

Kwantielanalyse en de Negatief Binomiale Verdeling

Jan Engel*

Samenvatting

Veronderstel, dat een aan subjecten uit een populatie in principe meetbare variabele bij benadering een logistische kansverdeling heeft. Indien waarnemingen uit deze verdeling alleen indirect te verkrijgen zijn door telkens voor een subject uit de populatie vast te stellen of de ermee corresponderende variabele een gekozen niveau (stimulus) overschrijdt, kan men het schatten van de parameters van deze logistische verdeling op verschillende manieren aanpakken. Aantrekkelijk is de minimum logit chikwadraat methode van Berkson. Indien bij een zekere stimulus voor een gefixeerd aantal subjecten helemaal geen responsie blijkt op te treden, kan men deze methode niet zonder meer gebruiken. Voorgesteld wordt het experiment zo op te zetten, dat bij sommige waarden van de stimulus niet het aantal te onderzoeken subjecten wordt gefixeerd, maar het aantal responsies.

* Nederlandse Philips Bedrijven B.V.
Centre for Quantitative Methods
Gebouw VN-714
5600 MD Eindhoven
Tel. 040-735951

1. Inleiding

In 1955 bestudeerde Joseph Berkson het volgende probleem. Stel dat van subjecten uit een populatie een interessante variabele, bijvoorbeeld de letale dosis gif voor een mug uit een populatie muggen, of de ontsteekspanning van een bepaald type lamp uit een partij lampen een logistische kansverdeling bezit. Echter, de genoemde variabele is slechts indirect meetbaar en wel op de volgende wijze. Men kiest een bepaalde dosis gif, respectievelijk spanning en onderwerpt een aselekt uit de populatie genomen n -tal muggen/lampen aan deze dosis. Vervolgens wordt het aantal gedode muggen/ontstoken lampen, ofwel het aantal responsies y in het uitgevoerde binomiale experiment, geteld. Door zo'n experiment uit te voeren voor verschillende doses kan men een schatting bepalen van de parameters van de logistische verdeling; de beschreven techniek noemt men kwantiel of logit analyse.

Ook anderen hebben zich met dit probleem beziggehouden. Zo zijn er minstens drie manieren om schatters te berekenen: via maximum likelihood (Finney (1971)), via minimum Pearson chikwadraat en via minimum logit chikwadraat, ontwikkeld door de reeds genoemde Berkson (1955). Deze methode, ook beschreven in de handleiding van Corsten (1976) is erg aantrekkelijk: rekentechnisch eenvoudig (een niet-iteratieve schattingsmethode) en meestal nauwkeuriger dan de andere. Hij heeft echter het volgende bezwaar: indien bij een bepaalde dosis x volledig geen respons ($y=0$) of volledige respons ($y=n$) optreedt, kan men deze uitkomsten niet ongemodificeerd betrekken in het parameter schattingsproces. Dit zal vooral het geval zijn als x in een staart van de verdeling wordt gekozen. Berkson stelt als aanpassing voor dan maar $y=0.5$ respectievelijk $y=n-0.5$ in te vullen, hetgeen echter een vrij willekeurige keuze is.

In dit artikel wordt een andere oplossing voorgesteld, namelijk door het experiment iets anders op te zetten. Vaak heeft men namelijk al enig inzicht in de "plaats" van de logistische verdeling, en dus ook wel enig idee waar zich de staarten bevinden. Het is dan bijvoorbeeld nog nodig, nauwkeurig de parameters te bepalen van de verdeling.

Indien nu een dosis x in een der staarten van de verdeling wordt gekozen, stellen we voor niet het aantal te onderzoeken subjecten te fixeren en het aantal responsies te tellen, maar het aantal responsies te fixeren en het aantal te onderzoeken subjecten te tellen tot de s -de responsie verkregen is. In dat geval wordt bij de bewuste dosis x het binomiale deelexperiment vervangen door een negatief binomiaal. Men spreekt dan van inverse sampling, wel vertaald als de omgekeerde methode (Rümke en Kuik (1981)). Van te voren is nu de steekproefomvang niet bekend en daarom is zulks natuurlijk niet altijd te verwezenlijken. Maar indien dit wel het geval is, en daar waar inverse sampling zelfs voor de hand ligt, is hiermee het bezwaar van substitutie in de methode van Berkson opgeheven.

In paragraaf 2 zal het schattingsprobleem nader worden uitgewerkt, en een toepassing vindt men in paragraaf 3. Enkele wat technische afleidingen zijn tenslotte in de 3 aanhangsels opgenomen.

2. Het schattingsprobleem

- 2.1. Voor een uitvoeriger beschrijving van de manier waarop men parameters kan schatten van de logistische verdeling via de minimum logit chikwadraat methode van Berkson (1955), een speciaal geval van de gewogen kleinste kwadraten methode (Corsten (1976)), zij verwezen naar deze literatuur. Op deze plaats zal worden volstaan met een kort overzicht.

De schattingsmethode gaat als volgt in zijn werk. Zij $G(x) = F(a+bx) = [1+\exp(-(a+bx))]^{-1}$ de logistische verdelingsfunctie met parameters a en b , die een variabele eigenschap van subjecten uit een populatie beschrijft, bijvoorbeeld de ontsteekspanning van lampen. Men kiest nu verschillende doses x_i van bijvoorbeeld elektrische spanning en bepaalt de responsie y_i , het aantal ontstoken lampen, van telkens n_i subjecten uit de populatie.

Als i de waarden $1, 2, \dots, I$ doorloopt heeft men een I -tal schattingen van de kansen $p_i = F(a + bx_i)$ op responsie middels de waargenomen fracties $\hat{p}_i = y_i/n_i$. Schattingen van a en b verkrijgt men dan door een gewogen regressie uit te voeren van $u_i = F^{-1}(\hat{p}_i) = \log \hat{p}_i / (1 - \hat{p}_i)$ op de doses x_i via het model $F^{-1}(p_i) = a + bx_i$, waarbij de optredende gewichten g_i zo gekozen worden, dat g_i^{-1} gelijk is aan de variantie van $u_i = F^{-1}(\hat{p}_i)$ in een lineaire benadering, en derhalve gelijk aan $g_i = n_i p_i (1 - p_i)$ (zie aanhangsel 2). De gewichten g_i zijn onbekend en worden geschat uit de fracties $\hat{p}_i$ via $\hat{g}_i = n_i \hat{p}_i (1 - \hat{p}_i)$. De gewogen kleinste kwadraten schatters van a en b , verkregen door minimalisatie van de gewogen kwadraatsom

$$SS = \sum_{i=1}^I \hat{g}_i (u_i - a - bx_i)^2 \quad \dots\dots\dots (2.1)$$

zijn gelijk aan

$$\hat{a} = \bar{u} - \hat{b} \bar{x}$$

$$\hat{b} = \frac{\sum_i \hat{g}_i (x_i - \bar{x})(u_i - \bar{u})}{\sum_i \hat{g}_i (x_i - \bar{x})^2} \quad \dots\dots\dots (2.2)$$

waarin de gemiddelden $\bar{x}$ en $\bar{u}$ gewogen gemiddelden zijn, bijvoorbeeld $\bar{x} = \sum_i \hat{g}_i x_i / \sum_i \hat{g}_i$. Voor $\min_i n_i \rightarrow \infty$ zijn de schatters $\hat{a}$ en $\hat{b}$ asymptotisch raak, zuiver en normaal verdeeld, terwijl de geschatte asymptotische varianties gelijk zijn aan

$$\begin{aligned} \hat{\text{var}} \bar{u} &= 1 / \sum_i \hat{g}_i \\ \hat{\text{var}} \hat{a} &= 1 / \sum_i \hat{g}_i + \bar{x}^2 \hat{\text{var}} \hat{b} \\ \hat{\text{var}} \hat{b} &= [\sum_i \hat{g}_i (x_i - \bar{x})^2]^{-1} \end{aligned} \quad \dots\dots\dots (2.3)$$

Tenslotte is het minimum van SS, te weten

$$\hat{SS} = \sum_i g_i (u_i - \hat{a} - \hat{b} x_i)^2 \quad \dots\dots\dots (2.4)$$

asymptotisch chikwadraat verdeeld met I-2 vrijheidsgraden als het logistisch model juist is, en hiermee hebben we dan ook een modeltoets.

- 2.2 Tot zover gaat alles goed, maar als de responsie y_i gelijk aan 0 of n_i blijkt te zijn, hetgeen vaak gebeurt als de dosis x_i in een staart van de verdeling wordt gekozen, is bijvoorbeeld $u_i = -\infty$ en $g_i = 0$ en derhalve niet te gebruiken in de schattingsmethode. Voor deze doses is het vaak lastig een voldoende hoge " n_i met succesgarantie" vast te stellen. Toch is het juist voor een dergelijke x_i zinvol informatie over de verdeling te verkrijgen. Wenst men namelijk een nauwkeurige schatter van bijvoorbeeld parameter b , dan lijkt (zie (2.3)) juist een x_i in een staart van de verdeling daartoe bij te dragen, mits het bijbehorend gewicht g_i voldoende groot is.

De modificatie van Berkson (1955) door maar $y_i = 0.5$ of $y_i = n_i - 0.5$ in te vullen is tamelijk arbitrair en we stellen een iets andere opzet van het experiment voor, in een ander verband ook al eens door Finney (1971, p.211) gesuggereerd. Vaak heeft men al enig idee van de ligging van de staarten van de verdeling. Indien nu een dosis x_i in de linkerstaart van de verdeling wordt gekozen, dan onderwerpen we niet een vast aantal van n_i subjecten aan dosis x_i waarna we de responsie y_i vaststellen, maar we gaan uit van een vast aantal responsies s_i en tellen het aantal subjecten nodig voor het verkrijgen van die s_i responsies. In de rechterstaart tellen we, analoog, het aantal subjecten tot en met de s_i -de non-responsie. Noemen we (in het eerste geval) dit aantal $\underline{m}_i$, dan is $\underline{m}_i \stackrel{\text{def}}{=} m_i' - s_i$ een negatief binomiaal verdeelde grootte, met parameters s_i en p_i .

We zullen nu in 2.3 de schattingsmethode van Berkson modificeren door, indien mogelijk, het experiment voor sommige x_i op deze wijze uit te voeren.

- 2.3 Van een negatief binomiaal verdeelde grootheid $\underline{m}$, het aantal non-responsies tot de s -de responsie, zijn de volgende eigenschappen welbekend (Johnson en Kotz (1969), p.122)

$$E \underline{m} = sq/p \text{ en } \text{var } \underline{m} = sq/p^2$$

waarin p de kans op responsie is, en $q=1-p$.

Voorts is

$$\underline{m}/s \xrightarrow{s \rightarrow \infty} q/p \text{ voor } s \rightarrow \infty \text{ (stochastische convergentie).}$$

De schatter $1/(1+\underline{m}/s)$ van p is asymptotisch raak voor $s \rightarrow \infty$, en voor $s > 1$ zuiver te maken tot de schatter $(s-1)/(\underline{m}+s-1)$.

Voorts is $\underline{m}^* = (\underline{m} - E \underline{m})/\sqrt{(\text{var } \underline{m})}$ asymptotisch $N(0,1)$ verdeeld voor $s \rightarrow \infty$, op grond van de CLS. Dit resultaat is nog wat te verfijnen. Uit de stelling van Lindeberg volgt namelijk dat

$$\underline{m}^* \text{ is asymptotisch } N(0,1) \text{ verdeeld d.e.s.d.a. } sq \rightarrow \infty$$

(zie aanhangsel 1).

Laten we met I_2 de verzameling aangeven van die indices i , waarbij een negatief binomiaal deexperiment wordt gedaan.

We definiëren dan bij zo'n index i : $\underline{v}_i = F^{-1}(\hat{p}_i) = \log(\hat{p}_i/(1-\hat{p}_i)) = -\log(\underline{m}_i/s_i)$, waarbij $\hat{p}_i = 1/(1+\underline{m}_i/s_i)$ als schatter van p_i fungeert. Natuurlijk kan nu $m_i = 0$ optreden, maar dit zal voor kleine p_i en $s_i > 1$ niet zo gauw het geval zijn.

De kwadraatsom (2.1) krijgt de gedaante

$$SS = \sum_{i \in I_1} \hat{g}_i (u_i - a - bx_i)^2 + \sum_{i \in I_2} \hat{h}_i (v_i - a - bx_i)^2 \dots\dots\dots (2.5)$$

waarin $I_1 = \{1, 2, \dots, I\} \setminus I_2$,

en in deze kwadraatsom treden nieuwe gewichten $\hat{h}_i = s_i \hat{q}_i$, ($\hat{q}_i = 1 - \hat{p}_i$) op, die schattingen zijn van gewichten $h_i = s_i q_i$, in aanhangsel 2 afgeleid uit $\text{var } v_i$.

De schatters $\hat{a}$ en $\hat{b}$ verkrijgt men weer door minimalisatie, nu van (2.5), waarin gesommeerd wordt over het totaal aantal groepen I . Aangezien deze uitdrukking qua vorm identiek is aan (2.1), worden $\hat{a}$ en $\hat{b}$ weer door (2.2) gegeven, en geschatte asymptotische varianties door (2.3) met dien verstande, dat voor $i \in I_2$, gewicht $\hat{g}_i$ wordt vervangen door $\hat{h}_i$ en u_i door v_i . Genoemde asymptotische eigenschappen gelden als zowel $\min_{i \in I_1} n_i \rightarrow \infty$ als $\min_{i \in I_2} s_i \rightarrow \infty$. De schatters $\hat{a}$ en $\hat{b}$ zijn dan weer asymptotisch raak en zuiver.

Tenslotte is ook het minimum van (2.5) asymptotisch chikwaadraat verdeeld, met $I-2$ vrijheidsgraden.

Dit volgt uit het feit dat (2.5) te schrijven is als:

$$SS = \sum_{i \in I_1} (u_i - a - bx_i) \sqrt{(n_i \hat{p}_i \hat{q}_i)} \sqrt{(n_i \hat{p}_i \hat{q}_i)} (u_i - a - bx_i) + \\ + \sum_{i \in I_2} (v_i - a - bx_i) \sqrt{(s_i \hat{q}_i)} \sqrt{(s_i \hat{q}_i)} (v_i - a - bx_i)$$

Zowel $\sqrt{(n_i \hat{p}_i \hat{q}_i)} (u_i - a - bx_i)$, voor $n_i \rightarrow \infty$

als $\sqrt{(s_i \hat{q}_i)} (v_i - a - bx_i)$, voor $s_i \rightarrow \infty$

is asymptotisch standaard normaal verdeeld (zie aanhangsel 3). Derhalve gedraagt (2.5) zich wat limietverdeling betreft hetzelfde als (2.1), en voor het minimum $\hat{SS}$ van de kwadraatsom is dit ook weer zo, zodat een toets op goodness-of-fit verkregen is.

2.4 Samenvattend geven we een overzicht van de opgetreden formules.

Responsiemodel: bij dosis x_i is de kans op responsie

$$p_i = F(a+bx_i) = \{1+\exp(-(a+bx_i))\}^{-1} \Rightarrow F^{-1}(p_i) = a+bx_i$$

Binomiaal experiment: $y_i \sim \text{binomiaal}(n_i, p_i)$

(voor $i \in I_1$)

$$E y_i = n_i p_i$$

$$\text{var } y_i = n_i p_i (1-p_i)$$

$$\hat{p}_i = y_i/n_i; u_i = F^{-1}(\hat{p}_i)$$

$$\text{gewicht } g_i = \text{var}^{-1} \underline{u}_i \approx n_i p_i (1-p_i)$$

Negatief binomiaal

experiment:

$$\underline{m}_i \sim \text{negatief binomiaal}(s_i, p_i)$$

$$E \underline{m}_i = s_i (1-p_i)/p_i$$

(voor $i \in I_2$)

$$\text{var } \underline{m}_i = s_i (1-p_i)/p_i^2$$

$$\hat{p}_i = \frac{1}{1+\underline{m}_i/s_i}; v_i = F^{-1}(\hat{p}_i)$$

$$\text{gewicht } h_i = \text{var}^{-1} \underline{v}_i \approx s_i (1-p_i)$$

Kwadraatsom SS:

$$\begin{aligned} SS &= \sum_{i \in I_1} \hat{g}_i (u_i - a - bx_i)^2 + \sum_{i \in I_2} \hat{h}_i (v_i - a - bx_i)^2 \\ &= \sum_i k_i (w_i - a - bx_i)^2 \end{aligned}$$

waarin in een verkorte notatie:

$$g_i \text{ als } i \in I_1$$

$$k_i =$$

$$h_i \text{ als } i \in I_2$$

$$u_i \text{ als } i \in I_1$$

$$w_i =$$

$$v_i \text{ als } i \in I_2$$

Minimum van SS treedt op voor:

$$\hat{a} = \bar{w} - \hat{b} \bar{x}$$

$$\hat{b} = \frac{\sum_i \hat{k}_i (w_i - \bar{w})(x_i - \bar{x})}{\sum_i \hat{k}_i (x_i - \bar{x})^2}$$

waarin

$$\bar{x} = \sum_i \hat{k}_i x_i / \sum_i \hat{k}_i$$

$$\bar{w} = \sum_i \hat{k}_i w_i / \sum_i \hat{k}_i$$

en schattingen van asymptotische varianties zijn:

$$\text{var } \bar{w} = 1 / \sum_i \hat{k}_i$$

$$\text{var } \hat{a} = 1 / \sum_i \hat{k}_i + \bar{x}^2 \text{var } \hat{b}$$

$$\text{var } \hat{b} = \left\{ \sum_i \hat{k}_i (x_i - \bar{x})^2 \right\}^{-1}.$$

3. Toepassing

- 3.1 Als toepassing bezien we het schattingsprobleem voor een populatie met een standaard logistische verdeling met verdeelingsfunctie $F(x) = [1 + \exp(-x)]^{-1}$ waarvoor de te schatten parameters derhalve $a=0$ en $b=1$ zijn. We kiezen $I=5$ doses x_i , met de waarden $-2, -1, 0, 1$ en 2 . De bij deze doses behorende F -waarden zijn dan $F(-2)=0.119$, $F(-1)=0.269$ en $F(0)=0.500$. Voor $x_1=-2$ en $x_5=2$ zullen we een negatief binomiaal deelexperiment uitvoeren en wel met $s_i=2$. Bij $x_5=2$ tellen we het aantal responsies tot de 2^e non-responsie en wisselen p en g . In de binomiale deelexperimenten bij $x_i=-1, 0$ en 1 is telkens $n_i=10$.

De verkregen waarnemingsuitkomsten staan in tabel 1. Hierin is ook opgenomen de schatting en theoretische waarde van kansen en gewichten; voor $i=1$ en $i=5$ wordt p_i zuiver geschat door $\hat{p}_i = (s_i - 1) / (s_i + \underline{m}_i - 1)$.

Dosis x_i	n_i resp. s_i	y_i resp. m_i	$\hat{p}_i$	p_i	$\hat{g}_i$ resp. $\hat{h}_i$	g_i resp. h_i
$x_1=-2$	$s_1= 2$	$m_1=15$	0.062	0.119	$\hat{h}_1=1.876$	1.762
$x_2=-1$	$n_2=10$	$y_2= 2$	0.200	0.269	$\hat{g}_2=1.600$	1.966
$x_3= 0$	$n_3=10$	$y_3= 5$	0.500	0.500	$\hat{g}_3=2.500$	2.500
$x_4=+1$	$n_4=10$	$y_4= 8$	0.800	0.731	$\hat{g}_4=1.600$	1.966
$x_5=+2$	$s_5= 2$	$m_5= 7$	0.875	0.881	$\hat{h}_5=1.750$	1.762

Tabel 1: Waarnemingsuitkomsten, theoretische en geschatte kansen en gewichten.

De schattingen $\hat{a}$ en $\hat{b}$ en hun geschatte asymptotische standaardafwijkingen vindt men in tabel 2, alsmede de bijbehorende theoretische waarden.

experimenteel				theoretisch			
$\hat{a}$	$\hat{b}$	$\hat{\sigma}(\hat{a})$	$\hat{\sigma}(\hat{b})$	a	b	$\sigma(\hat{a})$	$\sigma(\hat{b})$
-0.148	1.209	0.328	0.238	0	1.000	0.317	0.236

Tabel 2: Parameterschattingen, theoretische en geschatte standaardafwijkingen.

Tenslotte blijkt de modeltoets $\hat{SS}$ de uitkomst $\hat{SS} = 0.452$ te hebben, waarbij 3 vrijheidsgraden horen; het model blijkt redelijk te passen.

4. Discussie en conclusies

- 4.1 Na het voorgaande is een interessante vraagstelling de volgende:

"Bij welke doses x_i voeren we een deelexperiment uit en is dit dan binomiaal of negatief binomiaal?"

We veronderstellen dat de geschatte regressielijn $\hat{a} + \hat{b}x$, in de notatie $\hat{u} + \hat{b}(x - \bar{x})$ nauwkeurig bepaald is als zowel $\text{var } \hat{u}$ als $\text{var } \hat{b}$ klein zijn. Bovengenoemde vraagstelling zullen we splitsen in:

Vraag 1: Wat is de effectiviteit van de negatief binomiale schattingsmethode t.o.v. de binomiale, gemeten in termen van $\text{var } \hat{u}$ en $\text{var } \hat{b}$?

Vraag 2: Voor welke doses x_i kunnen we het beste een deelexperiment uitvoeren, opdat $\text{var } \hat{u}$ en $\text{var } \hat{b}$ zo klein mogelijk worden?

Deze vragen zullen we aan een nader onderzoek onderwerpen.

4.2 Ad vraag 1

We herinneren ons dat

$$\text{var}^{-1} \hat{b} = \sum_i k_i (x_i - \bar{x})^2, \text{ waarin } k_i = g_i = n_i p_i q_i \text{ voor } i \in I_1$$

$$= h_i = s_i q_i \text{ voor } i \in I_2$$

$$\text{en } \bar{x} = \sum_i k_i x_i / \sum_i k_i$$

$$\text{var}^{-1} \hat{u} = \sum_i k_i$$

Als we bij een dosis x_i als conditie stellen dat het vaste aantal subjecten n_i in een binomiaal deelexperiment gelijk is aan het verwachte aantal subjecten in het corresponderende negatief binomiale deelexperiment, met andere woorden dat $n_i = s_i/p_i$, dan is ook $g_i = n_i p_i q_i = s_i q_i = h_i$.

Dit betekent dat de var $\hat{b}$ en var $\bar{u}$ voor beide schattingsmethoden gelijk zijn: de methoden hebben theoretisch dezelfde efficiency. Onder de genoemde conditie is het dan voldoende voor het zoeken naar een antwoord op vraag 2 (dat doen we in 4.3) voorlopig alleen het binomiale geval te beschouwen. Echter nog een praktische opmerking. Bij het schatten van var $\hat{b}$ werken we met geschatte gewichten $\hat{g}_i$ en $\hat{h}_i$ i.p.v. de onbekende g_i en h_i . Nu is:

$\hat{g}_i = n_i \hat{p}_i (1 - \hat{p}_i)$ als $y_i \neq 0$ en $y_i \neq n_i$; de kans hierop is

$$1 - (1 - p_i)^{n_i} - p_i^{n_i}, \text{ terwijl } \hat{p}_i = y_i / n_i$$

$\hat{g}_i = n_i \frac{1}{2n_i} (1 - \frac{1}{2n_i})$ als $y_i = 0$ of $y_i = n_i$; de kans hierop is

$$(1 - p_i)^{n_i} + p_i^{n_i} \text{ (Berkson correctie)}$$

Derhalve geldt, omdat $E(n_i \hat{p}_i (1 - \hat{p}_i)) = (n_i - 1)p_i (1 - p_i)$

$$\begin{aligned} E \hat{g}_i &= \{(1 - p_i)^{n_i} + p_i^{n_i}\} n_i \frac{1}{2n_i} (1 - \frac{1}{2n_i}) + \\ &+ \{1 - (1 - p_i)^{n_i} - p_i^{n_i}\} \frac{(n_i - 1) p_i (1 - p_i)}{\{1 - (1 - p_i)^{n_i} - p_i^{n_i}\}} = \\ &= \{(1 - p_i)^{n_i} + p_i^{n_i}\} n_i \frac{1}{2n_i} (1 - \frac{1}{2n_i}) + (n_i - 1)p_i (1 - p_i). \end{aligned}$$

Ook is:

$$E \hat{h}_i = s_i (1 - p_i) = n_i p_i (1 - p_i) \text{ als } s_i = n_i p_i.$$

Enkele waarden van $E \hat{g}_i$ zijn gegeven in tabel 3.

		$E \hat{g}_i$	$g_i = n_i p_i (1-p_i)$
$n_i = 10$	$p_i = 0.1$	0.976	0.900
	0.2	1.491	1.600
	0.5	2.251	2.500
$n_i = 20$	$p_i = 0.1$	1.769	1.800
	0.2	3.046	3.200
	0.5	4.750	5.000

Tabel 3: Waarden van $E \hat{g}_i$ en g_i voor enkele n_i en p_i .

Aangezien $E \hat{h}_i = n_i p_i (1-p_i)$, is $\hat{h}_i$ een zuivere schatter van h_i . Daarmee is dan bijvoorbeeld $\text{var } \hat{u}$ een overschatting van $\text{var } \underline{u}$, want met de ongelijkheid van Jensen geldt:

$$E \text{var } \hat{u} = E \{ \sum_i \hat{h}_i \}^{-1} > \{ \sum_i E \hat{h}_i \}^{-1} = \{ \sum_i h_i \}^{-1} = \text{var } \underline{u}$$

Voor $\hat{g}_i$ komt het soms voor (zie tabel 3), dat $\hat{g}_i$ de waarde g_i overschat, waardoor de overschatting van $\text{var } \underline{u}$ gereduceerd wordt. Vaak is echter $\hat{g}_i$ een onderschatting van g_i ; het gevolg is een nog grotere overschatting van $\text{var } \underline{u}$ in het binomiale experiment met Berkson correctie dan in het negatief binomiaal deelexperiment. Ook is het helemaal weglaten van een waarneming $y_i=0$ of $y_i=n_i$ niet aan te bevelen: dan is $\hat{g}_i$ met $E \hat{g}_i = (n_i-1)p_i(1-p_i)$ altijd een onderschatting van g_i .

4.3 Ad vraag 2

In hoofdstuk 2 is gesuggereerd dat we, als we een nauwkeurige schatting van parameter b wensen, wel één of meer doses x_i in de staart van de logistische verdeling zullen moeten kiezen. We zullen nu eens nagaan voor welke doses x_i de $\text{var } \hat{b}$ (en later ook de $\text{var } \underline{u}$) klein wordt.

We veronderstellen daarbij dat de parameters a en b , voordat het experiment wordt uitgevoerd, al ruwweg bekend zijn: stel dat $a \approx a_0$ en $b \approx b_0$, en kies voor het rekenwerk $a_0=0$ en $b_0=1$. Is dan voor deze a_0 en b_0 een dosis x_i vastgesteld, dan volgt voor willekeurige a_0 en b_0 een corresponderende dosis x_i^0 uit x_i via $x_i^0 = b_0^{-1} (x_i - a_0)$. Gezien de symmetrie in de probleemstelling zullen we doses symmetrisch t.o.v. $x=0$ kiezen.

Dan is $\bar{x} = \sum g_i x_i / \sum g_i = 0$, $\hat{a} = \bar{u}$ en $\text{var}^{-1} \hat{b} = \sum g_i (x_i - \bar{x})^2 =$

$$= \sum_i n_i p_i (1-p_i) x_i^2 = \sum_i n_i \frac{e^{-x_i}}{(1+e^{-x_i})^2} x_i^2.$$

Na berekening blijkt dat $\frac{e^{-x_i}}{(1+e^{-x_i})^2} x_i^2$, de bijdrage per sub-

ject aan $\text{var}^{-1} \hat{b}$ bij dosis x_i , maximaal is voor $x_i = \pm 2.399$. Het is derhalve zinvol rond deze doses x_i waarnemingen te verrichten en omdat bij $x_i = -2.399$, $F(-2.399) = 0.083$ de succeskans is, is een negatief binomiaal experiment op z'n plaats.

Zouden we alléén waarnemingen verrichten bij $x_i = \pm 2.399$, met (verwacht) aantal N , dan is

$$\text{var}^{-1} \hat{b} = 2 \left\{ (2.399)^2 * \frac{N}{2} * (0.083) * (0.917) \right\} = 0.438N$$

zodat

$$\text{var} \hat{b} = \frac{2.283}{N} \text{ een ondergrens is voor } \text{var} \hat{b}.$$

Het bovengenoemd maximum is overigens niet zeer kritisch, zoals moge blijken uit onderstaande tabel 4.

Hierin zijn voor enkele doses x_i waarden uitgezet van

$$\frac{e^{-x_i} x_i^2}{(1+e^{-x_i})^2} \text{ in kolom 4.}$$

x_i	p_i	$\frac{e^{-x_i}}{(1+e^{-x_i})^2}$	$\frac{x_i^2 e^{-x_i}}{(1+e^{-x_i})^2}$
0	0.5	0.250	0
- 0.5	0.378	0.236	0.059
- 1.0	0.269	0.197	0.197
- 1.5	0.182	0.149	0.336
- 2.0	0.119	0.100	0.420
- 2.399	0.083	0.076	0.439
- 3.0	0.048	0.045	0.406

Tabel 4: Kans op responsie p_i , bijdrage per subject aan $\text{var } \bar{u}$ en $\text{var } \hat{b}$ voor verschillende waarden van dosis x_i .

In de toepassing van hoofdstuk 3 vonden we $\hat{\text{var } \bar{b}} = 0.057$ bij $N = 64$. De ondergrens voor $\text{var } \bar{b}$ bij deze N is 0.036.

Concurrerend met een kleine $\text{var } \hat{b}$ is een kleine $\text{var } \bar{u}$. Aangezien

we voor $\text{var}^{-1} \bar{u}$ kunnen schrijven $\text{var}^{-1} \bar{u} = \sum_i g_i = \sum_i n_i \frac{e^{-x_i}}{(1+e^{-x_i})^2}$

zoeken we nu het maximum van $\frac{e^{-x_i}}{(1+e^{-x_i})^2}$, hetgeen ligt bij $x_i = 0$.

Waarden van deze grootheid voor enkele x_i 's staan in tabel 4. Worden alleen waarnemingen bij $x_i=0$ verricht (N stuks) dan is $\text{var } \bar{u} = \frac{4}{N}$ een ondergrens voor $\text{var } \bar{u}$.

In de toepassing was $\hat{\text{var } \bar{u}} = 0.107$ bij een ondergrens van $\text{var } \bar{u} = 0.062$.

4.4 Concluderend kunnen we stellen:

- 4.4.1 De efficiency van een negatief binomiaal deelexperiment is (theoretisch) die van het corresponderende binomiale experiment, als verwacht resp. gefixeerd aantal subjecten gelijk zijn. Praktisch geeft een binomiaal deelexperiment met Berkson correctie vaak een grotere overschatting van $\text{var } \bar{u}$ en $\text{var } \hat{b}$ dan een negatief binomiaal deelexperiment.
- 4.4.2 Voor een kleine $\text{var } \hat{b}$ is het nuttig waarnemingen rond $x_i = b_0^{-1} (\pm 2.40 - a_0)$ te verrichten, als a_0 en b_0 ruwweg de parameters van de regressielijn voorstellen. Gezien de kleine of juist grote kans op responsie is een negatief binomiaal deelexperiment gediensig.
- 4.4.3 Voor een kleine $\text{var } \bar{u}$ zijn waarnemingen rond $x_i = -a_0 b_0^{-1}$ zinvol; hier is een binomiaal deelexperiment, waarbij het aantal subjecten vast wordt gekozen, goed uit te voeren.

Nawoord

De auteur is de redactie dankbaarheid verschuldigd voor enkele waardevolle opmerkingen.

Aanhangsel 1

De limietstelling van Lindeberg (zie bijvoorbeeld Loeve (1963), p. 295) zegt dat voor een negatief binomiaal verdeelde grootheid $\underline{m}$ de conditie

$$\lim_{s \rightarrow \infty} \left\{ \sum_{i=1}^s \sum_{|\underline{m}| > \varepsilon} m^2 P(\underline{m}^{(i)*} = m) \right\} = 0 \text{ voor } \forall \varepsilon > 0$$

equivalent is met asymptotische normaliteit van $\underline{m}$.

In deze conditie is $\underline{m} = \underline{m}^{(1)} + \underline{m}^{(2)} + \dots + \underline{m}^{(s)}$, waarin $\underline{m}^{(i)}$

o.o. en geometrisch verdeeld, terwijl $\underline{m}^{(i)*} = (\underline{m}^{(i)} - q/p) / \sqrt{(sq/p^2)}$,

zodat $E \underline{m}^{(i)*} = 0$ en $\text{var } \underline{m}^{(i)*} = 1/s$ voor $i=1, 2, \dots, s$.

Genoemde som schrijven we als

$$s \sum_j \left\{ \frac{j-q/p}{\sqrt{(sq/p^2)}} \right\}^2 p q^j \dots\dots\dots (*)$$

waarbij we $m = \frac{j-q/p}{\sqrt{(sq/p^2)}}$ gesteld hebben en we alleen sommeren over

die termen, waarvoor $j > 0$ en $|m| = \left| \frac{j-q/p}{\sqrt{(sq/p^2)}} \right| = \left| \frac{p(j+1)-1}{\sqrt{(sq)}} \right| > \varepsilon$ is.

We laten zien dat de Lindeberg conditie equivalent is met de voorwaarde $sq \rightarrow \infty$.

- i) Stel dat $sq \rightarrow \infty$. Dan geldt voor $\forall \varepsilon > 0$ en voor $\forall J \in \mathbb{N}$ dat $\exists s_0(\varepsilon, J)$ zodat $\left| \frac{p(j+1)-1}{\sqrt{(sq)}} \right| < \varepsilon$ voor $j=1, 2, \dots, J$ en $s > s_0(\varepsilon, J)$.

We sommeren in (*) dan hoogstens over $j=J+1, J+2, \dots$. Door J voldoende groot te kiezen wordt de som (*) willekeurig klein.

- ii) Stel dat $sq \not\rightarrow \infty$. Dan $\exists M > 0$ en een oneindige deelrij met $sq < M$. Voor deze deelrij geldt $q \rightarrow 0$ en dus $p \rightarrow 1$. Noem de indexverzameling van deze deelrij S .

Kies een q_0 , zo dat $0 < q_0 < 1$, en laat s_0 bepaald zijn door dat $q < q_0$ voor $s > s_0$, $s \in S$.

Dan is ook $p > 1 - q_0 > 0$; noem $p_0 = 1 - q_0$.

Kies nu j_0 , zo dat $p_0(j+1) - 1 > 1$ voor $j > j_0$.

Voor $s > s_0$, $s \in S$ en $j > j_0$ is dan zeker

$$|p(j+1)-1| > 1 \text{ en dus } \left| \frac{p(j+1)-1}{\sqrt{(sq)}} \right| > \frac{1}{\sqrt{M}} > 0.$$

Kies $\epsilon = \frac{1}{\sqrt{M}}$. Voor $\forall s_1 > s_0 \exists s > s_1$, $s \in S$ waarvoor in de som (*) bij deze ϵ tenminste gesommeerd wordt over

$j=j_0, j_0+1, \dots$. Dit levert voor deze som een waarde $> \delta_0$, voor zekere $\delta_0 > 0$. Er is voor som (*) derhalve geen limiet nul bij deze ϵ voor $s \rightarrow \infty$.

Met (i) en (ii) is de equivalentie bewezen.

Aanhangsel 2

Voor de bepaling van het negatief binomiale gewicht $h = \text{var}^{-1} \underline{v}$, optredend in de kwadraatsom (2.5), zoeken we een benadering voor de variantie van $\underline{v} = F^{-1}(\frac{1}{1+\underline{m}/s})$. Daartoe ontwikkelen we de func-

tie $F^{-1}(\frac{1}{1+x})$ in een stukje Taylorreeks t.o.v. $x = q/p$:

$$F^{-1}(\frac{1}{1+x}) = F^{-1}(\frac{1}{1+q/p}) + (x - q/p) \frac{d}{dx} F^{-1}(\frac{1}{1+x}) \Big|_{x=q/p} + \frac{(x - q/p)^2}{2} \frac{d^2}{dx^2} F^{-1}(\frac{1}{1+x}) \Big|_{x=y}$$

waarin y ligt tussen x en q/p .

Is $F(x)$ de logistische verdelingsfunctie $F(x) = \{1+\exp(-x)\}^{-1}$, dan komt er

$$F^{-1}(\frac{1}{1+x}) = -\log x = -\log q/p + (x-q/p)(-p/q) + \frac{(x-q/p)^2}{2} (1/y^2) \dots\dots\dots (A.1)$$

We negeren de derde term, hetgeen zinvol is voor x in de buurt van q/p , vullen voor x de schatter $\underline{m}/s$ van q/p in en benaderen de variantie van $\underline{v} = F^{-1}(\frac{1}{1+\underline{m}/s})$ door

$$\begin{aligned} \text{var } F^{-1}(\frac{1}{1+\underline{m}/s}) &\approx \text{var } (\underline{m}/s) \left\{ \frac{d}{dx} F^{-1}(\frac{1}{1+x}) \right\}_{x=q/p}^2 = \\ &= q/(sp^2) \{p^2/q^2\} = 1/sq \end{aligned} \dots\dots (A.2)$$

Gewicht h wordt dan benaderd door $h = sq$.

De procedure is analoog aan een in bijvoorbeeld Corsten (1976) genoemde, waarin voor gewicht g de benadering $g \approx np(1-p)$ wordt afgeleid.

Aanhangsel 3

Uitgaande van de zwakke convergentie

$$\sqrt{s}(\underline{m}/s - q/p) \rightarrow N(0, q/p^2) \text{ voor } s \rightarrow \infty \quad \dots\dots\dots(A.3)$$

voor negatief binomiaal (s, p) verdeelde $\underline{m}$ kan men laten zien dat voor $\underline{v} = F^{-1}(\frac{1}{1+\underline{m}/s})$, waarin F de logistische verdelingsfunctie is, geldt

$$\sqrt{s}(\underline{v} - F^{-1}(\frac{1}{1+q/p})) \rightarrow N(0, q^{-1}) \text{ voor } s \rightarrow \infty.$$

Gebruik hiertoe bijvoorbeeld Theorem 14.6-1 uit Bishop c.s. (1975). De limieteigenschap (A.3), het feit dat de asymptotische variantie van $\sqrt{s} \underline{v}$ gelijk is aan q^{-1} volgens (A.2) en de uit (A.1) volgende eigenschap

$$F^{-1}(\frac{1}{1+x}) = F^{-1}(\frac{1}{1+q/p}) + (x-q/p)(-p/q) + o(|x-q/p|)$$

voor $x \rightarrow q/p$, die geldig is voor iedere positieve q/p , leveren met elkaar het bewijs.

Literatuurverwijzingen

- Berkson, J. (1955),
Maximum likelihood and minimum χ^2 estimates of the logistic
function,
J. Amer. Statist. Ass., 50, 130 - 162.

- Bishop, Y.M.M.; Fienberg, S.E. and Holland, P.W. (1975),
Discrete multivariate analysis,
The MIT - Press, Cambridge.

- Corsten, L.C.A. (1976),
Handleiding kwantielanalyse,
Landbouwhogeschool Wageningen.

- Finney, D.J. (1971),
Probit analysis,
Cambridge University Press.

- Johnson, L. and Kotz, S. (1969),
Distributions in statistics I: discrete distributions,
John Wiley & Sons, New York.

- Loève, M. (1963),
Probability Theory,
Van Nostrand, New York.

- Rümke, Chr.L. en Kuik, D.J. (1981),
De omgekeerde bepalingsmethode voor een percentage of een
kans,
Ned. Tijdschr. v. Geneesk., 125^e jaargang, 2044 - 2050.

Ontvangen: 17-8-1982