SCHÄTZEN DES MEDIANS BEI GRUPPIERTEN DATEN

von Ian Cook, University of Essex, England
Originaltitel in "Teaching Statistics" Vol.9 (1987) Nr.1:
Estimation of the Median from Grouped Data
Übertragung: Bernd Wollring, Universität Münster

Sind Daten nicht gruppiert, i.e. in Klassen eingeteilt, so ist der Median die "Beobachtung in der Mitte", wenn die Zahl n der Beobachtungen ungerade ist und konventionsgemäß der Mittelwert der beiden zentralen Beobachtungen, wenn n gerade ist, die berühmte "(n/2 + 1/2)te Beobachtung".

Einfache Formeln, die auf diesem Ansatz beruhen, werden oft auch bei gruppierten Daten verwendet, aber die Betrachtung zweier einfacher Fälle wird zeigen, daß der Fall nicht ganz klar liegt.

Angenommen, man hat vier beobachtete Werte einer stetigen Zufallsgröße, und die Häufigkeitsverteilung nach dem Gruppieren ist folgende:

Klasse	$5 < x \leqslant 15$	$15 < x \leqslant 25$	$25 < x \leqslant 35$
Häufigkeit	1	2	1

Die naive Auffassung besagt, der Median sei die "2.5te Beobachtung", womit folgendes gemeint ist:

$$15 + \frac{1\frac{1}{2}}{2} \times (25 - 15) = 15 + \frac{3}{4} \times 10 = 22{\cdot}5$$

Unglücklicherweise erhalten wir, wenn wir von rechts her statt von links her arbeiten, als "2.5te Beobachtung":

$$25 - \frac{1\frac{1}{2}}{2} \times (25 - 15) = 25 - \frac{3}{4} \times 10 = 17{\cdot}5$$

Hier stimmt etwas nicht.

Gehen wir davon aus, daß die beiden Beobachtungen in der Klasse des Medians genausogut irgendwo in dem Intervall

$15 < x \leqslant 25$ liegen könnten, so wäre es vernünftig, als Wert des Medians 20 zu wählen. Wir könnten ebensogut die Beobachtungen in der Klasse des Medians möglichst ausgeglichen verteilen, würden sie demnach als 17.5 und 22.5 annehmen mit einem Abstand von 5 untereinander und von 2.5 zum jweiligen nächsten Rand. Die Mitte dieser beiden Werte ergibt den Median 20.

Nun nehmen wir folgende Häufigkeitstafel an:

Klasse	$5 < x \leqslant 15$	$15 < x \leqslant 25$	$25 < x \leqslant 35$
Häufigkeit	1	3	1

Dann ergibt die einfache Formel, "von links" gerechnet:

$$15 + \tfrac{2}{3} \times 10 = 21{\cdot}67$$

"Von rechts" geerechnet, erhält man:

$$25 - \tfrac{2}{3} \times 10 = 18{\cdot}33$$

Es ist klar, der Median sollte wieder 20 sein. Gleichmäßiges Verteilen der Werte ergibt 16.67, 20.00 und 23.33, was wieder den vernünftigen Wert 20 für den Median ergibt.

Im allgemeinen Fall sei n die gesamte Zahl der beobachteten Werte einer stetigen Zufallsgröße. Wir betrachten zunächst den Fall, daß n ungerade ist. Dann ist der Median der ((n+1)/2)te beobachtete Wert. Die Median-Klasse $a < x \leqslant b$ liegt so, daß wenn die Summenhäufigkeiten für $a < x$ und $x \leqslant b$ die Werte n_a bzw. n_b haben, gilt:

$$n_a < \tfrac{1}{2}(n+1) \leq n_b$$

Nehmen wir in der Median-Klasse eine Gleichverteilung an, und verteilen die $(n_b - n_a)$ Beobachtungen gleichmäßig mit der paarweisen Entfernung d, wobei

Stochastik in der Schule 9 (1989), Heft 2

$$d = \frac{b-a}{(n_b - n_a)}$$

und die erste Beobachtung a + d/2 und die letzte b - d/2 ist, so kann der Median geschätzt werden als:

$$a - \tfrac{1}{2}d + (\tfrac{1}{2}n + \tfrac{1}{2} - n_a)d = a + (\tfrac{1}{2}n - n_a)d$$

Nun betrachten wir den Fall, daß n gerade ist und nehmen als Median die Mitte zwischen der (n/2)ten und der ((n/2)+1)ten Beobachtung. Fallen diese beiden in dieselbe Klasse $a < x \leqslant b$ mit

$$n_a < \tfrac{1}{2}n < \tfrac{1}{2}n + 1 \leq n_b,$$

so kann man den Median wie folgt schätzen:

$$\tfrac{1}{2}\{a - \tfrac{1}{2}d + (\tfrac{1}{2}n - n_a)d + a - \tfrac{1}{2}d + (\tfrac{1}{2}n + 1 - n_a)d\} = a + (\tfrac{1}{2}n - n_a)d$$

Wir halten fest, daß dieselbe Formel (ohne den Zusatz "+1/2") sowohl bei geradem als auch bei ungeradem n anwendbar ist.

Zur Erläuterung zwei Beispiele:

Beispiel 1 : Gegeben sind 50 Beobachtungen mit folgender Verteilung

Klasse	$0 < x \leqslant 5$	$5 < x \leqslant 10$	$10 < x \leqslant 20$
Häufigkeit	10	25	15

Der geschätzte Median ist:

$$5 + \frac{25-10}{25} \times 5 = 8{\cdot}0$$

Beispiel 2 : Gegeben sind 51 Beobachtungen mit der Verteilung:

Klasse	$0 < x \leqslant 5$	$5 < x \leqslant 10$	$10 < x \leqslant 20$
Häufigkeit	11	25	15

Der geschätzte Median ist:

$$5 + \frac{25{\cdot}5 - 11}{25} \times 5 = 7{\cdot}9$$

Fallen im allgemeinen Fall bei geraden n die (n/2)te und die ((n/2)+1)te Beobachtung in verschiedene Klassen $a < x \leqslant b$ und $c < x \leqslant d$, die nicht notwendig benachbart sind, und wobei gilt:

$$n_a < n \leq n_b = n_c < n + 1 \leq n_d,$$

so kann als Median die Mitte der Werte

$$b - \frac{b-a}{2(n_b - n_a)} \qquad c + \frac{d-c}{2(n_d - n_c)}$$

genommen werden. Im symmetrischen Fall mit b - a = d - c und $n_b - n_a = n_d - n_c$ ergibt dies 0.5 (b+c) für den Median.

Zum Schluß drei Bemerkungen:

Erstens: Gibt es irgendeinen Grund zu der Annahme, daß die Ausgangsverteilung in der Median-Klasse schief ist, so ist die gleichmäßige Verteilung der Werte keine vernünftige Basis für die Schätzung, und das Verfahren ist zu revidieren.

Zweitens: Ist die Zufallsgröße diskret, so kann man die Klassen-Intervalle als disjunkt und von der Form $a_i < x \leqslant b_i$ ansehen und die Methode wie oben anwenden.

Drittens mag man fragen, ob der hier angegebene Wert zu kompliziert zu bestimmen ist. Nun, das Bestimmen der "((n+1)/2)ten Beobachtung" ist auch kompliziert, und wie oft benötigen wir einen derart genauen Median? Aber wenn schon kompliziert, dann auch korrekt. Das Schöne an diesem Verfahren ist, daß dieselbe Formel für gerades wie für ungerades n arbeitet, und daß die Analyse des Problems uns zu der Frage führt, inwieweit es vernünftig ist, die gruppierten Daten zur Darstellung der Rohdaten zu wählen.

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:

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

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

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

$\sum_{x_i>c} x_i$ 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.