EINE KURZFORMEL ZUR BERECHNUNG DER MITTLEREN ABSOLUTEN ABWEICHUNG

von Nicholas Farnum, California State University, Fullerton, USA.

Originaltitel in "Teaching Statistics" Vol.10 (1988) Nr.3: A Short-cut Formula for the Mean Absolute Deviation
Übertragung: Bernd Wollring, Universität Münster

Einführung

Fast jeder moderne Text zur Statistik stellt im Anschluß an die Stichproben-Varianz

$$s^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$$

die folgende "Kurzformel" vor, die die Berechnung in nur einem Arbeitsgang erlaubt:

$$s^2 = \frac{1}{n-1}\left[\sum_1^n x_i^2 - \frac{1}{n}\left(\sum_1^n x_i\right)^2\right] \qquad (1)$$

Sie reduziert die Anzahl der zur Berechnung von s^2 erforderlichen arithmetischen Operationen erheblich. Dagegen erwähnen viele Texte nicht die mittlere absolute Abweichung (mean absolute deviation)

$$\mathrm{MAD} = \frac{1}{n}\sum_1^n |x_i - \bar{x}| \qquad (2)$$

oder nennen sie nur im Rahmen der begrifflichen Entwicklung der Stichprobenvarianz, in keinem Fall wird jedoch eine Kurzformel zur Berechnung angegeben. (Manche Autoren notieren MD anstelle von MAD.)

In diesem Aufsatz stellen wir eine anscheinend wenig bekannte Kurzformel für die mittlere absolute Abweichung vor

und verallgemeinern sie für Summen von Abweichungen von einem beliebigen Wert c. Wir geben Beispiele aus den Bereichen der Vorhersagen und der Bestandskontrolle, zwei Bereichen, in denen die mittlere absolute Abweichung der Varianz oft vorgezogen wird.

Die Kurzformel

Die mittlere absolute Abweichung einer Serie x_i von einer beliebigen Konstanten c kann wie folgt berechnet werden:

$$\frac{1}{n}\sum_1^n |x_i - c| = \frac{1}{n}\left[\sum_{x_i>c}(x_i - c) + \sum_{x_i \le c}(c - x_i) + \sum_{x_i>c}(c - x_i) - \sum_{x_i>c}(c - x_i)\right]$$

$$= \frac{2}{n}\sum_{x_i>c}(x_i - c) + \frac{1}{n}\sum_1^n (c - x_i)$$

$$= \frac{2}{n}\left(\sum_{x_i>c} x_i - n_1 c\right) + (c - \bar{x}) \qquad (3)$$

$\sum_{x_i>c} x_i$ ist die Summe und n_1 die Anzahl derjenigen x_i, die größer sind als c.

Für den Spezialfall c = $\bar{x}$ erhalten wir aus (3):

$$\mathrm{MAD} = \frac{2}{n}\left(\sum_{x_i>\bar{x}} x_i - n_1\bar{x}\right) \qquad (4)$$

Man findet sie bei SACHS (1982; S. 252). Ansonsten scheint sie weitgehend unbekannt zu sein. Ein anderer interssanter Fall liegt vor, wenn als c der Median m der Daten gewählt wird:

$$\frac{1}{n}\sum_1^n |x_i - m| = \frac{2}{n}\left(\sum_{x_i>m} x_i\right) - \bar{x} + \frac{m}{n}(n - 2n_1) \qquad (5)$$

Es ist allgemein bekannt (siehe etwa ABEL, 1985) daß der Ausdruck

$$\frac{1}{n}\sum_{1}^{n}|x_i - c|,$$

an der Stelle c = m minimal wird. Wir halten zudem fest, daß bei praktischen Anwendungen n_1 in der Regel nahe bei n/2 liegt, d.h. etwa die Hälfte der beobachteten Werte x_i liegt oberhalb des Medians. Damit erhalten wir für (5) folgende brauchbare Approximation:

$$\frac{1}{n}\sum_{1}^{n}|x_i - m| \approx \frac{2}{n}\left(\sum_{x_i > m} x_i\right) - \bar{x}. \qquad (6)$$

Diese Approximation ist dann besonders gut, wenn n groß ist und die Anzahl mehrfach auftretender Werte klein.

Beispiele

Man könnte Formel (4) einen "Eineinhalb-Schleifen"-Algorithmus nennen. Einen Durchlauf durch die Daten x_i benötigt man zur Berechnung von $\bar{x}$ und einen zweiten "teilweisen" Durchlauf, um diejenigen x_i zu bestimmen, die größer als $\bar{x}$ sind. In der Praxis kann man die Effektivität von (4) gut erkennen. Bei Vorliegen von Daten mit "Schiefe" oder Ausreißern, also genau da, wo man die mittlere absolute Abweichung der Varianz vorzieht, arbeitet diese "Kurzformel" sogar besser. Betrachten wir etwa die Datenserie (1;2;3;4;5;12). Der Mittelwert 4.5 wird von nur zwei Werten überschritten und wir erhalten aus (4):

$$\text{MAD} = \frac{2}{6}[(5 + 12) - 2(4{\cdot}5)] = 2{\cdot}67$$

Wir halten noch fest, daß man die Formeln (3), (4) und (5) ebensogut mit Hilfe der x_i hätte entwickeln können, für die $x_i < c$ gilt, so daß auch das Umgehen mit Daten, die "in der anderen Richtung schief" liegen, keine Schwierigkeiten bereitet.

Wie EHRENBERG (1983) dargestellt hat, wird die mittlere absolute Abweichung in der Praxis weit häufiger benutzt, als ihre Darstellung in heutigen Abhandlungen vermuten läßt. Bei der Kontrolle von Lagerbeständen zum Beispiel werden Vielfache der mittleren absoluten Abweichung regelmäßig benutzt, um die Warenqualität zu garantieren. Die kleinen Stichprobenumfänge und mögliche Ausreißer zusammen mit großen Lagerbeständen machen die mittlere absolute Abweichung zu einem besseren Maß als die Varianz (vgl. TUKEY, 1960, S. 448-485).

Eine weitere Anwendung: Bei der Auswertung von Vorhersage-Modellen benötigt man oft die Größe des mittleren Vorhersagefehlers

$$\frac{1}{n}\sum_{1}^{n}|e_i|,$$

wobei e_i der Fehler (auch negativ vorhergesagt) ist, den das Modell in der iten Periode macht. Aus (3) erhält man:

$$\frac{1}{n}\sum_{1}^{n}|e_i| = \frac{1}{n}\sum_{1}^{n}|e_i - 0| = \frac{2}{n}\left(\sum_{e_i > 0} e_i\right) - \bar{e}.$$

Der mittlere Fehler und die Summe der positiven Fehler (der "zu kleinen" Vorhersagen) sind alles, was man dazu benötigt.

LITERATUR

(1) ARBEL, T. (1985). Minimizing the Sum of Absolute Deviations, Teaching Statistics, 7, 88-89.

(2) EHRENBERG, A.S.C. (1983). We must Preach What is Practiced, The American Statistician, 37, 3, 248-250.

(3) SACHS, L. (1982). Applied Statistics, Springer-Verlag, New York.

(4) TUKEY, J.W. (1960). Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Stanford.