stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK S 7002-1c APRIL

A.P.B.M. VEHMEYER

HET QUOTIENT VAN DE MEDIAAN EN DE SPREIDINGSBREEDTE VAN NORMAALVERDEELDE STEEKPROEVEN

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

No scientific investigation is final; it merely represents the most probable conclusion which can be drawn from the data at the disposal of the writer. A wider range of facts, or more refined analysis, experiment, and observation will lead to new formula's and new theories. This is the essence of scientific progress.

K. Pearson 1898

1. INLEIDING

In het navolgende wordt verslag gedaan van een, voornamelijk op numerieke resultaten gericht, onderzoek van de stochastische variabele, gevormd door het quotient van de mediaan en de spreidingsbreedte van steekproeven van omvang 3, 5, 7 of 9 uit de standaardnormale verdeling. Vanwege de met toenemende steekproefomvang snel afnemende doeltreffendheid werden geen grotere waarden van de steekproefomvang beschouwd.
Eerder werden hiervan, tegelijk met een artikel van DE HAAN & RUNNENBURG [5], waarin theoretisch enige gegevens van deze stochastische variabele werden afgeleid, voorlopige resultaten gepubliceerd (BOUMA & VEHMEYER [1]). De percentielen van de verdelingsfunctie en de eerste vier momenten, voor zover deze bestaan, worden gegeven. Daarnaast wordt voor de verdeling van deze stochastische variabele en voor de bijpassende niet-centrale verdeling een benadering voorgesteld. Bovendien worden enige mogelijkheden voor toepassingen besproken.

Wanneer de volgende notatie wordt ingevoerd:

n	een oneven natuurlijk getal,
$\underline{u}_{(1)},\ldots,\underline{u}_{(n)}$	een naar opklimmende grootte geordend n-tal onafhankelijke standaard-normaal verdeelde stochastische variabelen,
$\underline{m}_n = \underline{u}_{(n/2+1/2)}$	de mediaan,
$\underline{r}_n = \underline{u}_{(n)} - \underline{u}_{(1)}$	de spreidingsbreedte,

dan wordt de beschouwde stochastische variabele $\underline{h}_n$ gedefinieerd door

$$\underline{h}_n = \frac{\underline{m}_n}{\underline{r}_n} .$$

Evenals bij de Student-grootheid, waarmee $\underline{h}_n$ duidelijk overeenkomst vertoont, kan ook hier een bijpassende niet-centrale verdeling worden gedefinieerd. Algemener wordt daarom gedefinieerd:

$$\underline{h}_n^{\delta} = \frac{\underline{m}_n + \delta}{\underline{r}_n} ,$$

waarbij δ de reële niet-centraliteitsparameter is. De dichtheid en de verdelingsfunctie van $\underline{h}_n^{\delta}$ zullen worden genoteerd door, respectievelijk, f_n^{δ} en F_n^{δ}. De index δ zal worden weggelaten als $\delta = 0$.

2. DE VERDELING VAN $\underline{h}_n$

Voor het verkrijgen van waarden van de verdelingsfunctie $F_n(t)$ van $\underline{h}_n$ werden meerdere methoden gevolgd. Enkele theoretische resultaten konden worden gebruikt. Voor het overige moest gebruik gemaakt worden van methoden, die minder exacte resultaten geven, zoals simulatie en meervoudige numerieke integratie, om de percentielen van $F_n(t)$, zoals die worden gegeven in tabel 1, te verkrijgen.
Na bespreking van de verschillende methoden voor het berekenen van waarden van de verdelingsfunctie zal ook aan de momenten van $\underline{h}_n$ aandacht worden geschonken.

2.1. THEORETISCHE RESULTATEN

DE HAAN & RUNNENBURG [5] leidden formules af voor $F_3(t)$ en $f_5(t)$:

$$F_3(t) = 2 - \frac{3}{2\pi}\left\{\arccos\left(\frac{t-1}{\sqrt{4t^2+2}}\right) + \arccos\left(\frac{t+1}{\sqrt{4t^2+2}}\right)\right\};$$

$$f_5(t) = \frac{15}{\pi^2}\frac{1}{2t^2+1} \times$$

$$\times\Bigg[\frac{1}{\sqrt{3t^2+2}}\left\{2\pi - \arccos\left(\frac{-\frac{1}{2}(3t^2+t)}{\sqrt{(2t^2+t+\frac{1}{2})(3t^2+1)}}\right) - \arccos\left(\frac{-\frac{1}{2}(t^2-2t)}{\sqrt{(2t^2+t+\frac{1}{2})(2t^2+2)}}\right) +\right.$$

$$\left. - \arccos\left(\frac{-\frac{1}{2}(3t^2+2t+2)}{\sqrt{(3t^2+1)(2t^2+2)}}\right)\right\} + \frac{2t}{\sqrt{4t^2+1}}\left\{2\pi - \arccos\left(\frac{-\frac{1}{2}(2t^2+t)}{\sqrt{(2t^2+t+\frac{1}{2})(3t^2+\frac{1}{2})}}\right) +\right.$$

$$\left. - \arccos\left(\frac{-\frac{1}{2}(2t^2-t)}{\sqrt{(2t^2+t+\frac{1}{2})(3t^2+1)}}\right) - \arccos\left(\frac{-\frac{1}{2}(4t^2+t+1)}{\sqrt{(3t^2+\frac{1}{2})(3t^2+1)}}\right)\right\} +$$

$$- \frac{2t}{\sqrt{4t^2+1}}\left\{2\pi - \arccos\left(\frac{-\frac{1}{2}(2t^2-t)}{\sqrt{(3t^2+\frac{1}{2})(2t^2-t+\frac{1}{2})}}\right) - \arccos\left(\frac{-\frac{1}{2}(4t^2-t+1)}{\sqrt{(3t^2+1)(3t^2+\frac{1}{2})}}\right) +\right.$$

$$- \arccos\left(\frac{-\frac{1}{2}(2t+t)}{\sqrt{(3t^2+1)(2t^2-t+\frac{1}{2})}}\right)\} + \frac{1}{\sqrt{3t^2+2}}\{2\pi - \arccos\left(\frac{-\frac{1}{2}(3t^2-2t+2)}{\sqrt{(2t^2+t)(3t^2+1)}}\right) +$$

$$- \arccos\left(\frac{-\frac{1}{2}(t^2+2t)}{\sqrt{(2t^2+2)(2t^2-t+\frac{1}{2})}}\right) - \arccos\left(\frac{-\frac{1}{2}(3t^2-t)}{\sqrt{(3t^2+1)(2t^2-t+\frac{1}{2})}}\right)\}\ ].$$

Met behulp van deze formules kunnen de percentielen van $\underline{h}_3$ direct, en van $\underline{h}_5$ via enkelvoudige numerieke integratie, worden bepaald (zie tabel 1).
Voor andere waarden van de steekproefomvang zijn geen theoretische resultaten bekend. Voor steekproefomvang 7 en 9 moest dus voor het verkrijgen van de waarden van de verdelingsfunctie gebruik gemaakt worden van andere methoden, zoals simulatie en meervoudige numerieke integratie. Een belangrijk nadeel van deze methoden is echter dat de nauwkeurigheid van de via deze methoden verkregen resultaten moeilijk te beoordelen is. Door nu deze methoden ook te gebruiken voor steekproefomvang 3 en 5, en deze resultaten te vergelijken met de via bovenstaande formules verkregen resultaten, kon een inzicht worden verkregen in de behaalde nauwkeurigheid.

2.2. SIMULATIE

Voor de vier beschouwde waarden van de steekproefomvang n werden een groot aantal waarnemingen (tenminste 30.000) van $\underline{h}_n$ gegenereerd. Deze werden in 10.000 klassen ingedeeld en de cumulatieve frequentieverdeling werd berekend. Vanuit de zo verkregen schattingen voor waarden van de verdelingsfuncties werden de percentielen verkregen. Bij het gebruik van simulatie schuilt de moeilijkheid echter voornamelijk in het schatten van de nauwkeurigheid van de verkregen waarden. In dit geval waren er twee mogelijkheden om een idee te krijgen van de nauwkeurigheid:

a. Zoals eenvoudig is in te zien, is de verdeling van de stochastische variabele $\underline{h}_n$ symmetrisch. Dientengevolge moet ook iedere redelijke benadering die symmetrie vertonen. Aan deze voorwaarde voldeden de verkregen experimentele verdelingsfuncties met redelijk goede nauwkeurigheid.
b. De op deze manier verkregen resultaten voor n = 3 en n = 5 konden worden vergeleken met de waarden verkregen met behulp van de theoretische re-

sultaten, waarvan de nauwkeurigheid bekend was. De maximale afwijking was hiervoor kleiner dan .01.

Deze twee indicaties doen verwachten dat ook voor de steekproefomvang 7 en 9 de maximale fout voor de simulatieresultaten kleiner dan .01 is.

2.3. DIRECTE NUMERIEKE INTEGRATIE

Zoals al in [1] werd vermeld, is nadat de simulatieresultaten waren verkregen, nog op een andere manier getracht de percentielen van de verdelingsfunctie $F_n(t)$ voor n = 7 en n = 9 te berekenen. $F_n(t)$ kan immers geschreven worden als een drievoudige integraal, zodat de waarden van $F_n(t)$ mogelijk direct door numerieke integratie zouden kunnen worden berekend. Zij

$$g_n^*(x,y,z) = \frac{n!}{(((n-1)/2)!)^2}\{[\Phi(z) - \Phi(y)][\Phi(y) - \Phi(x)]\}^{(n-1)/2}\phi(z)\phi(y)\phi(x),$$

waarin Φ en ϕ respectievelijk de verdelingsfunctie en de dichtheid van de standaard-normale verdeling voorstellen, dan kan de marginale dichtheid van $\underline{u}_{(1)}$, $\underline{m}_n$, $\underline{u}_{(n)}$ geschreven worden als (zie e.g. Sarhan & Greenberg [9])

$$g_n(u_{(1)},m_n,u_{(n)}) = \begin{cases} g_n^*(u_{(1)},m_n,u_{(n)}) & \text{voor } u_{(1)} < m_n < u_{(n)} \\ 0 & \text{elders.} \end{cases}$$

Voor de verdelingsfunctie $F_n(t)$ geldt dan:

$$F_n(t) = P\{\underline{h}_n \leq t\} = P\left\{\frac{\underline{m}_n}{\underline{u}_{(n)} - \underline{u}_{(1)}} \leq t\right\}$$

$$= \int\int\int_{\frac{m_n}{u_{(n)} - u_{(1)}} \leq t} g_n(u_{(1)},m_n,u_{(n)})\, du_{(n)}\, dm_n\, du_{(1)}$$

$$= \int\int\int_{B(t)} g_n^*(x,y,z)\ dz\ dy\ dx$$

waarbij voor de verzameling B(t), waarover wordt geïntegreerd, geldt

$$B(t) = \{(x,y,z) \mid x < y < z,\ y/(z-x) \leq t\}.$$

Teneinde de beschikbare procedures voor numerieke integratie te kunnen gebruiken moet deze meervoudige integraal eerst geschreven worden als een herhaalde integraal. Door B(t) anders te schrijven, bijvoorbeeld als

$$B(t) = \{(x,y,z) \mid -\infty < x < +\infty,\ x < z < +\infty,\ x < y < \max(x,\min(z,(z-x)t))\},$$

kunnen de grenzen voor de herhaalde integraal en de integratievolgorde direct uit de voorwaarden van B(t) worden afgelezen. Afhankelijk van de waarde van t kunnen de ingewikkelde grenzen - er is altijd minstens één grensfunctie die de functie min(,) en/of max(,) bevat - worden vereenvoudigd. In sommige gevallen geeft dit aanleiding tot een splitsing van B(t). Vanwege de continuiteit van de integrand kan men zich bij splitsing beperken tot open verzamelingen, aangezien integratie over de rand toch steeds 0 oplevert. Eerst worden nu voor verschillende waardebereiken van t de vereenvoudigde verzamelingen gegeven in de hierboven aangegeven vorm, zodat de grenzen en de integratievolgorde hieruit kan worden afgelezen. Voor $t < 0$ is nog geen splitsing van B nodig. Wel kunnen de voorwaarden worden vereenvoudigd. Immers voor $t < 0$ geldt

$$B(t) = \{(x,y,z) \mid x < y < z,\ y/(z-x) \leq t\}$$

$$= \{(x,y,z) \mid -\infty < x < \infty,\ x < y < \infty,\ y < z < \infty,\ y < tz - tx\}$$

$$= \{(x,y,z) \mid -\infty < x < \infty,\ x < y < \infty,\ y < z < \infty,\ z < y/t + x\}$$

$$= \{(x,y,z) \mid -\infty < x < \infty,\ x < y < \infty,\ y < z < y/t + x\}.$$

Nu geldt $y < y/t + x \iff y < tx/(t-1)$. Invullen geeft

$$B(t) = \{(x,y,z) \mid -\infty < x < \infty,\ x < y < tx/(t-1),\ y < z < y/t + x\}.$$

Evenzo $x < tx/(t-1) \iff tx - x > tx \iff -x > 0 \iff x < 0$

$$B(t) = \{(x,y,z)\ -\infty < x < 0,\ x < y < tx/(t-1),\ y < z < y/t + x\}.$$

Evenzo geldt voor $t = 0$

$$B(0) = \{(x,y,z) \mid x < y < z,\ y < 0\}.$$

$$= \{(x,y,z) \mid -\infty < x < 0,\ x < y < 0,\ y < z < \infty\}.$$

Integratie hierover zal vanwege de symmetrie natuurlijk $\frac{1}{2}$ opleveren.

Het volgende geval dat onderscheiden moet worden is $0 < t < 1$. Hierbij treedt wel een echte splitsing op. Immers voor $0 < t < 1$ geldt

$$B(t) = \{(x,y,z) \mid x < y < z,\ y/(z-x) < t\} =$$

$$= \{(x,y,z) \mid -\infty < x < \infty,\ x < y < \infty,\ y < z < \infty,\ z > y/t + x\}$$

$$= \{(x,y,z) \mid -\infty < x < \infty,\ x < y < \infty,\ \max(y,y/t+x) < z < \infty\},$$

$$\max(y,y/t+x) = \begin{cases} y & \text{voor } y < tx/(t-1) \\ y/t + x & \text{voor } y > tx/(t-1) \end{cases}$$

Dit levert een splitsing van $B(t)$ in 2 disjuncte verzamelingen $B_1(t)$ en $B_2^*(t)$ met

$$B_1(t) = \{(x,y,z) \mid -\infty < x < \infty,\ x < y < \infty,\ y < z < \infty,\ y < tx/(t-1)\}$$

$$B_2^*(t) = \{(x,y,z) \mid -\infty < x < \infty,\ x < y < \infty,\ y/t + x < z < \infty,\ y > tx/(t-1)\}.$$

Als $x > tx/(t-1)$, ofwel als $x > 0$, is er geen y die aan de voorwaarden van $B_1(t)$ voldoet, zodat de extra voorwaarde $x < 0$ kan worden toegevoegd. Dus geldt

$$B_1(t) = \{(x,y,z) \mid -\infty < x < 0,\ x < y < tx/(t-1),\ y < z < \infty\}.$$

B_2^* kan op dezelfde manier weer gesplitst worden in B_2 en B_3. Voor $0 < t < 1$ is $B(t)$ met verwaarlozing van de randen dus zo te splitsen in $B_1(t)$, $B_2(t)$ en $B_3(t)$ met

$$B_1(t) = \{(x,y,z) \mid -\infty < x < 0,\ x < y < tx/(t-1),\ y < z < \infty\}$$

$$B_2(t) = \{(x,y,z) \mid -\infty < x < 0,\ tx/(t-1) < y < \infty,\ y/t+x < z < \infty\}$$

$$B_3(t) = \{(x,y,z) \mid 0 < x < \infty,\ x < y < \infty,\ y/t+x < z < \infty\},$$

dat de som van de integralen over B_1, B_2, B_3 met integrand $g_n^*(x,y,z)$ gelijk is aan de integraal over de oorspronkelijke B.

Voor $t = 1$ kan B op dezelfde manier gesplitst worden in

$$B_1(1) = \{(x,y,z) \mid 0 < x < \infty,\ x < y < \infty,\ x+y < z < \infty\}$$

$$B_2(1) = \{(x,y,z) \mid -\infty < x < 0,\ x < y < \infty,\ y < z < \infty\},$$

en voor $t > 1$ in

$$B_1(t) = \{(x,y,z) \mid -\infty < x < 0,\ x < y < \infty,\ y < z < \infty\}$$

$$B_2(t) = \{(x,y,z) \mid 0 < x < \infty,\ tx/(t-1) < y < \infty,\ y < z < \infty\}.$$

$$B_3(t) = \{(x,y,z) \mid 0 < x < \infty,\ x < y < tx/(t-1),\ y/t+x < z < \infty\}.$$

De oorspronkelijke integraal voor $F_n(t)$ kan dus nu, afhankelijk van t, geschreven worden als één herhaalde integraal of de som van herhaalde integralen, zó, dat de verzamelingen waarover geïntegreerd moet worden van eenvoudiger structuur zijn.

Vervolgens moeten de oneigenlijke grenzen worden weggewerkt. Dit kan door alle integratievariabelen te transformeren volgens Φ. De integrand wordt

hierdoor eenvoudiger, maar sommige grenzen krijgen een meer gecompliceerde vorm. Zo wordt de grens $tx/(t-1)$ dan $\Phi(t\Phi^{-1}(x')/(t-1))$ en gaat $y/t+x$ over in $\Phi(\Phi^{-1}(y')/t+\Phi^{-1}(x'))$. Het voorkomen van Φ^{-1} in deze grenzen maakt nog een wijziging noodzakelijk. Immers $\Phi^{-1}(0) = -\infty$ en $\Phi^{-1}(1) = +\infty$; er moet dus voor gezorgd worden dat het argument Φ^{-1} niet te dicht in de buurt van 0 of 1 komt. Na deze en enkele andere minder belangrijke wijzigingen konden tenslotte door directe numerieke integratie waarden van de verdelingsfunctie worden verkregen.

Een voordeel van deze methode is dat ze met een kleine wijziging ook gebruikt kan worden voor het verkrijgen van de waarden van de bijpassende niet-centrale verdelingsfunctie.

Opmerkelijk was dat voor negatieve t - dus in het gebied waar de oorspronkelijke integraal niet behoefde te worden gesplitst - voor $F_n(t)$ waarden werden verkregen die sterk afweken van de eerder door simulatie verkregen waarden. Voor positieve t, daarentegen, werden waarden gevonden die, gezien de precisie van simulatieresultaten, zeer goed hiermee overeenkwamen. Daarom werden verder slechts waarden van $F_n(t)$, berekend voor positieve t, gebruikt; vanwege de symmetrie van de verdeling van h_n was dit geen wezenlijke beperking.

Het rekenproces dat de waarden van $F_n(t)$ leverde werd vervolgens uitgebreid zó, dat direct de percentielen werden berekend. De nauwkeurigheid van het zo verkregen rekenproces wordt bepaald door een aantal, van te voren op te geven toleranties en de precisies van de te gebruiken benaderingen voor Φ en Φ^{-1}. Deze toleranties zijn echter niet onafhankelijk. Immers men kan, bijvoorbeeld, weliswaar opgeven dat de buitenste integraal met grote precisie moet worden berekend, maar de uitkomst zal alleen dan de vereiste nauwkeurigheid hebben als de integrand van de buitenste integraal, de twee overige integralen, ook met voldoend grote precisie wordt bepaald. Deze samenhang maakte het moeilijk de in feite gemaakte fout te schatten.

Door vergelijking met de eerder via andere methoden verkregen resultaten en door het effect te beschouwen van het variëren van de opgegeven toleranties en van het vervangen van de gebruikte benaderingen voor Φ en Φ^{-1} door andere, kon toch enig inzicht in de behaalde nauwkeurigheid worden verkregen.

Door verantwoorde verandering van de op te geven toleranties en gebruik

van nauwkeuriger benaderingen voor Φ en Φ^{-1} kan uiteraard iedere nauwkeurigheid worden bereikt. Beperkingen worden slechts gevormd door de rekenprecisie van de rekenautomaat en door de bij grotere precisie excessief snel toenemende rekentijd.

Voor de uiteindelijk gekozen toleranties en benaderingen konden met redelijke zekerheid waarden voor de percentielen met een maximale fout kleiner dan .001 worden verwacht. Deze waarden van de percentielen van $F_n(t)$ voor $n = 7$ en $n = 9$ werden opgenomen in tabel 1.

2.4. DE MOMENTEN

Aangezien $\underline{h}_n$ een symmetrische verdeling heeft zijn de oneven momenten alle gelijk aan nul. Verder toonden DE HAAN & RUNNENBURG [5] aan dat het α-de moment slechts bestaat als $0 < \alpha < n-1$. De bestaande tweede en vierde momenten konden door directe numerieke integratie gemakkelijk worden berekend. De waarden worden gegeven in tabel 2.

3. BENADERING MET BEHULP VAN DE STUDENTVERDELING

In het navolgende staat "$\hat{=}$" voor "heeft dezelfde verdeling als", "$\hat{\approx}$" voor "heeft ongeveer dezelfde verdeling als" en zijn $\underline{u}$, $\underline{t}$, $\underline{\chi}^2_\nu$ stochastische variabelen met respectievelijk de standaardnormale, Student- en chikwadraatverdeling, terwijl ν het aantal vrijheidsgraden voorstelt. Bovendien zal $\underline{t}^{\delta}_{\nu}$ een stochastische variabele zijn met een niet-centrale Studentverdeling met δ als noncentraliteitsparameter en ν weer het aantal vrijheidsgraden.

3.1. BENADERING VOOR $\underline{h}_n$

Benaderingen voor een steekproefgrootheid zijn meestal gebaseerd op zijn asymptotisch gedrag. Ze worden dan beter bij grotere steekproeven. Aangezien in dit geval de steekproefomvang maximaal slechts 9 is, kan niet zonder meer van asymptotische resultaten worden gebruik gemaakt. Voor zowel de mediaan als de spreidingsbreedte bestaan echter benaderingen waarvoor "extra" resultaten bekend zijn.

Het is bekend dat de mediaan asymptotisch normaal verdeeld is (zie e.g. WILKS [12]). CHU [2] was echter in staat door boven- en ondergrenzen te berekenen voor $P\{-x < \underline{m}_n/\sigma(\underline{m}_n) \leq y\}$ met x en y willekeurig, positief, aan te tonen dat de mediaan van normaal verdeelde onafhankelijke stochastische variabelen "snel" naar normaliteit convergeert. De meest hanteerbare relatie die hij voor de grenzen geeft, is van de vorm:

$$a_n[\Phi(y) - \Phi(-x)] \leq P\{-x < \underline{m}_n/\sigma(\underline{m}_n) \leq y\} \leq b_n[\Phi(y) - \Phi(-x)],$$

voor $x \geq 0$, $y \geq 0$,

waarin Φ de verdelingsfunctie is van de standaardnormale verdeling. Uitwerking van a_n en b_n voor $n = 9$ levert al

$$.97[\Phi(y) - \Phi(-x)] \leq P\{-x < \underline{m}_n/\sigma(\underline{m}_n) \leq y\} \leq 1.091[\Phi(y) - \Phi(-x)].$$

Voor de spreidingsbreedte $\underline{r}_n$ van een steekproef uit de standaardnormale verdeling zijn meerdere benaderingen bekend. PATNAIK [7] stelde voor de gemiddelde spreidingsbreedte voor k steekproeven van gelijke omvang n uit de standaardnormale verdeling te benaderen volgens

$$\bar{\underline{r}}_{n,k} \approx c_{n,k}\, \chi_{\nu_{n,k}} / \sqrt{\nu_{n,k}}\,.$$

De constante $c_{n,k}$ en het aantal vrijheidsgraden $\nu_{n,k}$ kunnen zo gekozen worden dat de verwachtingen en varianties van $\bar{\underline{r}}_{n,k}$ en $c_{n,k}\, \chi_{\nu_{n,k}}/\sqrt{\nu_{n,k}}$ gelijk zijn. De waarden van $c_{n,k}$ en $\nu_{n,k}$ werden voor kleine waarden van n en k getabelleerd (zie e.g. PATNAIK [7], DAVID [4], en voor grotere precisie, echter alleen voor k = 1, THOMSON [9]). In PEARSON [8] werd deze benadering voor kleine steekproeven vergeleken met de benadering van COX [3], die de gemiddelde spreidingsbreedte met een chi-kwadraat verdeling benaderde. Pearson vergelijkt ook speciaal het geval dat k = 1. Zijn conclusie is dat voor de spreidingsbreedte de benadering van Patnaik de voorkeur verdient voor een steekproefomvang kleiner dan 10, terwijl Cox' benadering voor steekproefomvang groter dan 10 moet worden geprefereerd.
Verder is bekend dat de mediaan en de spreidingsbreedte ongecorreleerd zijn. De voorafgaande overwegingen doen verwachten:

$$\underline{h}_n = \frac{\underline{m}_n}{\underline{r}_n} \approx \frac{\underline{u}\,\sigma(\underline{m}_n)}{c_n\,\underline{\chi}_{\nu_n}/\sqrt{\nu_n}} = \frac{\sigma(\underline{m}_n)}{c_n}\,\frac{\underline{u}}{\sqrt{\underline{\chi}^2_{\nu_n}/\nu_n}} \approx \gamma_n\,\underline{t}_\nu$$

met $\gamma_n = \sigma(\underline{m}_n)/c_n$ en $c_n = c_{n,1}$.

De variantie van $\underline{m}_n$ is ook getabelleerd (zie e.g. SARHAN & GREENBERG [9]). Deze benadering kan nog verbeterd worden door γ_n en ν_n opnieuw te kiezen. Voor de waarden van n, waarvoor het tweede en het vierde moment van $\underline{h}_n$ bestaan, kunnen γ_n en ν_n zo gekozen worden dat de tweede en vierde momenten van $\underline{h}_n$ en $\gamma_n\,t_{\nu_n}$ gelijk zijn. Voor de kleinere waarden van n, waarvoor de vierde momenten niet bestaan konden door numerieke methoden nieuwe waarden van γ_n en ν_n worden gevonden.

Een benadering met een Studentgrootheid met een niet geheel aantal vrijheidsgraden is in de praktijk niet erg handig, aangezien hiervoor geen handzame tabellen zijn, maar interpolatie noodzakelijk is. Daarom werd ook naar een benadering met behulp van de Studentverdeling met een geheel aantal vrijheidsgraden gezocht. Daartoe werd voor ν_n een geheel getal gekozen dicht bij de waarden die op de hierboven beschreven manier voor ν_n waren verkregen. De constanten γ_n werden zo gekozen dat voor $t = F_n^{-1}(.05)$ geldt

$$P\{\gamma_n\,\underline{t}_{\nu_n} < t\} = .05\ .$$

Dit, omdat in de praktijk vooral goede nauwkeurigheid wordt gevraagd voor de gebruikelijke onbetrouwbaarheden. Als er beslist moest worden tussen twee waarden voor ν_n, werd dan ook op grond van de fouten voor $t = F_n^{-1}(.1)$, $F_n^{-1}(.025)$ en $F_n^{-1}(.01)$ een van beide gekozen.

De aldus verkregen waarden van γ_n en ν_n zowel voor theoretische benadering als voor de verbeterde benadering en ook voor die met gehele ν_n, worden gegeven in tabel 3.

3.2. BENADERING VOOR $\underline{h}_n^\delta$

Waarden van de verdelingsfunctie van $\underline{h}_n^\delta$ kunnen ook worden verkregen door middel van simulatie en door directe numerieke integratie. Voor deze parameter-afhankelijke verdeling is tabellering echter nauwelijks de moeite waard. Volstaan wordt daarom met het geven van een benadering met behulp

van de niet-centrale tudentverdeling.

De al bekende benaderingen voor $\underline{h}_n$ kunnen gemakkelijk worden uitgebreid tot benaderingen voor $\underline{h}_n^{\delta}$. Immers

$$\underline{h}_n^{\delta} = \frac{\underline{m}_n + \delta}{\underline{r}_n} \hat{\approx} \frac{\underline{u}\,\sigma(\underline{m}_n) + \delta}{c_n\, \chi_{\nu_n}/\sqrt{\nu_n}} = \frac{\sigma(\underline{m}_n)}{c_n} \frac{\underline{u} + \delta/\sigma(\underline{m}_n)}{\sqrt{\chi^2_{\nu_n}/\nu_n}} \hat{\approx} \gamma_n\, t_{\nu_n}^{\beta_n \delta}$$

met γ_n en ν_n de constanten uit de theoretische benadering voor $\underline{h}_n$ en $\beta_n = 1/\sigma(\underline{m}_n)$, levert direct een theoretische benadering voor $\underline{h}_n^{\delta}$. Bovendien geldt dat, zowel als door een nieuwe keuze van γ_n, ν_n en β_n getracht wordt de benadering te verbeteren, als wanneer naar een verbeterde benadering van dezelfde vorm maar met gehele ν_n wordt gezocht, de waarden van ν_n en γ_n ontleend kunnen worden aan de al bekende benaderingen voor $\underline{h}_n$, aangezien de gezochte benaderingen voor $\underline{h}_n^{\delta}$ ook moeten gelden voor $\delta = 0$.

Uit een numeriek onderzoek, waarin, gebruik makend van de mogelijkheid $F_n^{\delta}(t)$ door directe numerieke integratie te berekenen, werd getracht verbeterde waarden voor β_n te vinden, bleek dat ook bij de verbeterde benaderingen voor β_n met redelijke resultaten de theoretische waarde $1/\sigma(\underline{m}_n)$ kan worden genomen en dat, mede door het ontbreken van een goed en handzaam criterium geen nieuwe waarden van β_n konden worden gevonden, die voor alle δ tot "betere" resultaten leiden.

In tabel 3 worden met de voor het drietal benaderingen geldende waarde van β_n ook voor ieder van de drie benaderingen de waarden van γ_n en ν_n gegeven.

In tabel 4 wordt voor een aantal waarden van δ en de beschouwde waarden van de steekproefomvang als illustratie gegeven het 95% percentiel van de voorgestelde benadering $t_{.95}$, de door numerieke integratie bepaalde waarde van de verdelingsfunctie $F_n^{\delta}(t_{.95})$ en de waarde die voor β moet worden gekozen om voor $t = t_{.95}$ te doen gelden $F_n^{\delta}(t_{.95}) = P\{\gamma_n\, \underline{t}_{\nu_n}^{\beta_n \delta} < t_{.95}\}$.

4. TOEPASSINGEN

4.1. DOELTREFFENHEID ALS VERVANGER VAN DE STUDENTGROOTHEID

Ter vervanging van de bekende Studentgrootheid

$$\bar{\underline{x}}\,\sqrt{n} \,/\, \underline{s}$$

met $\underline{\bar{x}} = \frac{1}{n} \sum_{n=1}^{n} \underline{x}_i$, het steekproefgemiddelde

en $\underline{s}^2 = \frac{1}{n-1} \sum_{n=1}^{n} (\underline{x}_i - \underline{\bar{x}})^2$, de steekproefvariantie,

werden al enige andere steekproeffuncties voorgesteld, die voor bepaalde waarden van de steekproefomvang slechts weinig minder doeltreffend zijn, eenvoudiger zijn te berekenen en soms ook te gebruiken zijn bij niet geheel bekende steekproef. Voor kleine steekproefomvang wordt bijvoorbeeld in plaats van de steekproefvariantie in de noemer de spreidingsbreedte van de steekproef genomen (zie e.g. PATNAIK [7], MOORE [6]). Vervanging van het zeer eenvoudig te berekenen steekproefgemiddelde vereenvoudigt de berekeningen nauwelijks. Slechts als de steekproef niet geheel bekend is, zal er reden tot vervanging zijn. Over het gebruik van de mediaan hiervoor schrijft DAVID (SARHAN & GREENBERG [9], hoofdstuk 7) dan ook:
"Although the use of midpoint (i.e. $(\underline{x}_{(n)}+\underline{x}_{(1)})/2$) and median is sometimes advocated in complete samples, the rapidly declining efficiency of these statistics with increasing n is a serous limitation, especially since they replace a quantity as simple as the sample mean."
WALSH [11] onderzocht voor gebruik in de plaats van de Studenttoets voor één steekproef een toets met toetsingsgrootheid

$$\frac{(\underline{x}_{(n)} + \underline{x}_{(1)})/2 - \mu}{\underline{r}_n} ,$$

waarin $\underline{x}_{(n)}$, $\underline{x}_{(1)}$, $\underline{r}_n$ respectievelijk het maximum, minimum en de spreidingsbreedte zijn van een n-tal onafhankelijke normaal verdeelde stochastische variabelen met verwachting μ. Voor steekproeven met omvang $n \leq 8$ vond hij deze toets nog redelijk efficient. Aangezien de beschouwde stochastische variabele $\underline{h}_n$ naast de twee buitenste waarnemingen ook de middelste waarneming gebruikt en bovendien de mediaan als schatter voor het gemiddelde nauwkeuriger is dan $(\underline{x}_{(n)}+\underline{x}_{(1)})/2$, is het te verwachten dat deze bij gebruik als toetsingsgrootheid iets beter zal zijn. Daarom kan worden verwacht dat de doeltreffendheid bij gebruik van $\underline{h}_n$ nog redelijk zal zijn voor $n \leq 9$.

Doeltreffendheid heeft overigens hier niet de gebruikelijke betekenis. De gewone grootheden voor de doeltreffendheid hebben immers een asymptotische karakter, terwijl hier slechts de bruikbaarheid van belang is voor zeer kleine steekproeven. Voor toetsen die voorgesteld worden ter vervanging van een op de Studentgrootheid gebaseerde toets bij kleine steekproefomvang wordt de doetreffendheid van een toets met steekproefomvang m gedefinieerd als het quotient van de steekproefomvang van de bijpassende op de Studentgrootheid gebaseerde toets met gelijke onbetrouwbaarheid en onderscheidingsvermogen en m. Een moeilijkheid bij het doen van uitspraken over deze efficientie is dat deze grootheid afhangt van zowel de toetsingsgrootheid en de beide hypothesen. Bij wijze van voorbeeld wordt voor één van de navolgende gebruiksmogelijkheden voor een speciaal geval de doeltreffendheid berekend.

4.2. GEBRUIKSMOGELIJKHEDEN

A. In het voorafgaande is steeds verondersteld dat de onderliggende stochastische variabelen standaardnormaal verdeeld zijn. Voor $\underline{x}_i$ $(i = 1,\ldots,n)$ normaal verdeeld met verwachting μ en variantie σ^2 zij $\underline{M}_n$ de mediaan en $\underline{R}_n$ de spreidingsbreedte, dan geldt

$$\frac{\underline{M}_n}{\underline{R}_n} = \frac{\underline{m}_n\sigma + \mu}{\underline{r}_n\sigma} = \frac{\underline{m}_n + \mu/\sigma}{\underline{r}_n} = \underline{h}_n^{\mu/\sigma} .$$

Hieruit blijkt dat ook voor de algemene normale verdeling de hier beschouwde stochastische variabele bruikbaar is.

B. In de plaats van de Studenttoets voor één steekproef kan een toets met een volgens $\underline{h}_n$ verdeelde toetsingsgrootheid worden gebruikt. Als immers gegeven zijn $\underline{M}_n$, en $\underline{R}_n$, gedefinieers als onder A, en de hypothesen zijn

$$H_0 : \mu = \mu_0 ,$$

$$H_1 : \mu = \mu_1 ,$$

dan geldt voor $(\underline{M}_n - \mu_0)/\underline{R}_n$ onder de nulhypothese

$$\frac{\underline{M}_n - \mu_0}{\underline{R}_n} = \frac{\underline{m}_n\sigma + \mu_0 - \mu_0}{\sigma\underline{r}_n} = \frac{\underline{m}_n}{\underline{r}_n} = \underline{h}_n$$

en onder het alternatief

$$\frac{\underline{M}_n - \mu_0}{\underline{R}_n} = \frac{\underline{m}_n \sigma + \mu_1 - \mu_0}{\sigma \underline{r}_n} = \underline{h}_n^{(\mu_1-\mu_0)/\sigma} ,$$

zodat $(\underline{M}_n - \mu_0)/\underline{R}_n$ als toetsingsgrootheid kan worden gebruikt. Juist zoals bij de Studenttoets kan de toets ook met samengestelde hypothesen zoals

$$H_0 : \mu \leq \mu_0$$

$$H_1 : \mu > \mu_0$$

worden gebruikt, en geldt dat het onderscheidingsvermogen slechts afhangt van n, $\mu-\mu_0$ en σ.

C. Bij kwaliteitscontrole aan de hand van continue variabelen toetst men vaak hypothesen omtrent μ/σ.
Van een partij goederen wenst men dan een uitspraak te doen over de fractie van de goederen, waarvoor de te controleren grootheid z groter is dan de norm L. Als men aanneemt dat $\underline{z}$ een normaal verdeelde stochastische variabele is met verwachting μ^* en variantie σ^{*2}, dan komt dit overeen met toetsen omtrent $p = P\{\underline{z} > L\}$ van, bijvoorbeeld, de hypothesen

$$H_0 : p \leq p_0$$

$$H_1 : p > p_0 .$$

Men kan H_0 herschrijven als

$$p = P\{\underline{z} > L\} = P\{\underline{u} \leq \frac{\mu^* - L}{\sigma^*}\} \leq p_0 ,$$

ofwel als

$$\frac{\mu^* - L}{\sigma^*} \leq \Phi^{-1}(p_0).$$

De variabele $\underline{y}$ met $\underline{y} = \underline{z} - L$ is nu normaal verdeeld met verwachting $\mu = \mu^* - L$ en variantie $\sigma^2 = \sigma^{*2}$. De oorspronkelijke toets komt dus overeen met het toetsen voor de verdeling van $\underline{y}$ van de hypothesen

$$H_0 : \mu/\sigma \leq \delta_0 ,$$

$$H_1 : \mu/\sigma > \delta_0 ,$$

met $\delta_0 = \Phi^{-1}(p_0)$. Voor deze toets op de verhouding van μ en σ kan voor een van de hier beschouwde waarden n van de steekproefomvang gebruik gemaakt worden van de toetsingsgrootheid $\underline{M}_n/\underline{R}_n$, waarvoor immers geldt

$$\frac{\underline{M}_n}{\underline{R}_n} = \underline{h}_n^{\mu/\sigma} .$$

Voor een speciaal geval van deze toets wordt nu de doeltreffendheid berekend. Als een steekproef $\underline{x}_i$ $(i = 1,\ldots,9)$ uit een niet-ontaarde normale verdeling wordt gebruikt voor het met onbetrouwbaarheid .05 toetsen van de hypothesen

$$H_0 : \mu = 0 ,$$

$$H_1 : \mu = \tfrac{1}{2}\sigma ,$$

dan wordt bij gebruik van de toetsingsgrootheid $\underline{M}_9/\underline{R}_9$ verworpen als

$$\frac{\underline{M}_9}{\underline{R}_9} > (F_9)^{-1}(.05) ,$$

dus als

$$\frac{\underline{M}_9}{\underline{R}_9} > .24 .$$

Het onderscheidingsvermogen is nu

$$P\{\frac{\underline{M}_9}{\underline{R}_9} > .24 \mid H_1\} = P\{\underline{h}_9^{1/2} > .24\} = .31 .$$

Voor de berekening van de doeltreffendheid moet nu de (eventueel niet gehele) waarde n worden gevonden uit

$$\begin{cases} P\{\underline{t}_{n-1} > t\} = .05 \\ P\{\underline{t}_{n-1}^{3/2} > t\} = .31 . \end{cases}$$

Door gebruik te maken van benaderingen voor de Student verdelingen kan dit numeriek worden opgelost en voor de doeltreffendheid e worden gevonden

$$e = \frac{m}{9} = .38 \ .$$

Voor de grotere der beschouwde waarden van de steekproefomvang moet gezien de gevonden lage waarde van de doeltreffendheid de bruikbaarheid toch worden betwijfeld.

D. Tot slot wordt vermeld dat de verdeling van $\underline{h}_n$ ook gebruikt kan worden bij het berekenen van voorspellings- en betrouwbaarheidsintervallen. Als $a = (F_n)^{-1}(1-\alpha/2)$, dan kan, bijvoorbeeld, als voorspellingsinterval voor μ worden geschreven

$$P\{\underline{M}_n - a\underline{R}_n < \mu \leq \underline{M}_n + a\underline{R}_n\} = 1 - \alpha \ .$$

Tabel 1. De percentielen van $\underline{h}_n$.

$F_n(t)\backslash t$	n = 3	n = 5	n = 7	n = 9
.50	0	0	0	0
.51	.01047	.00591	.004	.000
.52	.02096	.01184	.008	.001
.53	.03146	.01778	.011	.002
.54	.04199	.02375	.016	.006
.55	.05255	.02971	.020	.010
.56	.06317	.03570	.024	.014
.57	.07384	.04172	.028	.019
.58	.08458	.04778	.033	.022
.59	.09539	.05388	.038	.026
.60	.10630	.06002	.045	.030
.61	.11731	.06621	.047	.035
.62	.12843	.07246	.051	.039
.63	.13968	.07877	.056	.041
.64	.15107	.08517	.062	.045
.65	.16262	.09160	.067	.048
.66	.17435	.09814	.071	.051
.67	.18627	.10478	.082	.054
.68	.19840	.11151	.086	.059
.69	.21077	.11835	.090	.065
.70	.22339	.12532	.094	.069
.71	.23630	.13242	.098	.073
.72	.24952	.13966	.102	.078
.73	.26310	.14706	.107	.083
.74	.27705	.15464	.111	.087
.75	.29144	.16242	.117	.093
.76	.30630	.17040	.123	.099
.77	.32169	.17859	.129	.106
.78	.33768	.18705	.136	.112
.79	.35434	.19581	.144	.120

vervolg tabel 1.

$F_n(t)\backslash t$	n = 3	n = 5	n = 7	n = 9
.80	.37175	.20485	.153	.128
.81	.39002	.21427	.161	.134
.82	.40927	.22408	.165	.140
.83	.42965	.23432	.172	.146
.84	.45133	.24507	.182	.152
.85	.47453	.25638	.194	.158
.86	.49954	.26835	.201	.165
.87	.52668	.28111	.205	.171
.88	.55642	.29475	.215	.177
.89	.58933	.30946	.225	.184
.90	.62620	.32546	.236	.191
.91	.66811	.34303	.247	.200
.92	.71662	.36261	.258	.208
.93	.77401	.38476	.272	.218
.94	.84390	.41034	.289	.230
.95	.93229	.44072	.310	.240
.96	1.05019	.47834	.334	.256
.97	1.22064	.52781	.363	.276
.98	1.50336	.60033	.404	.304
.99	2.13585	.73495	.479	.381

Tabel 2. De bestaande tweede en vierde momenten.

	n = 3	n = 5	n = 7	n = 9
μ_2	-	.0824	.0373	.0224
μ_4	-	-	.0069	.0019

Tabel 3. Constanten β_n, γ_n, ν_n voor benaderingen van $\underline{h}_n$ en $\underline{h}_n^{\delta}$.

	Theoretische benadering		Verbeterde benadering		Benadering met gehele ν_n	
n	β_n	γ_n	ν_n	γ_n	ν_n	γ_n	ν_n
3	1.493	.350	2.0	.370	2.38	.319	2
5	1.867	.216	3.8	.228	4.61	.219	5
7	2.180	.162	5.5	.168	6.78	.164	7
9	2.454	.132	7.0	.136	9.17	.137	9

Tabel 4. Enige verbeterde waarden β_n^* van β.

n	β_n	δ	$t_{.95}$	$F_n^{\delta}(t_{.95})$	β_n^*
3	1.493	.1	1.049	.9483	1.70
		.3	1.291	.9441	1.74
		.5	1.550	.9400	1.76
5	1.867	.1	.526	.9559	1.16
		.3	.656	.9542	1.69
		.5	.792	.9522	1.80
7	2.180	.1	.366	.9535	1.79
		.3	.464	.9549	1.98
		.5	.567	.9557	2.03
9	2.454	.1	.289	.9586	1.47
		.3	.373	.9410	2.77
		.5	.459	.9552	2.32

Literatuur

1 Bouma, N. en A. Vehmeyer.
Percentiles and approximations of the sample median over the sample range of samples of size 3, 5, 7 and 9 from a standard normal distribution.
Statistica Neerlandica 23 (1969), 235-239.

2 Chu, J.T.
On the distribution of the sample median.
Ann. Math. Stat. 26 (1955), 112-116.

3 Cox, D.R.
The use of the range in sequential analysis.
J.R. Stat. Soc. B 11 (1949), 101-114.

4 David, H.A.
Further applications of range to the analysis of variance.
Biometrika 38 (1951), 393.

5 Haan, L. de en J.Th. Runnenburg
Some remarks concerning the quotient of sample median and sample range for samples of size 2n+1 from a normal distribution.
Statistica Neerlandica 23 (1969), 227-234.

6 Moore, P.G.
The two-sample t-test based on the range.
Biometrika 44 (1957), 482-489.

7 Patnaik, P.B.
The use of the mean range as an estimator of variance in statistical tests.
Biometrika 37 (1950), 78-87.

8 Pearson E.S.
Comparison of two approximations to the distribution of the range in small samples from the normal population.
Biometrika 39 (1952), 130-136.

9 Sarhan, A.E. and G. Greenberg.
Contribution to order statistics.
John Wiley & Sons, Inc., New York - London 1962.

10 Thomson, G.W.M.
Scale factors and degrees of freedom for small sample sizes for X-approximation to the range.
Biometrika 40 (1953), 449-450.

11 Walsh, J.E.
On the range-midrange test and some tests with bounded significance levels.
Ann. Math. Stat. 20 (1949), 257-267.

12 Wilks, S.S.
Mathmatical Statistics.
John Wiley & Sons, Inc., New York - London 1961.