$$
\begin{aligned}
n &\leftarrow n + 1 \\
d &\leftarrow (x - \bar{x})/n \\
\bar{x} &\leftarrow \bar{x} + d \\
S &\leftarrow S + n(n-1)d^2
\end{aligned}
$$

Ein Beispiel

Daß dieser Punkt von praktischer Bedeutung und nicht nur von akademischem Interesse ist, zeigt ein Beispiel aus einem A-level-Beitrag zur Statistik von 1978, wo eine Regression bezüglich der Zeit von 1963 bis 1972 benötigt wurde. Hier ist $\Sigma(x-\bar{x})^2 = 82{,}5$, aber $\Sigma x^2-(\Sigma x)^2/n = 38{\cdot}710{\cdot}645-38{\cdot}710{\cdot}562{,}5$, wobei der zweite Term auf einem achtstelligen Rechner nicht voll darstellbar ist, so daß das Ergebnis nur auf zwei Stellen genau ist.

Das Problem

Dieser Teil hat mit Absicht die gleiche Überschrift wie der erste, um den Teufelskreis bei diesem Problem zu zeigen. Mathematisch gesehen, gibt es das Problem nicht mehr; denn gute Lösungen sind seit Jahren bekannt. Das Problem liegt darin, daß zu viele Leute sich nicht dessen bewußt sind, daß hier überhaupt ein Problem vorliegt, und deshalb weiterhin die gefährliche Methode lehren. Das Argument, daß dies keinerlei Aufwand lohne, da die meisten Datenserien unproblematisch sind, kommt mir vor, als lehnte jemand es ab, an einem Straßenübergang stehenzubleiben und auszuschauen, weil ja noch nie ein Wagen gekommen sei.
Ob ein Artikel wie dieser irgendeine Wirkung haben wird? Wie kann man seinen Inhalt bekannt machen?

Ich danke dem Herausgeber für das Beispiel in vorletzter Minute.

Quelle

Wedford, B.P.: Note an a method for calculating sums of squares and products. Technometrics 4 (1962), S. 419-420

Maschinenausfälle und Geburtstage

von A.F.BISSELL
Übersetzt von G.König

Der Direktor einer Fabrik, in der etwa 100 Maschinen des gleichen Typs in Gebrauch sind, konfrontiert seine Statistikabteilung mit dem Betriebsprotokoll der Instandhaltung. Darin waren die Ausfallzeiten, dazugehörige Maschinennummer und verschiedene andere Informationen aufgezeichnet. Ein typisches Protokoll könnte so aussehen (mit erfundenen Daten):

Datum	Uhrzeit	Schicht	Maschinen-nummer	Defekt	Reparatur beendet	ausgeführt von
1.Jan.	1:35	B	28	Nocker	2:20	J.Smith
1.Jan.	5:20	B	7	Dichtungsring	5:30	D.Jones
1.Jan.	8:15	C	72	festgefressene Lager	10:40	P.Reilly
1.Jan.	8:55	C	57	Thermostat	9:10	I.McGregor
(Zwölf andere Einträge)						
2.Jan.	14:25	A	72	Antriebsriemen	14:45	W.Sykes

Der Direktor legte sein Problem dar: Jede Woche gibt es Maschinen, die ein 2. (oder sogar 3.) Mal infolge eines Schadens ausfallen oder die Reparaturzeiten beanspruchen, während die Mehrzahl der Maschinen keine Defekte hat. Betrachten wir z.B. diese Woche (s.obigen Auszug): Wir hatten 16 Ausfälle und dann mußte die Maschine mit der Nummer 72 innerhalb von 2 Tagen ein 2. Mal repariert werden. Bitte unterstützen Sie mich beim Beweis, daß durch die schlechte Wartung der Maschinen ihre Funktionstüchtigkeit beeinträchtigt wird.
Kann es also allgemein vorkommen, daß wiederholte Defekte einer Maschine früher auftreten können als man es von der Ausfallwahrscheinlichkeit erwarten könnte (unabhängig von Wartung und Instandsetzung)?
Betrachten wir uns den ersten Eintrag eines Defektes. Jede der N Maschinen (in diesem Fall sei N = 100) könnte hier verzeichnet sein. Der zweite Monteurruf kann wieder jeder der N Maschinen gelten, wenn die Fehlerhaftigkeit bzw. Brauchbarkeit einer Maschine unabhängige Ereignisse sind. Wir machen also die Annahme:
Die Ereignisse {Maschine i ist defekt}, (i=1,2,...,N) sind unabhängig. Die Wahrscheinlichkeit dafür, daß die zweite Maschine, die infolge eines Schadens ausfällt, von der ersten verschieden ist, ist (N-1)/N (den zweiten Defekt kann nach der Annahme jede Maschine außer einer haben). Genauso ergibt sich die Wahrschein-

lichkeit, daß die dritte defekte Maschine von den beiden vor ihr reparierten Maschinen verschieden ist, zu (N-2)/N. Die Wahrscheinlichkeit, daß die drei ersten Maschinenausfälle verschiedene Maschinen betreffen, berechnet sich zu $\frac{N-1}{N} \times \frac{N-2}{N}$. Der weitere Ansatz ist offensichtlich. Für die Wahrscheinlichkeit, daß die ersten R Ausfälle R verschiedene Maschinen betreffen, gilt nach der Produktregel für unabhängige Ereignisse:

$$Q_R = \frac{(N-1)}{N} \frac{(N-2)}{N} \cdots \frac{(N-R+1)}{N} = \frac{1}{N^{R-1}} \frac{(N-1)!}{(N-R)!}$$

Die Wahrscheinlichkeit des uns interessierenden Ereignisses - zum Zeitpunkt des R-ten Eintrages hatte mindestens eine Maschine mehrere Ausfälle - ist die Wahrscheinlichkeit des dazu komplementären Ereignisses. Also ist die gesuchte Wahrscheinlichkeit:

$$P_R = 1 - \frac{1}{N^{R-1}} \frac{(N-1)!}{(N-R)!}$$

Rechnet man $1-P_R$ sowie P_R für verschiedene R aus, so erhält man die Zahlen der folgenden Tabelle (N=100):

Eintrag	Wahrscheinlichkeit, daß alle Einträge verschiedenen Maschinen gelten	Wahrscheinlichkeit, daß mindestens eine Maschine mehrfache Einträge hat
1	(1)	(0)keine Wiederholung möglich
2	0.99	0.01
3	0.9702	0.0298
4	0.941094	0.058906
11	0.565341	0.434659
12	0.503153	0.496847
13	0.442775	0.557225
14	0.385214	0.614786

Aus dieser Tabelle ist ersichtlich, daß es unter der Annahme, die Maschinenausfälle seien zufällig, nicht erstaunlich ist, daß Wiederholungen sehr früh im Wartungs- bzw. Reparaturbuch auftreten können. Der Median liegt nämlich bei 12 und bei Fortsetzung der Rechnung würde sich zeigen, daß bei 22 Eintragungen die Wahrscheinlichkeit 90 % betrüge und bei 30 Eintragungen sogar 99 %; eine fast sichere Chance der Wiederholung (mit Wahrscheinlichkeit 0,999) ergäbe sich bei Eintrag Nr. 36.

Mit diesen Zahlen haben wir eben einige Quantile [1] einer Verteilungsfunktion bestimmt.

Die Zufallsvariable ist die Nummer des Eintrags, bei der die erste Wiederholung stattfindet. Ihr Wertebereich besteht aus den natürlichen Zahlen von 2bis (N+1), weil keine Wiederholung beim ersten Eintrag stattfinden kann und eine Wiederholung beim (N+1)-ten Eintrag stattfinden muß.

Der obige Ausdruck für Q_R stellt eine Summenfunktion dar, der durch Addition aufeinanderfolgender Wahrscheinlichkeiten bis zum R-ten Eintrag erhalten wird.

Die Wahrscheinlichkeit, daß die R-te Maschine infolge eines Schadens ausfällt unter der Bedingung, daß die vorherigen Ausfälle verschiedene Maschinen betraf, ergibt sich als Differenz der Funktionswerte für R und R-1 [2]. Es gilt somit:

$$P(R) = \left[\frac{1}{N^{(R-1)-1}} \times \frac{(N-1)!}{(N-(R-1))!}\right] - \left[\frac{1}{N^{R-1}} \times \frac{(N-1)!}{(N-R)!}\right]$$

$$= (N-1)! \left[\frac{1}{N^{R-2}(N-R+1)!} - \frac{1}{N^{R-1}(N-R)!}\right]$$

$$= \frac{(N-1)!(N-(N-R+1))}{N^{R-1}(N-R+1)!} = \frac{(N-1)!(R-1)}{(N-R+1)!N^{R-1}}$$

Diese Einzelheiten brauchte der Direktor nicht mehr, um überzeugt zu sein, daß frühe Wiederholungen nicht auf schlechte Wartung zurückgeführt werden können. Es war für ihn jedoch interessant zu erfahren, daß sein Problem eine Konkretisierung des bekannten Besetzungsproblems war (R Kugeln werden nacheinander zufällig auf N Urnen verteilt, wie groß ist die Wahrscheinlichkeit, daß keine Urne 2 oder mehr Kugeln enthält?) oder des Geburtstagsproblems [3]. Das Geburtstagsproblem lautet: Wie groß ist die Wahrscheinlichkeit, daß unter R Personen mindestens 2 am selben Tag Geburtstag haben? Hier stellt N - vorher die Zahl der Maschinen - die Zahl der Tage dar (N = 365, wobei Schaltjahre nicht berücksichtigt werden). Wie groß muß R sein, damit diese Wahrscheinlichkeit größer als 1/2 wird? Überraschenderweise ergibt die Rechnung, daß bereits für R = 23 Personen die Wahrscheinlichkeit, daß 2 Personen am gleichen Tag Geburtstag haben, größer als 50 % ist. [4] Wir müssen dazu natürlich jeden der 365 Tage für einen Geburtstag als gleichwahrscheinlich betrachten, eine Voraussetzung, die vielleicht gewagter ist als die Annahme gleicher Reparaturrisiken bei gleichen Maschinentypen.

Zum Schluß muß gesagt werden, daß das hier vorgestellte Modell nicht beweist, daß die Wartungsarbeiten zufriedenstellend ausgeführt wurden. Andere Aspekte

dieses Problems, die vielleicht noch untersucht werden sollten, wären, welche Maschinen über einen größeren Zeitraum öfters gewartet werden müssen, ob die Ausfallraten von Schicht zu Schicht oder Woche zu Woche sich ändern, oder ob weniger Ausfälle ein Ergebnis einer besseren Wartung sein können.

Anmerkungen

1) Als Quantil der Ordnung p (Quantil p-ter Ordnung) einer Zufallsgröße X mit der Verteilungsfunktion F_X bezeichnet man jede Zahl Q, für die $F_X(Q) \leq p \leq F_X(Q+0)$ gilt. Für p=0,5 ergibt sich der Median.

2) Es sei X eine Zufallsveränderliche, deren mögliche Werte $x_1, x_2, \ldots, x_n$ mit den Wahrscheinlichkeiten $p(x_1), p(x_2), \ldots, p(x_n)$ eintreten. Dann heißt die Funktion $F_X: R \to [0;1]$ mit $F_X(x) = \sum_{x_s<x} p(x_s)$ Verteilungsfunktion der diskreten Zufallsvariablen X. Sie gibt die Wahrscheinlichkeit dafür an, daß die Zufallsvariable X Werte annimmt, die x nicht überschreiten

$$F_X(x) = \sum_{x_s<x} p(x_s) = \sum_{s=1}^{r} f(x_s) = P(X=x_1)+P(X=x_2)+\ldots+P(X=x_r)$$
$$= P(X=x_1 \vee X=x_2 \vee \ldots \vee X=x_r)$$

3) Das Geburtstagsproblem läßt sich genauer ausgeführt in den meisten Büchern über Wahrscheinlichkeitsrechnung finden. Es sei hier jedoch noch auf den Aufsatz hingewiesen von SPENCER, N.: Celebrating the birthday problem. In: The Mathematics Teacher v. 70(4), S.348-353 (April 1977).

4) Das Geburtstagsparadoxon wird z.B. aufgelöst bei ENGEL, A.: Wahrscheinlichkeitsrechnung und Statistik, Bd. 1, Stuttgart, Klett 1973, S. 50-51.

Statistiken über verschiedene Stile
oder
Es war keine stilvolle Heirat

von J.Swift, übersetzt von M.Nuske

Nachdem ich kürzlich in meiner Unterrichtsklasse im Fach Statistik die Stile einiger Autoren untersuchte, muß meine Klasse Eliza große Sympathien entgegengebracht haben. Sie mußten also nicht nur im Englischunterricht mit Wörtern und Ausdrücken umgehen, sondern hatten sich damit auch im Statistik-Unterricht zu befassen.
Das Experiment wurde dadurch stimuliert, daß wir mit Jack Hodgins einen erfolgreichen Autor in unserem Kollegium besaßen. Unser Ziel war es herauszufinden, ob wir alleine unter dem Einsatz mathematischer Methoden Jack's Stil von dem anderer Autoren unterscheiden konnten. In "Probability und Statistics" behandelt John Durran Statistiken, die sich mit der Satzlänge und den unterschiedlich oft gebrauchten (=Variabilität) Substantiven befassen. Für die erste Untersuchung wurden deswegen diese beiden Gebiete gewählt.

Eine Statistik über den Gebrauch von Substantiven

Eine der Übungen in "Probability and Statistics" baut auf Material von G.U.Yule's "Statistical Study of Literary vocabulary (CUP 1944) auf. In der Aufgabe wird eine Statistik entwickelt, die den Gebrauch von Substantiven "mißt". Dies geschieht folgendermaßen:

a) Durch zufällige Auswahl wird ein Abschnitt herausgesucht.
b) Für den ausgewählten Abschnitt werden nun die darin erscheinenden Substantive aufgeschrieben.
c) Unter Berücksichtigung der Wiederholungen eines Substantives sammelt man so die ersten 100 verschiedenen Substantive. Dabei wird man viele Substantive haben, die nur einmal vorkommen; einige werden zweimal oder dreimal und einige noch öfters in diesem ausgezählten Teil enthalten sein.
d) Das Ergebnis wird nun in einer Häufigkeitstabelle angeordnet, wie dies für einen Abschnitt aus Hemingways Buch "For Whom the Bell Tolls" (dt.: Wem die Stunde schlägt) gezeigt wird: