MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 1956-10(2)

De log-normale verdeling bij de statistische verwerking van omzet cijfers

door

Prof.Dr J. Hemelrijk

Juni 1956

Wij beschouwen een populatie van winkeliers. De jaaromzet in guldens van een winkelier wordt aangegeven door y, zodat $\underline{y}$ op de beschouwde populatie een verdeling bezit.[1] Deze grootheid $\underline{y}$ wordt ondersteld een log-normale verdeling te bezitten, d.w.z. dat

$$(1) \qquad \underline{x} = c \log \underline{y}, \quad \underline{y} = e^{\underline{x}/c}$$

waarin c een willekeurige constante is en log de natuurlijke logarithme voorstelt, normaal verdeeld is. De verwachting resp. de spreiding van $\underline{x}$ geven wij aan met μ resp. σ en de verdelingsdichtheid resp. de verdelingsfunctie met $f(x)$ resp. $F(x)$.

Dus

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

en

$$(3) \qquad F(x) = \int_{-\infty}^{x} f(u)\, du .$$

Voor de verdelingsfunctie $G(y)$ van $\underline{y}$ geldt dan:

$$G(y) = P[\underline{y} \leqq y] = P[\underline{x} \leqq c \log y] = F(c \log y),$$

of wel

$$(4) \quad G(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{c \log y} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\, dx .$$

De verdelingsdichtheid van $\underline{y}$ wordt hieruit verkregen door naar y te differentiëren. Dit geeft

$$(5) \quad g(y) = \frac{c}{y}\, \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{(c \log y - \mu)^2}{\sigma^2}} .$$

De verwachting van de omzet $\underline{y}$ (dus de gemiddelde opbrengst over de populatie) is dan gelijk aan

$$(6) \quad \mathcal{E}\,\underline{y} = e^{\frac{\mu}{c} + \frac{\sigma^2}{2c^2}} . \qquad (*)$$

---

1) Stochastische grootheden worden aangegeven door onderstreepte symbolen, terwijl dezelfde symbolen, niet onderstreept, waarden voorstellen, die zij aan kunnen nemen.

De vraag, waar het nu om gaat is de volgende. Op welke wijze kan berekend worden:

1) $Z(y_0)$, de gezamelijke omzet van winkeliers, die ieder een omzet $\geq y_0$ bezitten,
2) $U(y_0)$, het percentage, dat $Z(y_0)$ van de totale omzet van alle winkeliers tezamen vormt.
3) Hoe kunnen deze grootheden geschat worden uit aan een steekproef van winkeliers ontleende omzet-cijfers?

De oplossing voor vraag 1) wordt gegeven door:

$$(7) \quad Z(y_0) = \mathcal{E}\, \underline{y} \left\{ 1 - F(x_0 - \tfrac{\sigma^2}{c}) \right\} , \quad (*)$$

waarin $x_0 = c \log y_0$ is en $\mathcal{E}\, \underline{y}$ en F door (6) en (3) gegeven worden.

Het antwoord op vraag 2) luidt:

$$(8) \quad U(y_0) = 100 \left\{ 1 - F(x_0 - \tfrac{\sigma^2}{c}) \right\} = \frac{100\, Z(y_0)}{\mathcal{E}\, \underline{y}} . \quad (*)$$

Het antwoord op vraag 3) is niet in één formule te geven, daar hiervoor verschillende methoden gevolgd kunnen worden.

Om te beginnen merken wij op, dat μ en σ de enige onbekende parameters zijn, immers c kan willekeurig gekozen worden. Worden de aan de steekproef ontleende waarnemingen aangegeven door

$$(9) \quad y_1, y_2, \cdots, y_n ,$$

dan kan men μ schatten door

$$(10) \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (x_i = c \log y_i)$$

en σ^2 door

$$(11) \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Deze methode, die zuivere schattingen voor μ en σ^2 geeft en die nauw aansluit bij de methode der meest aannemelijke schattingen, is de meest doeltreffende, doch eist nogal wat rekenwerk voor de berekening van s^2.

(*) De bewijzen van de formules, die van het teken (*) voorzien zijn, worden in een appendix gegeven.

Minder rekenwerk is noodzakelijk voor directe schatting van μ en $\mathcal{E}_{\underline{y}}$, waaruit dan een andere schatting voor σ^2 volgt. Men kan nl. μ schatten door $\bar{x}$ (zie (10)) en $\mathcal{E}_{\underline{y}}$ door

$$(12) \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad ,$$

terwijl dan door invullen van deze schattingen in (6) volgt

$$\bar{y} = e^{\bar{x}/c + s_1^2/2c^2} \quad ,$$

waarin s_1^2 de schatting voor σ^2 voorstelt. Deze schatting is, zoals door oplossen direct blijkt

$$(13) \qquad s_1^2 = 2c\,(c \log \bar{y} - \bar{x}) \quad .$$

Volgt men één van deze beide methoden, dan dient men dus de x_i in een log-tafel op te zoeken. Beschikt men daarbij over een tabel van de natuurlijke logarithmen, dan kan men $c = 1$ kiezen, hetgeen de formules vereenvoudigt. Gebruikt men Briggse logarithmen, dan kieze men

$$(14) \qquad c = {}^{10}\log e = 0{,}434\,29 \quad ,$$

waardoor

$$(15) \qquad x = {}^{10}\log y \qquad \text{en} \qquad y = 10^x$$

wordt, dus ook

$$(15') \qquad x_i = {}^{10}\log y_i \quad .$$

Heeft men schattingen voor μ en σ^2 berekend, dan kan men die in (7) en (8) invullen en met behulp van een tabel van de normale verdeling voor iedere gewenste waarde van y_0 de functies $Z(y_0)$ en $U(y_0)$ berekenen.

Eenvoudiger, doch minder nauwkeurig, is een grafische methode, waarbij gebruik gemaakt wordt van log-normaal waarschijnlijkheidspapier.

Bij dit papier staat (b.v. op de horizontale as) een percentage uitgezet. De schaal die daarbij gebruikt is, is aangepast aan de normale verdeling. Dit percentage zullen wij aangeven met P. Op de andere as (de verticale) staat een logarithmische schaal aangegeven. Dit betekent, als wij de bij deze schaal vermelde cijfers als waarden van y opvatten, dat, lineair beschouwd, juist $x = c \log y$ uitgezet wordt, waarin c willekeurig

gekozen kan worden zolang geen lineaire (x-) schaal naast de y -schaal aangegeven is.

Zetten wij op dit papier de functie

$$(16) \quad P(y) = 100\left\{1 - F(x)\right\} = 100\left\{1 - F(c\log y)\right\}$$

uit, dan geeft dit een rechte lijn. Een schatting van deze rechte kan verkregen worden door een aantal punten uit te zetten, met y_i als ordinaat en als abscis het percentage der waargenomen omzetten, die minstens gelijk aan y_i zijn. Door de verkregen punten kan dan op het oog een rechte lijn getrokken worden.

De vergelijking (8) wordt nu, zoals in de appendix bewezen wordt, voorgesteld door een aan de vorige evenwijdige lijn, waarvan één punt op de volgende wijze geconstrueerd kan worden.

Trek 2 verticale lijnen, één door het punt $P=50$ en één door het punt $P=16$ op de horizontale as. Lees de bijbehorende waarden van y op de lijn, die (16) voorstelt af. Noem deze y_{50} en y_{16} .

Dan is

$$(17) \quad x_{50} = c\log y_{50} \qquad (*)$$

een schatting van μ en

$$(18) \quad s_2 = c\log\frac{y_{16}}{y_{50}} \qquad (*)$$

een schatting van σ.

Bereken nu

$$(19) \quad y'_{50} = y_{50}\cdot e^{\left(\log\frac{y_{16}}{y_{50}}\right)^2} = y_{50}\cdot 10^{\frac{\left(^{10}\log\frac{y_{16}}{y_{50}}\right)^2}{0,43429}},$$

dan is y'_{50} een schatting van het 50% -punt van de lijn, die (8) voorstelt.

<u>Voorbeeld</u>

Stel $y_{50}=3,05$ en $y_{16}=8,5$, dan is

$y'_{50} = 3,05\cdot e^{\left(\log\frac{8,5}{3,05}\right)^2} = 3,05\cdot e^{1,05} = 3,05\cdot 2,86 = 8,72$.

Ook kan men de schattingen (10) en (11) voor μ en σ^2 in (8) invullen en voor twee waarden van y_0 de bijbehorende waarden van $\mathcal{U}(y_0)$ berekenen. Dit geeft twee punten op het grafische papier, waardoor dan een rechte getrokken kan worden, die een schatting van (8) voorstelt.

Appendix

Uit de definitie van $Z(y_o)$ volgt:

$$Z(y_o) = \int_{y_o}^{\infty} y\, g(y)\, dy = \frac{c}{\sigma\sqrt{2\pi}} \int_{y_o}^{\infty} e^{-\frac{1}{2}\frac{(c\log y - \mu)^2}{\sigma^2}}\, dy .$$

Stellen wij nu weer

(1) $\quad x = c \log y \quad$, dus $\quad y = e^{x/c}$,

dan is

$$\frac{dy}{dx} = \frac{1}{c}\, e^{x/c}$$

en stellen wij $x_o = c \log y_o$ dan gaat $Z(y_o)$ door de substitutie (1) over in

$$Z(y_o) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_o}^{\infty} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} + \frac{x}{c}}\, dx .$$

De exponent van e laat zich herleiden tot

$$-\frac{1}{2}\frac{\left(x-\mu-\frac{\sigma^2}{c}\right)^2}{\sigma^2} + \left(\frac{\mu}{c} + \frac{\sigma^2}{2c}\right) ,$$

zodat

$$Z(y_o) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{\mu}{c} + \frac{\sigma^2}{2c^2}} \int_{x_o}^{\infty} e^{-\frac{1}{2}\frac{\left(x-\mu-\frac{\sigma^2}{c}\right)^2}{\sigma^2}}\, dx = e^{\frac{\mu}{c} + \frac{\sigma^2}{2c^2}} \left\{1 - F\left(x_o - \frac{\sigma^2}{c}\right)\right\}$$

is. Voor $x_o \to -\infty$ verkrijgen wij, daar dan $y_o \to o$ en $F(x_o - \frac{\sigma^2}{c}) \to o$:

$$Z(o) = \int_{o}^{\infty} y\, g(y)\, dy = \mathcal{E}\, y = e^{\frac{\mu}{c} + \frac{\sigma^2}{2c^2}} .$$

Hiermede zijn de formules (6), (7) en (8) bewezen.

Vervolgens gaan wij over tot de beschouwing van de in het vorige beschreven grafische methode. Zoals reeds vermeld, wordt vergelijking (16) op dit log-normaal waarschijnlijkheidspapier door een rechte lijn voorgesteld. De verklaring hiervan laten wij achterwege, doch wij wijzen erop, dat de verticale schaal, lineair beschouwd, juist x aangeeft. Voor x is dit papier dus gewoon waarschijnlijkheidspapier, hetgeen overeenstemt met het feit, dat $\underline{x}$ normaal verdeeld is.

Derhalve kan de verwachting μ van $\underline{x}$ geschat worden door het 50 %-punt van (16), dat wij met x_{50} aangegeven hebben. De waarde van x_{50} kan niet afgelezen worden, daar in de regel geen lineaire schaal op dit papier aanwezig is, doch de bijbehorende waarde y_{50} van y kan wel afgelezen worden. Volgens (1) geldt dan (17).

Beschouwen wij nu vergelijking (18), waarbij $U(y_o)$ op de horizontale schaal van het grafiekenpapier betrekking heeft, dan zien wij, dat het tweede lid van deze vergelijking geheel overeenstemt met dat van (16), als wij

(20) $\quad x = x_o - \frac{\sigma^2}{c} \qquad$ of $\qquad x_o = x + \frac{\sigma^2}{c}$

stellen. Dit betekent echter, dat $U(y_o) = P(y)$ is, indien voor de bijbehorende waarden x_o en x aan (20) voldaan is, dus dat de beide lijnen, die (8) en (16) voorstellen in verticale richting steeds op gelijke afstand van elkaar lopen, en wel - gemeten in de x-schaal - op een afstand $\frac{\sigma^2}{c}$. Zij zijn dus, daar de x-schaal lineair is, evenwijdig en wel ligt (8) boven (16). In de y-schaal gemeten is de afstand niet constant en het is juist deze y-schaal, die op het papier aangegeven is. Beschouw nu echter het 50%-punt van de beide lijnen.

Voor (16) is, afgezien van de schattingsfout in de lijn, die (16) voorstelt,

(21) $\quad P(y_{50}) = 50$

en voor (8) stellen wij

(22) $\quad U(y'_{50}) = 50$.

Dan is volgens het bovenstaande

(23) $\quad x'_{50} = x_{50} + \frac{\sigma^2}{c}$.

Hierin moet nu voor de onbekende σ^2 nog een schatting ingevuld worden. Een grafisch te verkrijgen schatting is de volgende. Noemen wij y_{16} het 16%-punt van de reeds geschatte lijn (16), dan is dus

(24) $\quad P(y_{16}) = 16$,

doch volgens een tabel van de normale verdeling is dan

(25) $\quad x_{16} - x_{50} = \sigma$,

zodat het eerste lid als schatting voor σ gebruikt kan worden. Dit leidt, tezamen met (1), direct tot (18).

Invullen in (23) geeft

(26) $\quad x'_{50} = x_{50} + \frac{(x_{16} - x_{50})^2}{c} = c \log y_{50} + \frac{(c \log y_{16} - c \log y_{50})^2}{c}$

en voor de overeenkomstige y-waarde

$$(27)\quad y'_{50} = e^{x'_{50}/c} = e^{\log y_{50}} \cdot e^{(\log y_{16} - \log y_{50})^2} = y_{50} \cdot e^{(\log \frac{y_{16}}{y_{50}})^2}.$$

Werkt men met $^{10}\log$, dan stelle men weer

$$(14)\quad c = 0{,}4342\, g$$

en dan komt er

$$(26')\quad x'_{50} = {}^{10}\log y_{50} + \frac{({}^{10}\log y_{16} - {}^{10}\log y_{50})^2}{c}$$

en

$$(27')\quad y'_{50} = y_{50} \cdot 10^{\frac{({}^{10}\log y_{16} - {}^{10}\log y_{50})^2}{c}} = y_{50} \cdot 10^{\frac{1}{c}({}^{10}\log \frac{y_{16}}{y_{50}})^2}.$$

Opmerking

Het is niet juist uit het voorafgaande te concluderen, dat de grootheid $\underline{z} = \underline{y}\, g(\underline{y})$ nu een logarithmisch normale verdeling bezit. Want weliswaar heeft $Z(y_0) = \int_{y_0}^{\infty} y\, g(y)\, dy$ op een constante na de vorm van een geïntegreerde log-normale verdeling, zodat $U(y_0)$ op log-normaal papier door een rechte voorgesteldwordt, maar dit betekent niet dat hetzelfde geldt voor $\underline{z}$. De reden hiervan is, dat $z = y\, g(y)$ geen monotone functie van y is. Zou men het percentage $W(y_0)$ berekenen, voorstellende de som der omzetten, waarvoor $\underline{z} \geq z_0$ (dus $y\, g(y) \geq y_0\, g(y_0)$) is, dan wordt een andere uitkomst verkregen. Hieronder zouden nl. de zeer grote waarden van y niet begrepen zijn, daar $g(y)$ voor die omzetten zeer gering is.