~Kf~~ Vf van V, de kans dat $V \leqq v$ is, wordt dan

$$(12) \qquad L(v) = \int K_n(x|v)\, dx_1 \dots dx_n$$

geïntegreerd over de eenheidskubus. Deze kubus bestaat uit een deelverzameling A, waarop (2) geldt, en een rest B, waarop (2) niet geldt. Op A kunnen we K_n door $H_n\left(\frac{v}{m}\right)$ vervangen; we maken dan een fout $< n\varepsilon = \eta$ in de integrand en dus ook in de integraal. Op B kunnen we ook K_n door H_n vervangen; we maken dan een fout kleiner dan 1 geïntegreerd over B, dus een fout kleiner dan het volume van B. Dan het volume van B is $< \eta$, zodra m ~~voldoende~~ groot genoeg is. In totaal maken we dus een fout $< 2\eta$, als we in de hele kubus K_n door H_n vervangen. De integraal wordt dan $H_n\left(\frac{v}{m}\right)$.

Hiermee is bewezen:

Stelling 4. Bij gegeven n en bij voldoende grote m kan de Vf van V met een willekeurig kleine fout benaderd worden door $H_n\left(\frac{v}{m}\right)$, waarin H_n door de recursieformule (10) gedefinieerd wordt.

We willen nu even nagaan, hoe groot m bij gegeven η gekozen moet worden. Daarbij beperken we ons eenvoudigheidshalve tot de gevallen 1) en 2), waarin $\delta = \varepsilon$. Dan is volgens (1) en (11)

$$\theta = \varepsilon^2 = \frac{\eta^2}{n^2}$$

dus luidt de voorwaarde (3)

(13) $$2me^{-\frac{\eta^2 m}{n^2}} < \eta$$

Dit vereist minstens dat $\eta^2 m$ groot moet zijn t.o.v. n^2, dus m zeer groot t.o.v. n^2. Onze schatting geldt dus niet voor het geval dat m en n dezelfde orde van grootte hebben. Dit geval vereist een afzonderlijk onderzoek.

§3. De verdelingsfunctie van V voor grote m en n.

$K_1(v)$ is een Vf met eindige middelwaarde M_x en variantie σ^2. Hiermit volgt volgens de „centrale limietstelling", dat $V = V_1 + \dots + V_n$, altijd nog bij vastgehouden $x_1, \dots, x_m$, voor grote n bijna een GauBverdeling heeft.

Hoe goed deze benadering zelfs voor matige n is, dat kan men goed beoordelen door $K_1(v)$ te vervangen door de functie $H(\frac{v}{m})$, die er immers slechts weinig van verschilt. In geval 1) is $H(t) = t$, en het zal later blijken dat de Vf van een som van 3 of 4 termen, die elk tussen 0 en 1 gelijkverdeeld zijn, praktisch niet van een GauBverdeling te onderscheiden is. Voor kleine n zal het in de gevallen 2) en 3) wel net zo gaan. Voor grote n kunnen we dus veilig $K_n(v)$ door een GauBverdeling vervangen, met middelwaarde nM_H en variantie $n\sigma^2$:

$$(14) \qquad K_n(v) \sim \Phi\left(\frac{v - nM_x}{\sigma\sqrt{n}}\right) \text{ met fout} < \tfrac{1}{2}\eta$$

We moeten nu ~~zorg~~ M en σ^2 als functies van $x_1, \dots, x_m$ exact uitrekenen en dan de integratie (12) uitvoeren:

$$(15) \qquad L(v) \sim \int \Phi\left(\frac{v - nM}{\sigma\sqrt{n}}\right) dx_1 \cdots dx_n \quad \text{met fout} < \tfrac{1}{2}\eta$$

De waarden, die V_1 aan kan nemen, zijn $g = 0, 1, \dots, m$, met waarschijnlijkheden

$$H(x_1),\ H(x_2) - H(x_1),\ \dots,\ 1 - H(x_m)$$

Het eerste en tweede moment zijn dus

$$M = \sum_{1}^{m-1} g\left\{H(x_{g+1}) - H(x_g)\right\} + m\left\{1 - H(x_m)\right\}$$

$$M_2 = \sum_{1}^{m-1} g^2\left\{H(x_{g+1}) - H(x_g)\right\} + m^2\left\{1 - H(x_m)\right\}$$

en de variantie is

$$\sigma^2 = M_2 - M^2$$

Voor het gemak nemen we $H(x) = x$ aan (§ 2, geval 1). Dan is

$$M = (x_2 - x_1) + 2(x_3 - x_2) + \ldots + (m-1)(x_m - x_{m-1}) + m(1 - x_m)$$

$$= m - x_1 - x_2 - \ldots - x_m = m - \sum x_g$$

$$M_2 = (x_2 - x_1) + 2^2(x_3 - x_2) + \ldots + (m-1)^2(x_m - x_{m-1}) + m^2(1 - x_m)$$

~~$= m^2 - x_1 - 3x_2 - 5x_3 - \ldots - (2m-1)x_m$~~

$$= -x_1 + (1^2 - 2^2)x_2 + (2^2 - 3^2)x_2 + \ldots + \left\{(m-1)^2 - m^2\right\}x_m + m^2$$

$$= m^2 - x_1 - 3x_2 - 5x_3 - \ldots - (2m-1)x_m$$

Nu stellen we

(16) $$x_g = \frac{g - \frac{1}{2}}{m} + u_g$$

waarin $|u_g| < \varepsilon$ volgens (4). Dan wordt

$$M = \frac{1}{2}m - \sum_{1}^{m} u_g = \frac{1}{2}m - a$$

$$M_2 = \frac{2m^2+1}{6} - \sum_{1}^{m}{}' (2g-1)u_g = \frac{2m^2+1}{6} - b$$

(17) $$\sigma^2 = M_2 - M^2 = \frac{m^2}{12} + (ma - b) - a^2 + \frac{1}{6}$$

Hierin is

$$ma - b = (m-1)u_1 + (m-3)u_2 + \ldots + (3-m)u_{m-1} + (1-m)u_m$$

dus, als alle $|u_g| < \varepsilon$ zijn,

$$|ma - b| < \tfrac{1}{2}m^2\varepsilon$$

en

$$a^2 = \left(\sum u_g\right)^2 < m^2\varepsilon^2$$

Dus levert (17)

$$\frac{n^2}{12} - \frac{1}{2} n^2 \varepsilon - n^2 \varepsilon^2 < \sigma^2 < \frac{n^2}{12} + \frac{1}{2} n^2 \varepsilon + \frac{1}{6}$$

Hieruit volgt gemakkelijk, als $\varepsilon < \frac{1}{15}$ en $n\varepsilon > 4$ is,

$$\frac{n}{\sqrt{12}} (1 - 4\varepsilon) < \sigma < \frac{n}{\sqrt{12}} (1 + 3\varepsilon)$$

Stelt men nu

$$\sigma_0 = \frac{n}{\sqrt{12}}$$

dan volgt $\sigma = \sigma_0 \beta$ met

~~$\sigma = \sigma_0 \beta$ met~~ $1 - 4\varepsilon < \beta < 1 + 4\varepsilon$

Nu gaan we na, welke fout we in de integraal (15) ~~integrand~~ maken, als we in de integrand σ door σ_0 vervangen. De integrand is

$$\Phi\left(\frac{v - n M_x}{\sigma \sqrt{n}}\right) = \Phi\left(\frac{v - n M}{\sigma_0 \beta \sqrt{n}}\right) = \Phi\left(\frac{t}{\beta}\right)$$

Differentiëren we dit naar β, dan komt er

$$\frac{d}{d\beta} \Phi\left(\frac{t}{\beta}\right) = -\frac{1}{\beta \sqrt{2\pi}} \frac{t}{\beta} e^{-\frac{1}{2} \frac{t^2}{\beta^2}}$$

Nu wordt de functie $x e^{-\frac{1}{2} x^2}$ maximaal groter dan $e^{-\frac{1}{2}}$, dus is voor $\beta > 1 - 4\varepsilon$ en $\varepsilon < \frac{1}{4} - \frac{1}{\sqrt{2\pi}}$

$$\left| \frac{d}{d\beta} \Phi\left(\frac{t}{\beta}\right) \right| < \frac{1}{\beta \sqrt{2\pi e}} < \frac{1}{(1 - 4\varepsilon)\sqrt{2\pi e}} < \frac{1}{4}$$

Dit geldt ook, als β vervangen wordt door een getal tussen 1 en β. Dus is volgens de middelwaardestelling der differentiaalrekening

$$(18) \quad \left| \Phi\left(\frac{t}{\beta}\right) - \Phi(t) \right| < |\beta - 1| \cdot \frac{1}{4} < 4\varepsilon \cdot \frac{1}{4} = \varepsilon$$

Om nu de integraal (15) te benaderen, verdelen we weer de kubus in de deelverzamelingen A en B als in §2. Op A gelden (2) en (4), dus ook (18). Vervangen we dus op A de σ door σ_0, dan maken we een fout $< \varepsilon$. Als we ook op B de noemer $\sigma\sqrt{n}$ door $\sigma_0\sqrt{n}$ vervangen, dan maken we een fout $< \eta$, omdat B een volume $< \eta$ heeft. Kiezen we $\varepsilon = \eta$, dan krijgen we ~~tot staat~~ dus voor L de benadering

$$(19) \qquad L(v) \sim \int \Phi\left(\frac{v - n\,M}{\sigma_0\sqrt{n}}\right) dx_1 \cdots dx_n$$

met een fout kleiner dan

$$\varepsilon + \eta + \tfrac{1}{2}\eta = 2\tfrac{1}{2}\eta$$

Laat nu $\Psi(t)$ de verdelingsfunctie van $M = m - \sum x_g$ zijn. Dan kunnen we voor (19) ook schrijven

$$(20) \qquad L(v) \sim \int_0^m \Phi\left(\frac{v - n\,t}{\sigma_0\sqrt{n}}\right) d\Psi(t)$$

Immers zowel ~~(19) als (20)~~ het rechterlid van (19) als dat van (20) stellen de verwachtingswaarde

$$E\left\{\Phi\left(\frac{v - n\,t}{\sigma_0\sqrt{n}}\right)\right\}$$

voor.

N_n is $\sum x_{g\nu}$ een som van m onafhankelijke termen, die alle gelijkverdeeld zijn tussen 0 en 1. De functie Ψ is dus voor grote m met zeer goede benadering een verdelingsfunctie van Gauß. De middelwaarde van

$M = m - \sum x_j$ is $\frac{1}{2}m$, de variantie is

$$\sigma_1^2 = \frac{m}{12}$$

immers de variantie van elke x_j is $\frac{1}{12}$. Dus kunnen we $\Psi(t)$ benaderen door $\Phi\left(\frac{t-\frac{1}{2}m}{\sigma_1}\right)$ met een fout $< \frac{1}{2}\eta$. Zo vinden we voor $L(v)$ de benadering

$$(21) \qquad L(v) \sim \int_{-\infty}^{\infty} \Phi\left(\frac{v-nt}{\sigma_0 \sqrt{n}}\right) d\Phi\left(\frac{t-\frac{1}{2}m}{\sigma_1}\right)$$

met een fout $< 3\eta$. ~~Stellen~~

Stellen we

$$\sigma_2 = \frac{\sigma_0}{\sqrt{n}} = \frac{m}{\sqrt{12n}} \quad \text{en} \quad w = \frac{v}{n}$$

dan kunnen we voor (21) ook schrijven

$$(22) \qquad L(v) \sim \int_{-\infty}^{\infty} \Phi\left(\frac{w-t}{\sigma_2}\right) d\Phi\left(\frac{t-\frac{1}{2}m}{\sigma_1}\right)$$

(22) betekent, dat $\frac{V}{n}$ bij benadering dezelfde ~~verdeling~~ Vf heeft als een som van 2 onafhankelijke, volgens Gauß verdeelde grootheden S en T, waarvan S de middelwaarde 0 en variantie σ_2^2 heeft, T de middelwaarde $\frac{1}{2}m$ en variantie σ_1^2. Maar dan heeft de som $S+T$ ook een Gaußverdeling, met middelwaarde $\frac{1}{2}m$ en variantie

$$\sigma_3^2 = \sigma_1^2 + \sigma_2^2 = \frac{m}{12} + \frac{m^2}{12n} = \frac{m(m+n)}{12n}$$

Dus heeft V zelf een benaderende Gaußverdeling (fout $< 3\eta$ in de Vf) met middelwaarde $\frac{1}{2}mn$ en variantie $\frac{1}{12}mn(m+n)$. Dit alles geldt voor

$H(t) = t$, voor matige of grote n en zeer grote m.

Als we de veronderstelling $H(t) = t$ laten vallen, ~~dan~~ wordt de berekening ingewikkelder, maar het principe blijft het zelfde. We moeten nu als in § 2 tussen ε en δ onderscheiden, waarin δ bij gegeven ε door (6) gedefinieerd wordt en m aan de voorwaarde (3) met $\theta = \delta^2$ onderworpen wordt. In plaats van (16) stellen we nu

$$H(x_g) = H\left(\frac{2g-1}{2m}\right) + u_g$$

Dan is volgens (4) met δ in plaats van ε

$$\left| x_g - \frac{2g-1}{2m} \right| < \delta$$

dus volgens (6)

$$|u_g| < \varepsilon$$

De uitdrukkingen voor M en M_2 worden nu

$$M = m - \sum_1^m H\left(\frac{2g-1}{2m}\right) - \sum u_g$$

$$M_2 = m^2 - \sum_1^m (2g-1) H\left(\frac{2g-1}{m}\right) - \sum (2g-1) u_g$$

Deze sommen kunnen voor grote m worden benaderd door integralen

$$\sum_1^m H\left(\frac{2g-1}{2m}\right) \sim m \int_0^1 H(t)\,dt$$

$$\sum_1^m (2g-1) H\left(\frac{2g-1}{2m}\right) \sim 2m^2 \int_0^1 t\,H(t)\,dt$$

Door partiële integratie volgt hieruit

$$m - \sum_1^m H\left(\frac{2g-1}{2m}\right) \sim m \int_0^1 t\,dH(t) = m\,N$$

$$m^2 - \sum_1^m (2g-1) H\left(\frac{2g-1}{2m}\right) \sim m^2 \int_0^1 t^2\,dH(t) = m^2 N_2$$

Zo vinden we dus

$$M \sim m N - \sum u_g = m N - a$$

$$M_2 \sim m^2 N - \sum (2g-1) u_g = m^2 N_2 - b$$

en hiermit analoog (17)

$$\sigma^2 = M_2 - M^2 \sim m^2 (N_2 - N^2) + (2mNa - b) + a^2 \tag{23}$$

De coëfficiënt van m^2 is juist de variantie van een grootheid Y met de Vf $H(t)$:

$$N_2 - N^2 = \int_0^1 (t-N)^2 \, dH(t) = \sigma_Y^2$$

De tweede term in (23)

$$2mNa - b = 2mN \sum u_g - \sum (2g-1) u_g$$

is van de orde van grootte εm^2, omdat de $|u_g|$ alle $< \varepsilon$ zijn. De derde term

$$a^2 = \left(\sum u_g \right)^2$$

is kleiner dan $\varepsilon^2 m^2$, dus veel kleiner dan de tweede term. Zo vinden we meer

$$\sigma^2 \doteq m^2 \sigma_Y^2 + \mathcal{O}(\varepsilon m^2)$$

$$\sigma = m \sigma_Y \{ 1 + \mathcal{O}(\varepsilon) \}$$

waarin $\mathcal{O}(\varepsilon)$ een correctieterm van de orde ε (dwz. absoluut kleiner dan $C\varepsilon$) is.

Hiermit volgt weer, dat we in de integrand van (15) een factor van slechts $\mathcal{O}(\varepsilon)$ maken, als we σ door $m\sigma_Y$ vervangen. Zo vinden we

$$L(v) \sim \int \Phi \left(\frac{v - nM}{m \sigma_Y \sqrt{n}} \right) dx_1 \dots dx_n$$

$$= \int \Phi \left(\frac{v - nt}{m \sigma_Y \sqrt{n}} \right) d\Psi(t)$$

waarin Ψ de verdelingsfunctie van

$$M = m - \sum H(x_g) = \sum \{1 - H(x_g)\}$$

is. Nu zijn de termen $H(x_g)$ weer onafhankelijke grootheden met middelwaarde

$$\int_0^1 \{1 - H(x)\}\, dx = \int_0^1 x\, dH(x) = N$$

en met variantie

$$\sigma_H^2 = \int_0^1 \{1 - H(x)\}^2\, dx - N^2$$

De som M heeft dus bij benadering een Gaußverdeling met middelwaarde mN en variantie $m\sigma_H^2$. Stelt men dus weer

$$w = \frac{v}{n}\,, \quad \sigma_1 = \sigma_H \sqrt{m}\,, \quad \sigma_2 = \frac{m}{\sqrt{n}}\,\sigma_Y$$

dan vindt men voor $L(v)$ de benadering

$$L(v) \sim \int \Phi\left(\frac{w - t}{\sigma_2}\right) d\,\Phi\left(\frac{t - mN}{\sigma_1}\right)$$

die als functie van w weer een Gauß-verdeling voorstelt, met middelwaarde mN en variantie

$$\sigma_3^2 = \sigma_1^2 + \sigma_2^2 = m\sigma_H^2 + \frac{m^2}{n}\,\sigma_Y^2$$

Voert men nu weer $v = nw$ in plaats van w in, dan wordt de middelwaarde mnN en de variantie

$$mn^2\sigma_H^2 + m^2 n\sigma_Y^2 = mn\left(n\sigma_H^2 + m\sigma_Y^2\right)$$

Hiermit volgt dus:

Stelling 5. Voor natuurlijke getallen n en m heeft V een benaderende Gaußverdeling met middelwaarde mnN

en variantie

(24) $$mn\,(m\,\sigma_Y^2 + n\,\sigma_H^2)$$

Hierin betekenen N.
en σ_Y de middelwaarde en standaard-
afwijking van een grootheid Y met Vf $H(t)$,
en σ_H de standaardafwijking van een
grootheid $H(x)$, waarin x tussen 0 en 1
gelijkverdeeld is. In geval 1) is $\sigma_Y^2 = \sigma_H^2 = \frac{1}{12}$.

§4. Middelwaarde en Standaardafwijking van V.

Een middel om de benaderingen van §2 en van §3 te toetsen is, hun middelwaarde en variantie te vergelijken met de exakt berekende middelwaarde en variantie van V.

Laat z_{ik} een toevallige grootheid zijn, die #1 is wanneer $x_i \leqq y_k$, en 0 wanneer $x_i > y_k$.
De kans op $x_i \leqq y_k$ is, als y_k vastgehanden wordt, juist $F(y_k)$. Gemiddeld over alle y_k geeft dit

$$p = \int_{-\infty}^{\infty} F(y)\, d\,G(y)$$

Voert men $t = F(y)$ als nieuwe veranderlijke in, dan wordt dit

$$p = \int_0^1 t\, d\,H(t)$$

Dit is dezelfde middelwaarde, die in §3
N werd genoemd. De middelwaarde

van z_{ik} is dus

$$\hat{z}_{ik} = p \cdot 1 + q \cdot 0 = p$$

en de middelwaarde z_{ik}^2 is eveneens p.

Nu is

$$V = \sum_i \sum_k z_{ik}$$

Dus is de middelwaarde

(25) $$\hat{V} = m n \hat{z}_{ik} = m n p$$

Verder is

$$V^2 = \sum_i \sum_j \sum_k \sum_l z_{ik} z_{jl}$$

dus is de verwachtingswaarde van V^2

$$E(V^2) = \sum_i \sum_j \sum_k \sum_l E(z_{ik} z_{jl})$$

In deze som zijn nm ~~termen zoals~~ gelijke termen met $i=j$, $k=l$

$$E(z_{ik}^2) = p$$

verder $mn(n-1)$ termen met $k \neq l$

$$E(z_{ik} z_{il}) = P(x_i \leqq y_k \, , \, x_i \leqq y_l) = A$$

en $m(m-1)n$ termen met $i \neq j$

$$E(z_{ik} z_{jk}) = P(x_i \leqq y_k \, , \, x_j \leqq y_k) = B$$

Ten slotte $m(m-1)n(n-1)$ termen met $i \neq j$, $k \neq l$

$$E(z_{ik} z_{jl}) = \hat{z}_{ik} \hat{z}_{jl} = p^2$$

Dus wordt

$$E(V^2) = mnp + mn(n-1)A + m(m-1)nB + m(m-1)n(n-1)p^2$$

De variantie van V wordt dus

(26) $$\sigma_V^2 = E(V^2) - (mnp)^2$$

$$= mn\{pq + (n-1)(A-p^2) + (m-1)(B-p^2)\}$$

In het geval $F = G$ is elke volgorde van 2 of 3 grootheden zoals x_i, y_k, y_l even waarschijnlijk, dus

$p = \frac{1}{2}$
$A = B = \frac{1}{3}$
$\hat{V} = \frac{1}{2} mn$

$$\sigma_V^2 = mn\left(\frac{1}{4} + \frac{n-1}{12} + \frac{m-1}{12}\right) = mn\,\frac{m+n+1}{12}$$

De benaderingen van §2 ~~geeft~~ en §3 geven beide de juiste middelwaarde mnp. Die van §2 geeft voor $F = G$ de variantie

$$\sigma_V^2 \approx \frac{m^2 n}{12} \qquad \text{voor } m \gg n^2$$

Die van §3 geeft de veel betere benadering

$$\sigma_V^2 \approx mn\,\frac{m+n}{12}$$

Voor willekeurige F en G berekenen we A en B op dezelfde manier als zojuist p.
Voor vaste y_k is de kans op $x_i \leqq y_k$ en $x_j \leqq y_k$

$$F(y_k)^2$$

Gemiddeld over alle y_k geeft dit

$$B = \int_{-\infty}^{\infty} F(y)^2\, dG(y) = \int_0^1 t^2\, dH(t)$$

Dit is juist het tweede moment N_2 van §3. De coëfficiënt $B - p^2$ in (26) is dus

$$B - p^2 = N_2 - N^2 = \sigma_V^2$$

Door verwisseling van de x en y, van $\leqq$ en $\geqq$ vinden we op dezelfde wijze

$$A = \int_{-\infty}^{\infty} \{1 - G(y)\}^2 \, dF(y)$$

of, wanneer $t = F(y)$ als nieuwe veranderlijke wordt gekozen,

$$A = \int_0^1 \{1 - H(t)\}^2 \, dt$$

Dus is A het tweede moment van een grootheid $H(Y)$, als Y zelf gelijk verdeeld is tussen 0 en 1. Het eerste moment

$$\int_0^1 \{1 - H(t)\} \, dt$$

is juist p, zoals door partiële integratie onmiddellijk blijkt:

$$\int_0^1 \{1 - H(t)\} \, dt = \int_0^1 t \, dH(t) = p$$

Dus is de spreiding

$$\sigma_H^2 = A - p^2$$

Met deze notatie wordt (26)

$$(27) \quad \sigma_V^2 = mn\{pq + (n-1)\sigma_H^2 + (m-1)\sigma_Y^2\}$$

Vergelijken we dit met (24)

$$\sigma_V^2 \sim mn(n\sigma_Y^2 + m\sigma_H^2)$$

dan zien we, dat de benadering van § 3 wat de variantie betreft goed is.