Postkarten aus Seoul

A.J. BACZKOWSKY

ÜBERSETZT UND BEARBEITET VON KARL RÖTTEL, EICHSTÄTT

Wish you were here: Postcards from Seoul. Teaching Statistics Volume 25 (Summer 2003)Number 2, p.46-48.

Zusammenfassung: *Was 73 Postkarten, die aus Seoul zwischen dem 23. August und dem 1. September 2001 an eine Adresse in England geschickt wurden, über den Postdienst kundtun.*

Einleitung

Von Donnerstag, 22. 8. 2001, bis Samstag, 1. 9. 2001, verschickte ich während einer Konferenz über Statistik 73 Postkarten aus Seoul an meine Heimatadresse in England. Tabelle 1 zeigt die Tage bis zur Ankunft der Postkarten.

Tag des Einwurfs		Insgesamt	Tage bis zur Ankunft										…			
			5	6	7	8	9	10	11	12	13	14		17	18	19
Donnerst..	23.8.	12				11	1									
Freitag	24.8.	6				5		1*								
Samstag	25.8	5		2				1								2
Sonntag	26.8	7	4											2		1
Montag	27.8.	6			4*				1						1	
Dienstag	28.8.	4					4									
Mittwoch	29.8.	6		1		4	1									
Donnerst.	**30.8.	5					4					1				
Freitag	31.8.	14				11					3					
Samstag	1.9.	8		1			1*		2	1	2	1				

Tabelle 1 Tage bis zur Ankunft von 73 Postkarten

** Am Montag angekommen (siehe „Diskussion"), ** Poststempel vom 31.8.*

Die meisten Postkarten waren dem Hotel täglich nach dem Frühstück zum Einwerfen beim Postamt gegeben worden. Und es wurde davon ausgegangen, dass man die Karten täglich zur gleichen Zeit zur Post brachte. Sie wurden mit „SL.Kangnam" gestempelt und im allgemeinen am Tag des Einwurfs von der Post gestempelt. Die fünf am Donnerstag, dem 30., eingeworfenen Karten wurden mit 31. gestempelt. Drei mit Stempel vom Freitag, dem 31., wurden vom Postamt Kwanghwamun versandt und mit diesem Datum gestempelt.

Dieser Aufsatz soll einige einfache Modelle zur Analyse der Daten von Tabelle 1 aufzeigen.

Consignia (pers. Mitt., 2002) weigerte sich, mehr Informationen darüber preiszugeben, wie die Post zwischen den zwei Ländern befördert wird – „aus Sicherheitsgründen". Um die Daten in Modellen verarbeiten zu können, muß man verschiedene vereinfachende Annahmen machen. Zum Beispiel wurden die am 30. 8. abgegebenen Postkarten so behandelt, als wären sie vom 31. 8., weil das Personal die Briefe am 30. 8. anscheinend nicht am gleichen Tag zur Post brachte. Die Gleichartigkeit der Ankunft der beiden Freitags- und Samstagsbriefe legte auch die Annahme nahe, dass Postkarten des gleichen Wochentages gleich behandelt werden, weshalb sie jeweils zusammengefasst werden.

Ein Modell für die Beförderung

Die kleinste Anzahl bis zur Ankunft betrug fünf Tage. Wie können diese Beförderungstage zustande kommen? Die Postkarte muss vom Hotel zur Post und von dort zum Flughafen gelangen, wo sie auf den Flug wartet. Es ist wenigstens ein Tag für den Flug nach London zu veranschlagen. Nach Ankunft in London muss sie zur zentralen Sortierstelle und von da zu meinem Heimatpostamt gebracht werden. Der Postbote wird sie am nächsten Tag zustellen.

Wie kann es zu Verzögerungen kommen? Wenn im Postamt zu viel zu tun ist, wird die Post oft zurückgehalten, bis Zeit für das Versenden verfügbar wird. Es ist auch denkbar, dass Postkarten eine geringere Priorität haben, wenn eine Menge anderer Post vorhanden ist.

Nehmen wir an, eine Postkarte werde im Postamt in Seoul zurückgehalten und es wird gewartet bis Platz in einem Transportflugzeug am Incheon-Flughafen (internationaler Flughafen von Seoul) ist. Ein einfaches Modell ist, dass die Wahrscheinlichkeit für das unverzügliche Weiterleiten der Postkarten p, für das Zurückhalten bis zum nächsten Tag 1 – p beträgt. Folglich wird eine Postkarte mit der Wahrscheinlichkeit p nicht verzögert. Eine Postkarte wird um einen Tag verzögert, wenn sie am ersten Tag (mit der Wahrscheinlichkeit 1-p) zurückgehalten und am nächsten Tag mit der Wahrscheinlichkeit p weitergeleitet wurde. Unter der Voraussetzung der Unabhängigkeit ist die Wahrscheinlichkeit, dass sie um einen Tag verzögert wird, $(1-p)\cdot p$.

Wenn P_x die Wahrscheinlichkeit bezeichnet, dass eine Postkarte um x Tage verzögert ist, ergeben entsprechende Schlußfolgerungen $P_x = p(1-p)^x$, x = 0,1,2,…, also die bekannte geometrische Verteilung.

Tabelle 1 zeigt, dass die kleinste Zahl der Tage bis zur Ankunft für das Absenden am Sonntag fünf, am Samstag sechs, usw. ist. Die Unterschiede könnten mit den Startzeiten der Flugzeuge zusammenhängen, die die Post nach England befördern. So könnte man aus den Daten für Samstag und Sonntag schließen, dass nur ein einziger Flug am Sonntag abends die Post befördert. Zieht man diese Minimalzahl für jeden Tag, an dem Post aufgegeben wurde, ab, erhält man die Anzahl x der Tage der Verzögerung jeder Postkarte. Tabelle 2 gibt die Häufigkeit an, die für jeden Wert x festgestellt wurde.

	Tage der Verzögerung														
Tag des Einwurfs	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Sonntag	4												2		1
Montag	4				1							1			
Dienstag	4														
Mittwoch	1		4	1											
Donners.	11	1													
Freitag	20		1			4									
Samstag	3			1	1	2	1	2	1					2	

Tabelle 2: Anzahl der Tage, um die 73 Postkarten verzögert wurden.

Wie kann nun p geschätzt werden? Für diese geometrische Verteilung ist der Erwartungswert (1-p)/p. Setzt man dies dem Stichprobenmittelwert x´ gleich, erhält man den Schätzwert $p' = 1/(1+x')$. Dies ist ein spezieller Fall der sogenannten Momentenmethode (s. Kreyszig, S.167).

Ein anderer Weg, p zu schätzen, ist die Methode der „Maximum-Likelihood-Schätzung“ (siehe z. B. Henze, Kapitel: Schätzprobleme). Wenn die n beobachteten Verzögerungen mit x_1, x_2, … bezeichnet und als unabhängig angesehen werden, ist die Wahrscheinlichkeit oder Likelihood für diese n auftretenden Werte:

$$L = \{p(1-p)^x 1\}\{p(1-p)^x 2\}\ldots\{p(1-p)^x n\} =$$
$$= pn(1-p)^{\Sigma xi}.$$

Der wahrscheinlichste Wert für p ist derjenige, für den die Funktion L maximal wird. Mit Logarithmen lautet die Formel

$$\ln L = n \ln p + n\, x' \ln (1-p).$$

Der Wert von p, für den L ein Maximum annimmt, ist derselbe, für den ln L maximal wird.

Deshalb ist der geschätzte Wert von p gleich $p^* = 1/(1+x')$, hier derselbe wie jener bei der Methode der Momentschätzung, obwohl dies im allgemeinen nicht der Fall ist.

Die Varianz dieses Maximum-likelihood-Schätzwertes kann mit dem erhaltenen Ergebnis gefunden werden:

$$\mathrm{Var}\,[p'] = -E\left[\left\{\frac{d^2 \ln L}{dp^2}\right\}^{-1}\right].$$

Alternativ kann die Näherungsformel

$$\mathrm{Var}\left[\frac{1}{Y}\right] = \frac{Var[Y]}{\{E[Y]\}^4}$$

verwendet werden.
Setzt man $Y = 1 + X'$ mit dem Erwartungs $1/p$ und der Varianz $(1-p)/(np^2)$, erhält man

$$\mathrm{Var}[p^*] = \frac{p^2(1-p)}{n}.$$

Dies kann geschätzt werden, wenn man p durch p' ersetzt.
Die Daten liefern $p^* = 0{,}316$ mit Standardabweichung STD(p*) = 0,0306. Ein näherungsweises 95%-Konfidenzintervall für p ist $p^* \pm 1{,}96$ STD(p^*) = (0,256 , 0,376).

Ein zweites Modell

Das einführende Modell ließ darauf schließen, dass für eine Postkarte, die Wahrscheinlichkeit an einem Tag liegen zu bleiben 0,684 beträgt.

In einem verfeinerten Modell möge angenommen werden, dass die Wahrscheinlichkeit für eine am Sonntag abgegebene Postkarte, die am gleichen Tag am Postamt ankommt, verzögert zu werden $1\text{-}p_1$ beträgt, während die Verzögerung für die am Montag abgegebenen Postkarten $1–p_2$ beträgt, usw.

Im vorangegangenen Modell waren alle diese Wahrscheinlichkeiten gleich, nämlich $1\text{-}p$.
Unter den am Sonntag abgegebenen Postkarten hatten vier keine Verzögerung, zwei hatten eine von 12 Tagen und eine von 14 Tagen. Die Wahrscheinlichkeit dieses Zustandes ist

$$p_1^4 \times \{(1-p_1)(1-p_2)\cdots(1-p_{11})(1-p_{12})p_{13}\}^2$$
$$\text{x } (1-p_1)(1-p_2)\cdots(1-p_{13})(1-p_{14})\,p_{15}.$$

Behandelt man die Sonntage gleich ($p_1 = p_8$) usw., dann ergibt das Obige vereinfacht:

$$p_1^5 p_6^2 (1-p_1)^6(1-p_2)^6(1-p_3)^6(1-p_4)^6 \text{ x}$$
$$\text{x } (1-p_5)^6(1-p_6)^4(1-p_7)^4.$$

Wiederholt man dies für alle Daten in Tabelle 2, erhalten wir die Gesamtwahrscheinlichkeit

$$L = p_1^7 p_2^4 p_3^5 p_4^6 p_5^{13} p_6^{32} p_7^6 (1-p_1)^{23}(1-p_2)^{25}$$
$$\text{x } (1-p_3)^{24}(1-p_4)^{24}(1-p_5)^{23}(1-p_6)^{16}(1-p_7)^{23}.$$

Ein Weg, die Werte $p_1, p_2, \ldots, p_8$ zu bestimmen, die diese Likelihood-Funktion maximieren, ist die Verwendung eines Computers, um Unmengen an Schätzungen von p_i-Werten zu generieren und so die Kombination zu finden, die L (oder ln L) maximiert. Ein einfacher Algorithmus, der dies bewerkstelligt, ist der folgende:

> Generiere $P_1, P_2, \ldots, P_7$ als gleiche Schätzvariablen zwischen 0 und 1.
>
> Bilde ihre Summe S.
>
> Setze $p_i = p_i / S$, wodurch sichergestellt ist, dass die Schätzwahrscheinlichkeiten zwischen 0 und 1 liegen und die Summe 1 ist.
>
> Berechne ln L und prüfe, ob dies größer als das vorangegangene Maximum ist.

Wenn man dies mit einer Vielzahl von p_i-Werten versucht, hat man eine einfache Methode, die Werte für maximales ln L zu finden. Tabelle 3 gibt diese geschätzten Wahrscheinlichkeiten zusammen mit ihren Standardfehlern wieder, die nach einer dem vorangehend geschilderten Verfahren ähnlichen Weise gefunden wurden.

	Tag des Einwurfs						
	So	Mo	Di	Mi	Do	Fr	Sa
p_i*	0,098	0,055	0,070	0,083	0,176	0,432	0,085
STD(p_i*)	0,038	0,028	0,032	0,035	0,051	0,091	0,035

Tabelle 3: Geschätzte Wahrscheinlichkeiten p_i und angenäherte Standardfehler für jeden Wochentag.*

Die Durchsicht der Tabelle 3 zeigt, dass die Wahrscheinlichkeit, dass eine Postkarte am gleichen Tag befördert wird, ziemlich klein ist, kleiner als 0,1. Nur am Ende der Woche hat eine Postkarte gute Chancen, sofort befördert zu werden.

In Tabelle 1 sieht man, dass 6 Postkarten sehr große Verzögerungen haben. Consignia (pers. Mitt., 2002) vermutet, dass diese wohl nicht als Luftpost befördert wurden. Das Streichen jener sechs Werte ändert das Ergebnis geringfügig. So erhielte man im ersten Modell $p^* = 0{,}447$ mit Standardabweichung STD(p^*) = 0,0406.

Im Originaltext (URL-Adresse s. Literatur) werden noch zwei weitere Modelle diskutiert. Eines der beiden geht davon aus, dass die Bearbeitung jeder Postkarte an zwei verschiedenen Orten verzögert werden kann und diese unabhängig voneinander sind. Ein viertes Modell schließlich geht davon aus, dass beim zweiten Modell (S.10) die Tage 1,2,,3,4,7 gleiche Wahrscheinlichkeit p_1 haben, dass die Postkarten unverzüglich weitergeleitet werden, die Wahrscheinlichkeit dagegen am Tag 5 und 6 jeweils p_2 bzw. p_3 betragen.

Diskussion

Tabelle 1 zeigt auch sechs Postkarten, die an einem Montag ankamen. Da es im Bestimmungsort keine Zustellung am Sonntag gibt, könnte dies im Modell als einen Tag zuvor angekommen berücksichtigt werden. Dem wurde hier nicht weiter nachgegangen

Das vorgestellte Modell enthält mehrere Voraussetzungen. Einige sind gültig, wie z.B. die, dass an Sonntagen keine Post ausgetragen wird, während andere, wie die Unabhängigkeit, einfach bequeme Voraussetzungen sind, um beim Modellbilden voranzukommen. Die scheinbar großen Wahrscheinlichkeiten für die Verzögerungen lassen vermuten, dass die zugrundegelegte Annahme einer geometrischen Verteilung nicht passend war, vielleicht auch einschließend, dass Postkarten eben nicht nach Zufall fürs Befördern an jedem Tag gewählt werden. Mehr Informationen über den Postdienst und mehr Daten würden erlauben, verfeinerte Modelle zu entwickeln.

Literatur

Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen. Göttingen: Vandenhoeck&Ruprecht

Henze, Norbert: Stochastik für Einsteiger. Braunschweig: Vieweg

Originaltext in ausführlicher Form:
http://www.maths.leeds.ac.uk/~sta6ajb/drep0111.pdf

Autor:

A.J. Baczkowski
Department of Statistics
University of Leeds
Leeds LS2 9JT
UK