Steekproeven uit de halve cauchy verdeling

door J. L. MIJNHEER *

Summary Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a sample from a distribution with infinite expectation, then for $n \to \infty$ the sample average $\bar{\underset{\sim}{x}}_n$ tends to $+\infty$ with probability 1 (see [4]).

Sometimes $\bar{\underset{\sim}{x}}_n$ contains high jumps due to large observations. In this paper we consider samples from the "absolute Cauchy" distribution. In practice, on may consider the logarithm of the observations as a sample from a normal distribution. So we found in our simulation. After rejecting the log-normality assumption, one will be tempted to regard the extreme observations as outliers. It is shown that the discarding of the outlying observations gives an underestimation of the expectation, variance and 99 percentile of the actual distribution.

1. Inleiding

Zij $\underset{\sim}{x}$ een stochastische variabele met verdelingsfunctie

$$F(x) = \begin{cases} \frac{2}{\pi} \operatorname{arctg} x & \text{als } x \geqslant 0 \\ 0 & x < 0. \end{cases}$$

Voor een steekproef $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ uit deze „absolute Cauchy" verdeling, ook wel halve Cauchy verdeling genoemd, geldt:

$$\frac{\underset{\sim}{x}_1 + \ldots + \underset{\sim}{x}_n}{n} \xrightarrow{\text{a.s.}} \infty \qquad (zie\ [4]).$$

De stochastische variabele $\underset{\sim}{y}$ met

$$\underset{\sim}{y} = \log \underset{\sim}{x}$$

heeft verdelingsfunctie

$$G(y) = P(\underset{\sim}{y} \leqslant y) = P(\underset{\sim}{x} \leqslant e^y) = \frac{2}{\pi} \operatorname{arc\,tg} e^y$$

en verdelingsdichtheid

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

Voor $\underset{\sim}{y}$ geldt dan:

1. Alle momenten zijn eindig.
2. Alle oneven momenten zijn nul. In het bijzonder $\mathscr{E}\underset{\sim}{y} = 0$.
3. $\sigma^2(\underset{\sim}{y}) = (\pi/2)^2$.

Uit figuur 1 blijkt dat de verdeling van $\underset{\sim}{y}$ grote overeenkomst vertoont met die van een $N(0, (\pi/2)^2)$-verdeelde stochastische variabele. In de praktijk zal men $\log \underset{\sim}{x}_1, \ldots, \log \underset{\sim}{x}_n$ dus kunnen aanzien voor een steekproef uit een normale verdeling. Dit betekent dat men bij verdere berekeningen doet alsof $\underset{\sim}{x}$ een log-normale verdeling bezit en dus een eindige verwachting en variantie heeft.

* Technische Hogeschool Twente.

De log-normaliteit zal gebruikt worden voor het berekenen van de schatters voor verwachting, variantie en percentielen en het toetsen op uitschieters.

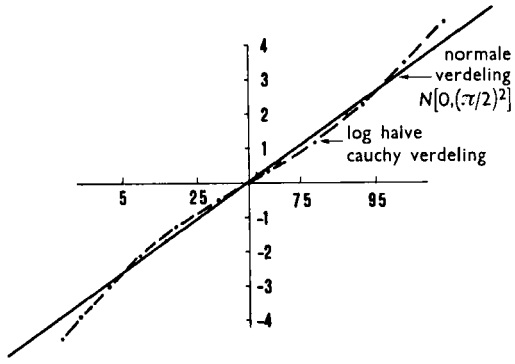

Fig. 1 Verdelingsfuncties van de log-halve Cauchy verdeling en de $N(0, (\pi/2)^2)$-verdeling uitgezet op waarschijnlijkheidspapier.

2. Numerieke resultaten

Voor verschillende steekproefomvang zijn steeds 100 steekproeven uit de halve Cauchy verdeling gesimuleerd. Voor de beschrijving van de simulatie zij verwezen naar [4]. De logaritmen van de waarnemingen uit elke steekproef zijn m.b.v. de toets van SHAPIRO en WILK (zie [5]) op normaliteit getoetst. De resultaten hiervan staan in tabel 2. Bij het toetsen is een onbetrouwbaarheidsdrempel $\alpha = 0{,}05$ gebruikt.

Tabel 2

$n =$	20	25	30	35	40	45	50
I	85	84	81	76	80	72	80
II	15	16	19	24	20	28	20

n = steekproefomvang
I = aantal keren dat normaliteit niet wordt verworpen
II = aantal keren dat normaliteit wel wordt verworpen

Verwerping van normaliteit kan een gevolg zijn van de aanwezigheid van uitschieters in de steekproef. Daarom zijn de steekproeven, die hiervoor in aanmerking kwamen, op de aanwezigheid van uitschieters getoetst met de toets van DOORNBOS (zie [2]). Daarbij blijken ook uitschieters naar beneden te kunnen optreden. In de steekproeven uit de halve Cauchy verdeling zijn dit kleine waarnemingen. Deze kleine waarnemingen hebben een minder opvallende invloed op het gedrag van het steekproefgemiddelde dan de grote waarnemingen, die de sprongen veroorzaken.

Een toets voor zowel uitschieters naar boven als naar beneden wordt niet gegeven in [2]. Voor dergelijke situaties is de uitschietertoets tweemaal toegepast.

We kunnen de steekproeven nu in vier typen indelen.

I: normaliteit van de hele steekproef wordt niet verworpen.
II: idem; toch zijn er „uitschieters" in de steekproef.
III: in eerste instantie wordt de normaliteit voor de hele steekproef verworpen; na het weglaten van de „uitschieters" wordt de normaliteit van de overige waarnemingen in de steekproef niet meer verworpen.
IV: blijvend niet normaal.

In tabel 3 staan de aantallen van ieder type bij elke steekproefomvang vermeld. Bij het toetsen op uitschieters is ook een onbetrouwbaarheidsdrempel $\alpha = 0{,}05$ gebruikt.

Tabel 3

$n =$	20	25	30	35	40	45	50
I	71	73	75	72	74	60	69
II	14	11	6	4	6	12	11
III	12	13	16	21	17	21	18
IV	3	3	3	3	3	7	2

In de figuren 2 t/m 5 is van ieder type één steekproef op waarschijnlijkheidspapier uitgezet. In 3. worden deze figuren nader bekeken en de betekenis van de rechten gegeven.

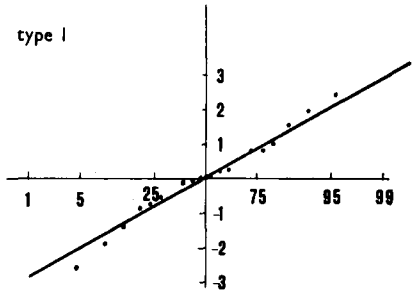

Fig. 2 Logaritmen van de waarnemingen uit een steekproef van de halve Cauchy verdeling uitgezet op waarschijnlijkheidspapier. Steekproef van type I.

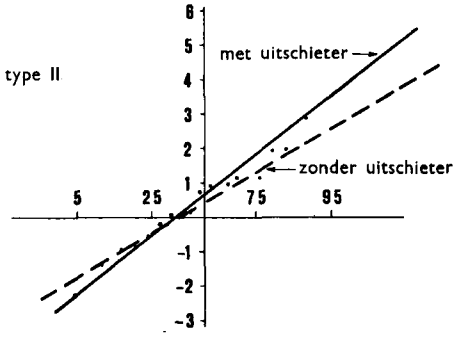

Fig. 3 Steekproef van type II. Bezit één uitschieter naar boven.

3. Effect van het weglaten van de uitschieters

AITCHISON en BROWN geven in [1] verschillende methoden om de verwachting en variantie van de log-normale verdeling te schatten. Hiervan zijn er een paar gebruikt.

a. Methode van de meest aannemelijke schatter

Deze geeft voor de parameters van de bijbehorende normale verdeling de schatters

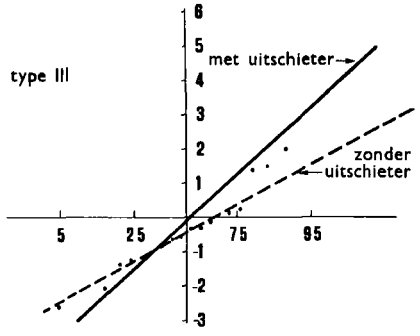

Fig. 4 Steekproef van type III. Bezit één uitschieter naar boven.

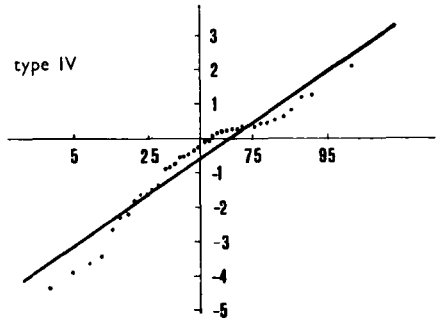

Fig. 5 Steekproef van type IV.

$\bar{y}$ en s_y^2. Door substitutie van $\bar{y}$ en s_y^2 in de formules voor de verwachting, variantie en het 99 percentiel van de log-normale verdeling krijgen we de meest aannemelijke schatters voor deze grootheden. Dit geeft:

$$a_1 = \exp(\bar{y} + \tfrac{1}{2}s_y^2),$$
$$b_1^2 = \exp(2\bar{y} + s_y^2)(\exp s_y^2 - 1),$$
$$P_{99} = \exp(\bar{y} + 2.33 s_y).$$

b. Momentenmethode

Alleen het steekproefgemiddelde $a_2 = \bar{x}$ als schatter voor de verwachting is berekend. De momentenschatter voor de variantie is veel te onnauwkeurig.

De schatters a_1 en b_1^2 zijn niet zuiver. FINNEY geeft in [3] een methode om zuivere schatters met minimale variantie te bepalen. In onze berekeningen verschillen deze schatters en de meest aannemelijke echter weinig.

Het is duidelijk dat, na het weglaten van de uitschieters, s_y^2 een veel kleinere waarde krijgt terwijl $\bar{y}$ slechts weinig verandert. De schatters a_1, b_1^2 en P_{99} zijn monotoon stijgende functies van s_y^2. Zij krijgen dus ook kleinere waarden. De schatter a_2 daarentegen is veel gevoeliger voor uitschieters naar boven dan voor die naar beneden.

In de figuren 2 t/m 5 zijn de logaritmen van de geordende waarnemingen op waarschijnlijkheidspapier uitgezet. Voor steekproeven uit een log-normale verdeling liggen deze punten bij benadering op de rechte

$$\log x = \sigma v + \mu,$$

waarbij v het quantiel in de $N(0,1)$-verdeling is.

In de figuren zijn deze lijnen getekend, waarbij voor μ en σ de meest aannemelijke schatters zijn gebruikt. In de figuren 3 en 4 zijn de lijnen ook getekend na het weglaten van de uitschieters. Deze lijnen lopen minder steil en „passen beter". Hierdoor zal men nog sterker de indruk van log-normaliteit krijgen. Deze procedure leidt echter tot verdere onderschatting.

4. Dankwoord

Voor alle medewerking ben ik Prof. Dr. J. HEMELRIJK veel dank verschuldigd.

5. Literatuur

[1] AITCHISON, J. and J. A. C. BROWN, The log-normal distribution. 1st ed., Cambridge, 1963.
[2] DOORNBOS, R., Slippage tests (Akademisch Proefschrift), 1966.
[3] FINNEY, D. J., On the distribution of a variate whose logarithm is normally distributed, Suppl. J.R. Statist. Soc. 7 (1941), p. 145.
[4] MIJNHEER, J. L., The conduct of the sample average when the first moment is infinite. Statistica Neerlandica **22** (1968), pp. 37–41.
[5] SHAPIRO, S. S. and M. B. WILK, An analysis of variance test for normality, Biometrika **52** (1965), pp. 591–611.