MATHEMATISCH CENTRUM,
Rekenafdeling,
A m s t e r d a m -O.

Mededeling MR 3.
Tevens ZW-(1950)-001.

AFRONDINGSFOUTEN

Door A. van Wijngaarden.

Dit rapport bevat een kort overzicht van het resultaat van onderzoekingen van W.L. Scheen en A. van Wijngaarden over enkele fundamentele kwesties met betrekking tot afrondingsfouten. Een uitvoerig rapport zal binnenkort worden gepubliceerd. Wij danken onze collega's van het Mathematisch Centrum voor hun medewerking op verschillende detailpunten.

1. Definities en nomenclatuur.

Ons gemakshalve beperkende tot de voorstelling van een getal in het decimale stelsel, definieren wij de op n decimalen afgeronde waarde $A_n f$ van een getal f als dat gehele getal maal 10^{-n}, waarvoor geldt $-5 \times 10^{-n-1} < f - A_n f < 5 \times 10^{-n-1}$. In het geval, dat $10^n f = 0{,}5 \pmod 1$, geldt de aanvullende definitie, dat $A_n f$ dat even gehele getal maal 10^{-n} is, waarvoor geldt $f - A_n f = \pm 5 \times 10^{-n-1}$. Deze definitie heeft het voordeel, dat a priori de waarschijnlijkheid om een getal naar boven of naar beneden af te ronden gelijk is, terwijl voorts de kans om door twee maal in successie af te ronden een ander resultaat te bereiken dan door inoens tot het uiteindelijke aantal decimalen af te ronden sterk verminderd wordt. Door nog wat zorgvuldiger definitie van het afrondingsprincipe is deze kans zelfs nog te verkleinen.

In het volgende zullen wij grote reeksen getallen tot hetzelfde aantal decimalen afgerond denken. Gemakshalve denken wij ze eerst met 10^n vermenigvuldigd zodat de afgeronde waarde een geheel getal is. Wij schrijven dan kortweg A voor de afrondingsoperator. De operator 1-A zullen wij α noemen. Voorts zullen wij vaak de afgeronde waarde aangeven door een corresponderende hoofdletter te bezigen, dus $Af = F$ en het resterende gedeelte van f, dus αf, dat wij het breukdeel van f zullen noemen, geven wij aan met een corresponderende Griekse letter, dus $\alpha f = \varphi$. Dus is:

$$f = Af + \alpha f = F + \varphi , \qquad F = 0 \pmod 1 \text{ als } |\varphi| < \tfrac{1}{2}.$$
$$\text{of} \quad F = 0 \pmod 2 \text{ als } |\varphi| = \tfrac{1}{2}. \qquad (1.1)$$

Een stelsel van N breukdelen $\varphi_k (k = 1,2,\ldots N)$ noemen wij homogeen als voor het aantal $M(N)$ der φ_k, waarvoor $-\tfrac{1}{2} \leqslant \varphi_k \leqslant \xi \leqslant \tfrac{1}{2}$, geldt $\lim_{N\to\infty} M(N,\xi)/N = \xi + \tfrac{1}{2}$.

Voorts noemen wij n opeenvolgende φ_k in ditstelsel, indien deze limiet cok nog geldt voor het gevalwij de φ_k, welke bijdragen tot $M(N,\xi)$, slechts dan tellen als $-\tfrac{1}{2} \leqslant a_j \leqslant \varphi_{k-j} \leqslant b_j \leqslant \tfrac{1}{2}$, $(j = 1,2\ldots n-1)$, waarin de a_j's en b_j's willekeurig gekozen zijn.

Hoewel onze problemen eigelijk niets met statistiek te maken hebben, omdat een begrip als waarschijnlijkheid van een bepaalde afronding zinneloos is (ze is immers volkomen bepaald) kunnen wij onder bepaalde omstandigehden met het oog op onze

onkunde genoegen nemen met pseudostatistische uitspraken over "waarschijnlijkheid van een fout". In het volgende zullen wij daarom ook gemakshalve statistische terminologie gebruiken.

2. Eerste standaardvraagstuk.

Zij gegeven een grote verzameling getallen $f_k = F_k + \varphi_k$, waarvoor geldt, dat de φ_k homogeen verdeeld is en voorts dat n opvolgende φ_k onafhankelijk zijn. Zij gevraagd te bepalen

$$f = \sum_{k=1}^{n} a_k f_k, \tag{2.1}$$

waarin de a_k's gegeven constanten zijn.

Als ons alleen de afgeronde waarden F_k ter beschikking staan is het beste war wij kunnen doen te vormen

$$g = \sum_{k=1}^{n} a_k F_k \tag{2.2}$$

Tussen beide antwoorden bestaat een discrepantie

$$\psi = f - g = \sum_{k=1}^{n} a_k \varphi_k = \sum_{k=1}^{n} \psi_k . \tag{2.3}$$

Wat is de waarschijnlijkheidsdichtheid $w(\psi)$ van ψ?

De waarschijnlijkheidsdichtheid $w_k'(\varphi_k)$ van φ_k is gegeven door:

$$w_k'(\varphi_k) = \begin{cases} 1 & \text{als } |\varphi_k| < \frac{1}{2} \\ 0 & \text{als } |\varphi_k| > \frac{1}{2} \end{cases} \tag{2.4}$$

en dus is die van ψ_k, nl. $w_k(\psi_k)$ gelijk aan

$$w_k(\psi_k) = \begin{cases} 1/|a_k| & \text{als } |\psi_k| < |a_k|/2 \\ 0 & \text{als } |\psi_k| > |a_k|/2 \end{cases} \tag{2.5}$$

Dus is

$$w(\psi) = \int_{-\infty}^{\infty} w_1(\psi_1)d\psi_1 \int_{-\infty}^{\infty} w_2(\psi_2)d\psi_2 \cdots \int_{-\infty}^{\infty} w_{n-1}(\psi_{n-1}) \cdot w_n\Big(\psi - \sum_{k=1}^{n-1} \psi_k\Big)d\psi_{n-1}$$

Wij vormen de tweezijdig Laplacegetransformeerde $L_{II}\, w(\psi)$ van $w(\psi)$:

$$L_{II}\, w(\psi) = \int_{-\infty}^{\infty} e^{-p\psi} w(\psi)d\psi =$$

$$= \int_{-\infty}^{\infty} e^{-p\psi} d\psi \int_{-\infty}^{\infty} w_1(\psi_1)d\psi_1 \int_{-\infty}^{\infty} w_2(\psi_2)d\psi_2 \cdots \int_{-\infty}^{\infty} w_{n-1}(\psi_{n-1}) \cdot w_n\Big(\psi - \sum_{k=1}^{n-1} \psi_k\Big)d\psi_{n-1}$$

$$= \int_{-\infty}^{\infty} e^{-p\psi_1} w_1(\psi_1)d\psi_1 \int_{-\infty}^{\infty} e^{-p\psi_2} w_2(\psi_2)d\psi_2 \cdots \int_{-\infty}^{\infty} e^{-p\psi_n} w_n(\psi_n)d\psi_n .$$

Dus is

$$L_{II}\ w(\psi)= \prod_{k=1}^{n} L_{II}\ w_k(\psi_k) \qquad (2.6)$$

Uit (2.5) volgt

$$L_{II}\ w_k(\psi_k)= \int_{-|a_k|/2}^{|a_k|/2} \frac{1}{|a_k|}\, e^{-p\psi_k}\, d\psi_k = \frac{e^{pa_k/2} - e^{-pa_k/2}}{pa_k}. \qquad (2.7)$$

Om hieruit $w(\psi)$ te vinden kan op twee wijzen geschieden, nl. door directe inversie van de Laplacetransformatie of door $w(\psi)$ te ontwikkelen naar Hermite-functies, waarbij de ontwikkelingscoefficienten dan juist gemakkelijk met de Laplace-getransformeerde bepaald kunnen worden. Hier zullen wij alleen de eerste methode bespreken.

Zijn de 2^n grootheden λ_j gedefinieerd door

$$\lambda_j = \sum_{k=1}^{n} \pm a_k/2, \qquad (2.8)$$

dan is blijkbaar

$$L_{II}\ w(\psi)= \frac{1}{p^n \prod_{k=1}^{n} a_k} \sum_{\lambda_j} \pm e^{p\lambda_j} \qquad (2.9)$$

waarin het plusteken dan wel het minteken gekozen dient te worden naar gelang bij de bepaling van de betreffende λ_j volgens (2.8) een even dan wel een oneven aantal mintekens is gebruikt.

De inversie is eenvoudig en levert

$$w(\psi)= \frac{1}{(n-1)! \prod_{k} a_k} \sum_{\lambda_j} \pm \left[\psi+\lambda_j\right]^{n-1}, \qquad (2.10)$$

waarin het symbool tussen vierkante haken gedefinieerd is door

$$\left[x\right]^n = \left(\frac{x+|x|}{2}\right)^n = \begin{cases} x^n & \text{als } x>0 \\ 0 & \text{als } x<0. \end{cases} \qquad (2.11)$$

Hieruit volgen de cumulatieve verdelingsfunctie $v(\psi)$:

$$v(\psi)= \int_{-\infty}^{\psi} w(t)dt = \frac{1}{n! \prod_{k} a_k} \sum_{\lambda_j} \pm \left[\psi+\lambda_j\right]^n \qquad (2.12)$$

en haar eerste integraal $u(\psi)$:

$$u(\psi)= \int_{-\infty}^{\psi} v(t)dt = \frac{1}{(n+1)! \prod_{k} a_k} \sum_{\lambda_j} \pm \left[\psi+\lambda_j\right]^{n+1}. \qquad (2.13)$$

In deze formules mogen de a_k's ook vervangen worden door hun absolute waarden. Blijkbaar hoeft de som over λ_j slechts uitgestrekt te worden over die λ_j, waarvoor geldt $\psi+\lambda_j>0$.

3. Tweede standaardvraagstuk.

Dit is een wijziging van het voorgaande vraagstuk. Laat ons niet f doch slechts F willen bepalen. In plaats hiervan kunnen wij slechts G bepalen. De discrepantie P

$$P = F - G \tag{3.1}$$

is nu een geheel getal. Gevraagd wordt de waarschijnlijkheid $\Omega(P)$ te bepalen dat P een voorgeschreven (gehele) waarde aanneemt. Dit vraagstuk is aanmerkelijk ingewikkelder, hoewel eigenlijk minder gevraagd wordt.

$$P = A\Sigma a_k f_k - A\Sigma a_k F_k = A(\Sigma a_k F_k + \Sigma a_k \varphi_k) - A\Sigma a_k F_k =$$

$$= A(g + \varphi) - G = A(g - G + \psi)$$

m.a.w.

$$P = A(\gamma + \psi) \tag{3.2}$$

Dus geldt

$$P - \tfrac{1}{2} - \psi \leqslant \gamma \leqslant P + \tfrac{1}{2} - \psi \tag{3.3}$$

Laat $V(\gamma)$ de verdelingsfunctie van γ zijn, dus $V(\gamma_0)$ de waarschijnlijkheid, dat $\gamma < \gamma_0$, zodat $V(\gamma) = 0$ als $\gamma < -\tfrac{1}{2}$ en $V(\gamma) = 1$ als $\gamma > \tfrac{1}{2}$. Voor gegeven ψ is dan

$$\Omega(P) = \begin{cases} 0 & \text{als } \psi < P-1 \text{ en als } \psi > P+1 \\ 1 - V(P - \tfrac{1}{2} - \psi) & \text{als } P-1 < \psi < P \\ V(P + \tfrac{1}{2} - \psi) & \text{als } P < \psi < P+1. \end{cases} \tag{3.4}$$

Dus totaal is

$$\Omega(P) = \int_{P-1}^{P} w(\psi)\left\{1 - V(P-\tfrac{1}{2}-\psi)\right\} d\psi + \int_{P}^{P+1} w(\psi)\, V(P+\tfrac{1}{2}-\psi) d\psi =$$

$$= \int_{-\frac{1}{2}}^{\frac{1}{2}} w(P-\tfrac{1}{2}+\gamma)\left\{1-V(-\gamma)\right\} d\gamma + \int_{-\frac{1}{2}}^{\frac{1}{2}} w(P+\tfrac{1}{2}-\gamma)\, V(\gamma) d\gamma. \tag{3.5}$$

In het belangrijke geval, dat $V(\gamma)$ voldoet aan de symmetrierelatie $V(\gamma) = 1 - V(-\gamma)$ is dus:

$$\Omega(P) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \left\{w(P-\tfrac{1}{2}+\gamma) + w(P+\tfrac{1}{2}-\gamma)\right\} V(\gamma) d\gamma. \tag{3.6}$$

Om $V(\gamma)$ te bepalen, bewijzen wij eerst twee hulpstellingen:

Hulpstelling 1. Zij gegeven een stelsel gehele getallen p_k, met $GGD(p_k) = 1$. Dan heeft de vergelijking $\Sigma F_k p_k = 1$ een oplossing met gehele F_k.

Bewijs: Zij d de kleinste positieve waarde, welke $\Sigma F_k p_k$ kan aannemen. Zij nu $p_k = q_k\, d + r_k$ met gehele q_k en $0 \leqslant r_k \lll d$, dan is $r_k = p_k - q_k\, d$ een getal van de vorm $\Sigma F_k p_k$, en omdat $r_k < d$, geldt $r_k = 0$. Dus $d \leqslant GGD(p_k)$, dus $d = 1$.

Hulpstelling 2. Als de F_k's (k=1,2...n) alle gehele waarden van $-\infty$ tot ∞ met gelijke waarschijnlijkheid aannemen, dat doet $F_k p_k$ dit ook, mits $GGD(p_k) = 1$.

Bewijs: Dit is bewezen, als wij hebben aangetoond, dat het rooster in de n-dimensionale ruimte gevormd door de punten $(F_1, F_2, \ldots F_n)$ in disjuncte configuraties verdeeld kan worden, zodanig, dat

a) in iedere configuratie $\Sigma F_k p_k$ elke gehele waarde éénmaal aanneemt;

b) alle configuraties congruent zijn;

c) het gehele rooster uitgeput is.

Eerst vormen wij zulk een configuratie door een punt $(F_1, F_2, \ldots F_n)$, waarvoor $\Sigma F_k p_k = 1$, te vermenigvuldigen met alle gehele getallen. De andere configuraties worden gevonden door alle translaties welke de oorsprong overvoeren in een punt, waarvoor ook geldt $\Sigma F_k p_k = 0$. Dan is aan a) en b) voldaan. Ook aan c) is voldaan, omdat als $\Sigma F_k p_k = q$ en $\Sigma F'_k p_k = q$ ook $\Sigma (F'_k - F_k) p_k = 0$.

Wij onderstellen nu alle a_k's rationaal en herleid tot breuken met kleinst mogelijke gelijke noemer q. Zij dan $a_k = t_k/q$ en GGD $t_k = r$, dus $t_k = r\, p_k$ met GGD $p_k = 1$. Dan is $g = \Sigma a_k F_k = (r/q) \Sigma p_k F_k = (r/q)K$, waarin $K = \Sigma p_k F_k$ alle gehele waarden met gelijke waarschijnlijkheid aanneemt als F_k dit doet, terwijl $GGD(r,q) = 1$.

Is q oneven en neemt K de waarden $0, 1, \ldots q-1$ aan, dan neemt γ de waarden $-\frac{1}{2} + \frac{1}{2q}, -\frac{1}{2} + \frac{3}{2q}, \ldots \frac{1}{2} - \frac{1}{2q}$ elk één maal aan, terwijl γ voorts periodiek in K is met periode q. Dan is dus:

$$V(\gamma) = \begin{cases} 0 & \text{als } \gamma < -\frac{1}{2} \\ k/q & \text{als } -\frac{1}{2} \leqslant -\frac{1}{2} + \frac{1}{2q} + \frac{k-1}{q} \leqslant \gamma < -\frac{1}{2} + \frac{1}{2q} + \frac{k}{q} \leqslant \frac{1}{2} \\ 1 & \text{als } \gamma > \frac{1}{2}. \end{cases} \qquad (3.7)$$

Is q even en wordt een symmetrische afrondregel voor $\frac{1}{2}$ en $-\frac{1}{2}$ gebruikt als beschreven in paragraaf 1, dan geldt (3.7) ook.

Maken wij gebruik van het periodiek voortgezette polynoom van Bernoulli $\bar{B}_1(x)$, dan is blijkbaar

$$V(\gamma) = \begin{cases} 0 & \text{als } \gamma \leqslant -\frac{1}{2} \\ -\frac{1}{2} + \gamma - \frac{1}{q}\bar{B}_1(q\gamma) & \text{als } -\frac{1}{2} \leqslant \gamma \leqslant \frac{1}{2} \\ 1 & \text{als } \gamma > \frac{1}{2}. \end{cases} \qquad (3.8)$$

Het is nu slechts een kwestie van rekenen om $\Omega(P)$ te vinden. Wij kunnen het antwoord nog in verschillende vormen krijgen. Onder gebruikmaking van de centrale differentieoperatoren δ en δ^2 gedefinieerd door

$$\delta f(x) = f(x + \tfrac{1}{2}) - f(x - \tfrac{1}{2})$$
$$\delta^2 f(x) = f(x + 1) - 2 f(x) + f(x-1),$$

en de functies v en u uit (2.12) en (2.13) en realiserende, dat δf alle $q\lambda_j = 0 \pmod 1$ of alle $q\lambda_j = \frac{1}{2} \pmod 1$, vinden wij tenslotte:

1) $q = 0 \pmod 2$

a) $q\lambda_j = 0 \pmod 1$ en $n = 2m$ of $n = 2m+1$, of $q\lambda_j = \frac{1}{2} \pmod 1$ en $n = 2m$:

$$\Omega(p) = \delta^2 \sum_{k=0}^{m} \frac{B_{2k}}{(2k)!q^{2k}} u^{(2k)}(P) \qquad (3.9)$$

b) $q\lambda_j = \frac{1}{2} \pmod 1$ en $n = 2m+1$:

$$\Omega(P) = \delta^2 \left\{ \sum_{k=0}^{m} \frac{B_{2k}}{(2k)!q^{2k}} u^{(2k)}(P) - \frac{B_{2m+2}(\frac{1}{2})}{(2m+2)!q^{2m+2}} u^{(2m+2)}(P) \right\} \qquad (3.10)$$

2) $q = 1 \pmod 2$

a) $q\lambda_j = 0 \pmod 1$ en $n=2m$ of $n= 2m+1$, of $q\lambda_j = \frac{1}{2} \pmod 1$ en $n = 2m$:

$$\Omega(P) = \delta^2 \sum_{k=0}^{m} \frac{B_{2k}(\frac{1}{2})}{(2k)!q^{2k}} u^{(2k)}(P) \qquad (3.11)$$

b) $q\lambda_j = \frac{1}{2} \pmod 1$ en $n= 2m+1$

$$\Omega(P) = \delta^2 \left\{ \sum_{k=0}^{m} \frac{B_{2k}(\frac{1}{2})}{(2k)!q^{2k}} u^{(2k)}(P) - \frac{B_{2m+2}}{(2m+2)!q^{2m+2}} u^{(2m+2)}(P) \right\} \qquad (3.12)$$

De numerieke waarden van optredende B_{2k} en $B_{2k}(\frac{1}{2})$ zijn:

$$\frac{B_2}{2!} = \frac{1}{12} \qquad \frac{B_4}{4!} = -\frac{1}{720} \qquad \frac{B_6}{6!} = \frac{1}{30240} \quad \ldots$$

$$\frac{B_2(\frac{1}{2})}{2!} = -\frac{1}{24} \qquad \frac{B_4(\frac{1}{2})}{4!} = \frac{7}{5760} \qquad \frac{B_6(\frac{1}{2})}{6!} = -\frac{31}{967680} \quad \ldots$$

Is minstens een van de a_k's irrationaal, dan is $q = \infty$ en geldt eenvoudig:

$$\Omega(P) = \delta^2 u(P). \qquad (3.13)$$

Een andere vorm van het resultaat is:

1) $q = 0 \pmod 2$

$$\Omega(P) = \frac{1}{q} \delta^2 \left\{ \sum_{k=-\infty}^{qP} v\left(\frac{k}{q}\right) - \tfrac{1}{2} v(P) \right\} \qquad (3.14)$$

2) $q = 1 \pmod 2$

$$\Omega(P) = \frac{1}{q} \delta^2 \sum_{k=-\infty}^{qP} v\left(\frac{k-\frac{1}{2}}{q}\right) \qquad (3.15)$$

Hieruit volgt juist voor het andere extreme geval, nl. dat alle a_k's geheel zijn, d.w.z. $q = 1$ is, een eenvoudige formule:

$$\Omega(P) = \delta v(P), \qquad (3.16)$$

welke trouwens ook direct is in te zien.

4. Afhankelijkheid van de breukdelen.

Tot dusverre hebben wij steeds onafhankelijke φ_k's beschouwd. Van groot belang is echter het geval, dat er nevencondities aan de φ_k's zijn opgelegd in de vorm van lineair onafhankelijke functionele relaties

$$\chi_j = \sum_{k=1}^{n} b_{j,k} \varphi_k, \qquad (4.1)$$

waarin de χ_j's gegeven constanten zijn. De bovengenoemde methodes falen dan en wij roepen een meer elementaire methode te hulp, welke wij voortaan kortweg de meetkundige zullen noemen, hoewel zij zuiver analytisch doorgevoerd kan worden. Construeer een n-dimensionale ruimte R_n met coordinaten φ_k. De grenzen $|\varphi_k| = \frac{1}{2}$ definieren een eenheidskubus om de oorsprong. Als de φ_k's homogeen verdeeld en onafhankelijk zijn is de waarschijnlijkheidsdichtheid om het punt $\Phi(\varphi_1, \varphi_2 \ldots \varphi_n)$ ergens binnen de kubus aan te treffen constant = 1. De vergelijking (2.3) definieert een hypervlak R_{n-1} in R_n en $v(\psi)$ is eenvoudig het volume ingesloten door R_{n-1} en de vlakken $|\varphi_k| = -\frac{1}{2}$. De λ_j's uit (2.8), welke zo'n belangrijke rol speelden in de oplossing van de standaardproblemen corresponderen met de waarden van ψ, waarvoor R_{n-1} een hoekpunt van de kubus bevat.

Is een aantal m voorwaarden (4.1) gegeven, dan kan het punt zich slechts bewegen over de binnen de kubus gelegen gemeenschappelijke deelruimte R_{n-m} der R_{n-1}'s voorgesteld door (4.1). Een gegeven ψ kan dusslechts worden gerealiseerd op de aan R_{n-m} en de R_{n-1} volgens (2.3) gemeenschappelijke deelruimte R_{n-m-1}.

Is overigens aan de homogene verdeling van de φ_k voldaan, dan vinden wij $v(\psi)$ als het volume van R_{n-m} begrensd door R_{n-m-1} en de vlakken $\varphi_k = -\frac{1}{2}$ gedeeld door het totale volume van R_{n-m} begrensd door de vlakken $|\varphi_k| = \frac{1}{2}$.

In vele gevallen kan op het aantal dimensies wat bezuinigd worden door geschikte projectie, doch de technische moeilijkheden der berekening zijn bij meer dan drie dimensies uit de aard der zaak niet gering. De methode verschaft evenwel naast een antwoord bovendien inzicht, wat van de hiervoor behandelde methoden niet gezegd kan worden.

5. Homogene verdeling en afhankelijkheid in een tabel van een functie.

Wij willen de hierboven geschetste methoden toepassen op vraagstukken die rijzen bij lineaire operaties verricht op een tafel van een functie f met gelijkmatig opklimmend argument $x_k = x_o + kh$. Twee vragen doen zich direct voor:

1) Zijn de φ_k's homogeen verdeeld tussen $-\frac{1}{2}$ en $\frac{1}{2}$?
2) Zijn n φ_k's al dan niet afhankelijk?

De eerste vraag is in wezen die van de gelijkverdeling modulo 1 van de functie en als zodanig i.h.a. niet te beantwoorden. Echter interesseren wij ons eigenlijk helemaal niet voor oneindige tafels, doch behoeven slechts te weten of de gelijkverdeling ongeveer optreedt over een groot aantal functiewaarden. Voorts tabelleert men gewoonlijk niet functies, welke practisch niet veranderen, d.w.z. in zulke gevallen vergroot men het interval h op passende wijze. Daarom ligt de zaak toch anders dan in de getallentheorie en blijkt aan de homogene verdeling meestal zeer goed voldaan te zijn.

De tweede vraag is belangrijker. Voor de n-de differentie $\Delta_1^n f$ van een n-maal continu differentieerbare reeele functie $f(x)$ geldt $\Delta_1^n f = h^n f^{(n)}(\xi)$ met $x_1 \leqslant \xi \leqslant x_{n+1}$. Als $f^{(n)}(\xi)$ begrensd is over het gebied van de tafel kan $\Delta^n f$ willekeurig klein gemaakt worden en vrijwel iedere functietafel, welke in de praktijk gebruikt wordt is met zo'n klein interval gemaakt, dat een bepaalde differentie verwaarloosbaar klein is. Nu is

$$\Delta^n \varphi = \Delta^n f - \Delta^n F = \sum_{k=1}^{n} (-1)^{n-k+1} \binom{n}{k-1} \varphi_k + \varphi_{n+1} =$$

$$\sum_{k=1}^{n} a_k \varphi_k + \varphi_{n+1} . \qquad (5.1)$$

dus

$$\varphi_{n+1} = \alpha(\Delta^n f - \sum_{k=1}^{n} a_k \varphi_k) \qquad (5.2)$$

Zijn de $\varphi_1,\ldots,\varphi_n$ onafhankelijk en is $\alpha\Delta^n f$ homogeen verdeeld en onafhankelijk van $\Sigma a_k\varphi_k$, dan is φ_{n+1} onafhankelijk van $\varphi_1 \ldots \varphi_n$. Is echter over het gehele grote gebied, waarover wij een uitspraak wensen te doen $\alpha\Delta^n f \approx$ constant, dan is φ_{n+1} afhankelijk van $\varphi_1 \cdots \varphi_n$.

De laagste differentie welks breukdeel $\approx$ constant is noemen wij de critische differentie. Differenties van lagere orde noemen wij subcritisch, die van hogere orde supercritisch. Opgemerkt dient te worden, dat wanneer $\alpha\Delta^n f$ exact constant is (zodat $\alpha\Delta^{n+p} = 0$ is), slechts voor irrationale $\alpha\Delta^n f$ geldt, dat de φ_n's homogeen verdeeld zijn. Polynomen met rationale coefficienten en rationale h voldoen dus niet aan de gelijkverdeling van de φ_k's en zullen worden uitgesloten, hoewel de resultaten, welke wij zullen afleiden, wel goed <u>kunnen</u> zijn.

In het volgende zullen wij onze tafels de volgende eisen opleggen:

1) De φ_k's zijn homogeen verdeeld.
2) Tot zekere n geldt, dat $\varphi_1 \ldots \varphi_n$ onafhankelijk zijn, terwijl van hogere n afhankelijkheid optreedt en wel direct in de functionele zin (4.1).

De vraag of een bepaalde tafel hieraan voldoet laten wij over aan degeen die de resultaten wil toepassen.

6. <u>De verdeling van de differenties in een tafel</u>.

De voorgaande overwegingen leiden al direct tot een practisch probleem, nl. dat van de bepaling van de waarschijnlijkheid $\Omega(P)$ van een voorgeschreven discrepantie P

$$P = A\,\Delta^n f - \Delta^n F. \qquad (6.1)$$

Dit vraagstuk is daarom van groot belang, omdat men de differenties $\Delta^n F$, waarvoor $\Delta^n f \approx 0$ is gebruikt als controle op de juistheid van de getabelleerde F-waarden. De extreme waarden die P dan kan aannemen zijn $\pm 2^{n-1}$, doch de waarschijnlijkheid van zulke grote en zelfs beduidend kleinere waarden is zo gering dat men met recht de F-waarden kan wantrouwen ook al is P een weinig binnen deze extreme grenzen. Men moet dus voor verschillende n practische grenzen aangeven. Comrie heeft dergelijke grenzen aangegeven gebaseerd op practische ervaring en Miller berekende, dat deze grenzen van Comrie juist die waren, waarvoor de verwachtingswaarde van een grotere P kleiner dan 1 % is. Wij zullen aantonen, dat de zaak ingewikkelder is en dat een dergelijke uitspraak alleen onder bepaalde veronderstellin-

gen over $\propto\Delta^n f$ gedaan kan worden.

Is $\Delta^n f$ supercritisch, dan zijn $\varphi_1 \ldots \varphi_{n+1}$ onafhankelijk en de oplossing van het tweede standaardvraagstuk levert ons het gezochte resultaat. Als a_k fungeren de binomiaalcoefficienten $(-1)^{n-k+1}\binom{n}{k-1}$ voor $k = 1,2\ldots n+1$. Omdat de a_k's geheel zijn kan de eenvoudige formule (3.16) gebruikt worden.

Is $\Delta^n f$ critisch, dan zijn $\varphi_1 \ldots \varphi_n$ onafhankelijk maar φ_{n+1} is er afhankelijk van. Uit (5.2) volgt dan

$$\Omega(P) = \delta v(P + \propto\Delta^n f), \qquad (6.2)$$

waarbij weer $a_k = (-1)^{n-k+1}\binom{n}{k-1}$, doch nu voor $k = 1,2\ldots n$. Het resultaat hangt nu af van $\propto\Delta^n f$. Voor $\propto\Delta^n f = 0$ vinden wij Miller's resultaten.

Is $\Delta^n f$ subcritisch, dan zijn $\varphi_{n+1} \ldots \varphi_{n-r}$ afhankelijk van de onafhankelijke $\varphi_1 \ldots \varphi_{n-r-1}$. De oplossing kan nu worden gevonden met de meetkundige methode.

Op deze wijze hebben wij $\Omega(P)$ onder vele omstandigheden uitgerekend. Als voorbeeld van de resultaten geven wij $\Omega(P)$ voor de derde differentie op pag. 12 . Men ziet wel, dat de waarschijnlijkheidsverdeling zeer sterk varieert met de omstandigheden.

7. De kans op afrondingsfouten bij interpolatie.

Een andere toepassing van de theorie is gelegen in de interpolatie van een functie. Hierbij wordt een functiewaarde f_p behorende bij een argument $x_o + p(x_1-x_o)$ gevormd als een lineair compositum van een aantal functiewaarden. Zijn de functiewaarden welke hierbij gebruikt w orden modulo 1 onafhankelijk, dan kan zonder meer de foutenkans berekend worden met de oplossing van het standaardprobleem 1. Van meer belang is echter het geval, dat wij het geinterpoleerde resultaat afronden tot het originele aantal decimalen en dit antwoord vergelijken met dat, wat verkregen zou zijn door de interpolatie te verrichten met behulp van de niet afgeronde functiewaarden en pas daarna af te ronden. Standaardprobleem 2 levert nu de oplossing. Daarbij komt het merkwaardige resultaat te voorschijn, dat de kans om een fout van een bepaald aantal eenheden te maken een discontinue functie van p is en wel zodat voor iedere rationale waarde van p een discontinuiteit optreedt Dit effect is overigens slechts van belang als n klein of p "erg rationaal" is. Als voorbeeld volgen hier de kansen

Waarschijnlijkheid $\Omega(P)$ voor de derde differentie.

P	$\Delta^3 f$ Supercritisch $\propto \Delta f \approx$						$\Delta^3 f$ Supercritisch $\propto \Delta^2 f \approx$						$\Delta^3 f$ Critisch $\propto \Delta^3 f \approx$						$\Delta^3 f$ Subcritisch	
	0	.1	.2	.3	.4	.5	0	.1	.2	.3	.4	.5	0	.1	.2	.3	.4	.5		
<-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	>4
-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.0003	3
-3	0	0	0	0	0	0	0	.0025	.0100	.0225	.0367	.0417	.0185	.0135	.0095	.0064	.0040	.0023	.0226	3
-2	0	.1	.2	.3	.4	.5	.1667	.1592	.1367	.0992	.0567	.0417	.1111	.1000	.0890	.0783	.0679	.0579	.1114	2
-1	0	0	0	0	0	0	.1667	.1742	.1967	.2342	.2767	.2917	.2222	.2111	.2000	.1889	.1778	.1667	.2216	1
0	1	.7	.4	.1	0	0	.3333	.3283	.3133	.2883	.2600	.2500	.2963	.2952	.2921	.2873	.2809	.2731	.2882	0
1	0	.2	.4	.6	.4	0	.1667	.1742	.1967	.2342	.2767	.2917	.2222	.2333	.2441	.2546	.2643	.2731	.2216	-1
2	0	0	0	0	.2	.5	.1667	.1592	.1367	.0992	.0567	.0417	.1111	.1222	.1333	.1444	.1556	.1667	.1114	-2
3	0	0	0	0	0	0	0	.0025	.0100	.0225	.0367	.0417	.0185	.0246	.0317	.0397	.0484	.0579	.0226	-3
4	0	0	0	0	0	0	0	0	0	0	0	0	0	.0000	.0001	.0005	.0012	.0023	.0003	-4
>4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	<-4
	0	-.1	-.2	-.3	-.4	-.5	0	-.1	-.2	-.3	-.4	-.5	0	-.1	-.2	-.3	-.4	-.5	$\Delta^3 f$ Subcritisch	P
	$\propto \Delta f \approx$						$\propto \Delta^2 f \approx$						$\propto \Delta^3 f \approx$							
	$\Delta^3 f$ Supercritisch												$\Delta^3 f$ Critisch							

op een fout +1 of -1 samen voor lineaire, drie- en vierpunts centrale Lagrange-interpolatie en wel is de eerste kolom de waarschijnlijkheid $\Omega(1)+\Omega(-1)$ gegeven voor de exacte waarde van p, terwijl er naast de kans is gegeven voor een p-waarde, welke er willekeurig weinig van verschilt.

p	Lineaire interpolatie		Driepuntsinterpolatie		Vierpuntsinterpolatie	
0	0.0000	0.2500	0.0000	0.2500	0.0000	0.2500
0.1	0.2278	0.2259	0.2479	0.2479	0.2358	0.2358
0.2	0.2000	0.2021	0.2417	0.2418	0.2208	0.2208
0.3	0.1881	0.1857	0.2320	0.2320	0.2058	0.2058
0.4	0.1667	0.1722	0.2190	0.2192	0.1938	0.1938
0.5	0.2500	0.1677	0.2083	0.2049	0.1897	0.1886

Vooral bij p =0 is het rationaliteitseffect zo groot, dat een eenvoudig experiment het gemakkelijk aantoont.

Ook nu kunnen wij weer het geval onderzoeken, dat de functiewaarden, welke bij de interpolatie gebruikt worden functioneel gecorreleerd zijn. Nog zonder veel moeite is dan het geval van lineaire interpolatie in een tabel van een lineaire functie te behandelen. Het blijkt dan, dat het breukdeel γ van de uit de tabel geinterpoleerde g nog slechts 2 waarden kan aannemen, (tenzij bij rationale p, waarbij het kan voorkomen, dat één van beide waarden $\frac{1}{2}$ wordt, wat dan verdeeld dient te worden over $\frac{1}{2}$ en $-\frac{1}{2}$), en nog wel met verschillende waarschijnlijkheid. Afgezien van het juist hierboven genoemde effect heeft rationaliteit of irrationaliteit van p nu geen invloed meer! Daarentegen speelt nu een rol of het breukdeel van de constante eerste differentie (dus $\alpha\Delta f$) al of niet rationaal is. Ook speelt de afgeronde waarde van de differentie (dus $A\Delta f$) nu een belangrijke rol. De kans op discrepanties 1 of -1 varieert nu sterk met de omstandigheden, maar als $\alpha\Delta f$ irrationaal is geldt $\Omega(1)+\Omega(-1)<\frac{1}{2}$, waarbij hetzij $\Omega(1)$ of $\Omega(-1)$ willekeurig dicht bij $\frac{1}{2}$ kan komen. Is $\alpha\Delta f$ rationaal, dan vervalt ook deze begrenzing. Uit de algemene theorie leidt men gemakkelijk extreem gunstige of ongunstige omstandigheden af. Is bijv. f = 40 x afgerond op eenheden getabelleerd voor gehele x, dan is iedere geinterpoleerde juist. Is daarentegen f = 40 x + 0.4 en p = 0.01, dan is het resultaat altijd fout!