ZUR BERECHNUNG DER STANDARDABWEICHUNG

I. D. HILL

Übersetzt von Bernd Wollring

Viele Rechner bestimmen die Standardabweichung auf Knopfdruck. Sind sie genau genug?

Das Problem

Der Hauptteil bei der Berechnung einer Standardabweichung besteht darin, die Summe der Quadrate aller Abweichungen vom Mittelwert, $\Sigma(x-\bar{x})^2$, zu bestimmen. Es zeigt sich, daß die Formel in dieser Gestalt zum Ausrechnen "von Hand" unbequem ist, da $\bar{x}$ im allgemeinen keinen einfachen Wert annimmt. Als einfaches Beispiel betrachten wir eine Stichprobe aus lediglich drei Werten: 0 , 5 und 8 . Der Mittelwert ist 4,333... , wendet man die Formel an, so muß man 4,333... , 0,666... und 3,666... quadrieren, und das ist wegen der periodischen Dezimalbrüche unbequem. Infolgedessen hat man Generationen von Statistik-Studenten beigebracht, lieber die algebraisch äquivalente Formel $\Sigma x^2-(\Sigma x)^2/n$ zu benutzen. In unserem Beispiel ergibt sie:

$$0^2 + 5^2 + 8^2 - 13^2/3,$$

und das ist wesentlich einfacher auszuführen, da nur an einer Stelle ein periodischer Dezimalbruch auftritt.

Als man noch "von Hand" ausrechnete, war dies ein guter Rat; denn er vereinfachte die Dinge. In den Fällen, da die Formel zu Ungenauigkeiten führte, konnte man von jemandem, der "von Hand" rechnete, erwarten, daß er dies bemerkte und andere Zahlen verwendete, oder daß er die Daten geeignet umskalieren würde. In der heutigen Zeit des elektronischen Ausrechnens ist es dagegen ein schlechter Rat. Diese Tatsache sollte heutzutage

Bestandteil jeder elementaren Statistik sein, aber man sieht wohl deutlich genug, daß das nicht so ist, trotz der Tatsache, daß man seit mindestens 16 Jahren davon weiß (siehe Welford, 1962). Man findet nicht nur Computerprogramme, die die Anweisung $\Sigma x^2-(\Sigma x)^2/n$ benutzen, und Lehrbücher, die nur diese erwähnen, sondern man findet inzwischen auch Taschenrechner auf dem Markt, die sie unter ihren fest verdrahteten Funktionen aufweisen.

Warum sollte das gefährlich sein, wo doch die beiden Formeln algebraisch völlig äquivalent sind? Die Antwort liegt in der Tatsache, daß die algebraische Äquivalenz von unbegrenzter Präzision in allen berechneten Werten ausgeht, aber bei praktischen Rechnungen nur eine begrenzte Anzahl geltender Ziffern zur Verfügung steht.

Probleme verursachen diejenigen Fälle, bei denen die Schwankung im Vergleich zum Mittelwert klein ist. Um zu sehen, was dann passiert, betrachten wir die Stichprobe aus den drei Zahlen 1000 , 1005 und 1008.

$$\Sigma x^2 = 3\cdot026\cdot089$$

$$(\Sigma x)^2/n = 3\cdot026\cdot056{,}333...$$

Subtrahieren liefert den korrekten Wert 32,666.... Nehmen wir aber an, wir benutzen einen Rechner, der nur mit 7 geltenden Ziffern arbeitet. Der erste Wert wird als $3\cdot026\cdot089$ bestimmt, aber der zweite wird auf $3\cdot026\cdot056$ gerundet. Folglich ergibt die Subtraktion 33, und obwohl die Maschine mit 7 geltenden Ziffern arbeitet, ist dieses Ergebnis nur in zwei geltenden Ziffern korrekt. Fünf Ziffern sind bei der Subtraktion weggeschnitten worden.

Wer einen Rechner zur Verfügung hat, dessen Repertoire die Standardabweichung umfaßt, sollte einmal versuchen, damit die Standardabweichung folgender Stichproben zu bestimmen: 1) 10 , 15 und 18 , 2) 100 , 105 und 108 und so weiter, also 10^n , $10^n + 5$ und $10^n + 8$ für wachsende Werte von n.

Mit einem Hewlett-Packard HP-67 erhält man:

n	berechnete Standardabweichung
1	4,041 451 886
2	4,041 452 091
3	4,041 472 504
4	4,043 513 324
5	3,872 983 346
6	0,000 000 000

Dabei ist das korrekte Resultat in jedem Fall 4, 041 451 884 . Ich möchte betonen, daß ich keinesfalls Hewlett-Packard im besonderen kritisieren möchte. Die Resultate dieser Maschine zitiere ich, da sie die einzige ist, die mir zur Verfügung steht.
Ich würde mich darüber freuen, von einem Hersteller zu erfahren, dessen Maschine nicht auf diese Art versagt.

Natürlich will ich nicht unterstellen, daß eine Stichprobe wie 100·000, 100·005 und 100·008 in der Praxis häufig auftritt. Sie wurde hier nur als einfaches Beispiel verwendet, um zu prüfen, ob ein Computerprogramm oder ein Rechner sichere oder gefährliche Methoden benutzt.

Man kann davon ausgehen, daß es keinem sachverständigen Statistiker je einfiele, die Standardabweichung von 100·000, 100·005 und 100·008 zu bestimmen, ohne die Daten zunächst passend umzuformen; das stimmt wohl. Aber diese Maschinen sind nicht nur für sachverständige Statistiker gedacht (und werden sicher nicht nur von solchen benutzt). Es sollte Aufgabe der Maschine sein, Methoden zu benutzen, die nicht einem solchen Irrtum anheimfallen, egal, welche Daten ein unwissender Benutzer eingibt. Insbesondere kann - bei fortschreitender Automatisierung - der unwissende Eingeber der Daten selbst wieder eine Maschine sein, so daß kein Mensch je die Zahlen sieht, die in die Rechnung eingehen. Das gilt besonders für ein Computerprogramm, das Zahlen verwendet, die innerhalb des Programms selbst ermittelt wurden.

Auch ein sachverständiger Statistiker kann nicht ständig so aufmerksam sein, daß er einhakt, wenn Daten wie 134,230 , 135,235 und 135,238 (, um mal eine andere Verkleidung unseres 0 , 5 , 8-Beispiels zu nehmen,) innerhalb einer langen Rechnung

auftreten. Mit einem HP-67 (, der mit 10 geltenden Ziffern arbeitet, mehr, als die meisten Taschenrechner haben,) würde er als Standardabweichung 0,004 472 136 erhalten, mehr als 10% abweichend vom korrekten Wert 0,004 041 452.

Die Lösung

Es gibt mindestens drei Wege, um diese Schwierigkeiten zu vermeiden. Der erste besteht darin, die ursprüngliche Formel zu benutzen. Er hat den Nachteil, daß man die Daten zweimal eingeben muß: zunächst, um den Mittelwert zu finden, dann wieder, um die Quadratsumme zu bestimmen. Das wäre nicht nur lästig, sondern auch eine Fehlerquelle beim Rechnen. Es würde bei einem Computer zudem untragbar sein, die Daten [von außen] zweimal in die Maschine einzulesen, aber wenn die Datenmenge nicht zu groß ist, um sie vorübergehend ganz zu speichern, gibt es keine Schwierigkeit, sie zweimal abzurufen. Darüber hinaus hat die Formel den Vorteil, daß man sie leicht behalten kann.

Die zweite Methode besteht darin, die "gefährliche" Formel zu benutzen, aber nicht mit den ursprünglichen Daten, sondern, nachdem man den ersten Wert von allen folgenden abgezogen hat. Man erhält den Mittelwert von x_1; x_2; x_3; ... wie vorher, aber zum Bestimmen der Standardabweichung verwendet man die Daten 0; x_2-x_1; x_3-x_1; .... Damit umgeht man fast alle Schwierigkeiten, und diese Prozedur wäre leicht als fest verdrahtete Operation in einen Taschenrechner einzubauen. (Allerdings ist es Unsinn, die Tatsache, daß der erste Wert bei dieser Methode stets Null ist, als "hübsches heuristisches Argument" für die Wahl von n-1 als passenden Teiler zu benutzen.)

Die dritte Methode ist um einiges komplizierter, und ihre Einzelheiten sind schwerer zu behalten, dennoch ist sie eine gute Methode für automatische Berechnungen. Anstelle der drei Speicher für n, Σx und Σx^2 werden in der Maschine drei Speicher für n, $\bar{x}$ und $\Sigma(x-\bar{x})^2$ benutzt. Wir bezeichnen den Speicherplatz für $\Sigma(x-\bar{x})^2$ kurz mit S. Bei der ersten Beobachtung initialisieren wir n mit 1, $\bar{x}$ mit dem beobachteten Wert und S mit 0. Bei jeder neuen Beobachtung x treffen wir dann folgende Zuweisungen:

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

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-