

ZUR WAHL PASSENDER PARAMETER
IN DER BESCHREIBENDEN STATISTIK

nach R. Beyth-Marom, Everyman's University,
Ramat Aviv, Israel

Originaltitel in 'Teaching Statistics' Vol. 5 (1983),

Nr. 3: Choosing an Appropriate Summary Statistic

Übersetzung: B. Wollring

Texte zur Einführung in die Statistik für College-Studenten beginnen meist mit beschreibender Statistik, wobei diverse Parameter für die "zentrale Tendenz" (Lageparameter) und für die Streuung dargestellt werden. Obwohl oft mehrere Parameter alternativ vorgestellt werden, wird der Student nur selten systematisch dazu angeleitet, zwischen diesen zu wählen. Wo Auswahlkriterien genannt sind, sind sie unvollständig oder beziehen sich nur auf die Lageparameter (z. B. HOPKINS und GLASS, 1978; SPENCE, COTTON, UNDERWOOD und DUNCAN, 1976).

Im folgenden geben wir sechs Auswahlkriterien an. Sie werden zunächst umgangssprachlich eingeführt und dann zur Auswertung diverser üblicher Lage- und Streuungsmaße verwendet (siehe Tabelle 1 und 2). Dieselben Kriterien kann man auch zur Auswahl passender Korrelationskoeffizienten benutzen.

1. In welchem Maßstab liegen die beobachteten Daten vor?

Ogleich es technisch möglich ist, für jeden Satz numerischer Daten alle Parameter auszurechnen, hängt die Bedeutung des Ergebnisses von der zugrunde liegenden Art der beobachteten Werte ab. Die Datensätze unterscheiden sich voneinander, je nachdem, wieviel Typisches der Beobachtung sie wiedergeben, und legen so einen "Maßstab" der Beobachtung fest. Zum Beispiel können die Zahlen 0 und 1 als Ersatz für zwei mögliche Antworten ("Links" oder "Rechts") auf eine Frage gemeint sein, die mehreren Personen gestellt ist. In einer solchen "Nominalskala" dienen die Zahlen nur zum Identifizieren der Antwort. In die Zahlen kann man nichts hinsichtlich der "Reihenfolge" oder "Größe" der Antworten

hineinlegen. Bei einem solchen Maßstab ist nur die Häufigkeit von Bedeutung: Haben die meisten Leute "1" geantwortet, so meinen die meisten Leute "Rechts". Es gibt dagegen keinen Sinn, den Median oder das arithmetische Mittel zu berechnen. ("Welche mittlere Entscheidung fällen Sie?")

2. Welches ist die angemessene Verlustfunktion?

Die Wahl desjenigen Lageparameters, der die gegebenen Daten "am besten" repräsentiert, hängt auch von den "Kosten" (oder dem vorhergesagten Fehler) ab, die man auf sich nimmt, wenn man davon ausgeht, daß alle Beobachtungswerte ihm gleich sind (FALK, 1980). Verschiedene Verlustfunktionen erfordern verschiedene Definitionen dieser "Kosten". Jede kann durch die Wahl eines passenden Lageparameters (Mittelwertes) minimal gehalten werden. Das arithmetische Mittel minimiert zum Beispiel die Summe der quadratischen Abweichungen, der Median dagegen minimiert die Summe der absoluten Abweichungen. Einige Streuungsmaße sind sehr stark auf Verlustfunktionen bezogen; meist wählt man das Streuungsmaß passend zu der Verlustfunktion, an der man interessiert ist, und bestimmt dazu dann durch Minimieren den geeignetsten Lageparameter.

Kriterien	Modalwert	Median	arithm. Mittel
bei welchem Skalentyp und differenziertere anwendbar	Nominalskala und differenziertere	Ordinalskala und differenziertere	Intervallskala und differenziertere
minimiert welche Verlustfunktion	Zahl der Fehler	Summe der absoluten Abweichungen	Summe der quadratischen Abweichungen
wie beeinflußt von sehr seltenen Werten	gar nicht beeinflußt	leicht beeinflußt	stark beeinflußt
für welche Verteilung am passendsten	symmetrische und asymmetrische; unimodale	symmetrische und asymmetrische; unimodale und multimodale	symmetrische; unimodale und multimodale
wie leicht zu berechnen	leicht	aufwendiger	aufwendiger
wie gut weiter zu verwenden	wenig	einigermaßen	gut

Tabelle 1: Kriterien zur Wahl eines Lageparameters (Mittelwertes)

Kriterien	Anteil der Fehler	Spannweite	Quartil-abstand	Durchschnitt der absoluten Abw.	Durchschnitt der quadratischen Abw.
bei welchem Skalentyp anwendbar	Nominalskala und differenziertere	Intervallskala und differenziertere	Intervallskala und differenziertere	Intervallskala und differenziertere	Intervallskala und differenziertere
Verlustfunktion	Zahl der Fehler	./.	./.	Summe der absoluten Abw.	Summe der quadratischen Abw.
wie beeinflusst von sehr seltenen Werten	nicht beeinflusst	nur von diesen beeinflusst	wenig beeinflusst	von ihnen wie von allen anderen beeinflusst	von ihnen mehr als von allen anderen beeinflusst
für welche Verteilung am passendsten	symmetrisch und asymmetrisch; unimodal	symmetrisch und asymmetrisch; unimodal und multimodal	symmetrisch und asymmetrisch; unimodal und multimodal	symmetrisch und asymmetrisch; unimodal und multimodal	symmetrisch; unimodal und multimodal
wie leicht zu berechnen	leicht	leicht	aufwendig	aufwendig	aufwendig
wie gut weiter zu verwenden	wenig	wenig	einigemaßen	wenig	gut

Tabelle 2: Kriterien zur Wahl von Streuungsmaßen

3. Sollte das gewählte Maß von sehr seltenen Beobachtungen beeinflusst werden?

Die spezielle Verwendung des Maßes bestimmt darüber, ob eines gewählt werden sollte, das von extrem seltenen Beobachtungen mehr beeinflusst wird oder eines, für das dies weniger gilt. Das Durchschnittsgehalt der Angestellten einer Firma ist zum Beispiel ein geeigneter Mittelwert, wenn die Lohnkosten der Firma in erster Linie interessieren. Dagegen ist es nicht repräsentativ für das, was die meisten Angestellten verdienen, da es sehr stark von wenigen sehr hohen Gehältern beeinflusst wird. Die Varianz, der Durchschnitt der quadratischen Abweichungen, wird von "Ausreißern" (stark abweichenden seltenen Werten) wesentlich mehr beeinflusst als der Durchschnitt der absoluten Abweichungen. Sind andererseits große Abweichungen Ursache großer Kosten, so kann es sehr wohl sinnvoll sein, sie "überproportional" zu berücksichtigen.

4. Wie sind die Daten verteilt?

Die Form einer Verteilung kann die Größe der verschiedenen Parameter stark beeinflussen und damit das Ausmaß, mit dem diese Parameter die "Wirklichkeit" beschreiben. Die Anzahl extremer Beobachtungen und ihre Häufigkeit sind ein Indikator, mit dem man Verteilungen unterscheiden kann und die die Parameter verschieden beeinflussen. Andere unterscheidende Charakteristika sind Symmetrie bzw. Asymmetrie einer Verteilung und die Zahl der Spitzen. Ist eine Verteilung zum Beispiel gewiß nicht symmetrisch, so unterscheiden der Median und das arithmetische Mittel sich stark, da das arithmetische Mittel von "Ausreißern" sehr stark beeinflusst wird. Die Anzahl der Spitzen einer Verteilung beeinflusst jeden Parameter, der nur von den Häufigkeiten abhängt (wie den Modalwert und den Anteil der Fehler).

5. Will man eine schnelle, grobe Schätzung oder eine genauere?

Eine selbstverständliche, oft sehr praktische Überlegung betrifft die Zeit, die zum Berechnen der Parameter nötig

ist. Diese Zeit hängt stark von der Form ab, in der die Daten vorliegen: als Urliste oder bereits nach Häufigkeiten sortiert. Einige Parameter lassen sich viel leichter als andere berechnen; gewöhnlich ist der Preis dafür eine größere Auswertung der Daten. Aber als erste Näherung sind sie bisweilen ganz nützlich.

6. Welche Untersuchungen sind weiterhin geplant?

Die sorgfältigsten Untersuchungen von Daten stützen sich auf das arithmetische Mittel als "Parameter der zentralen Lage" und auf die Standardabweichung als Streuungsmaß. Wenn solche Analysen beabsichtigt sind, so gibt es für den Mittelwert und die Standardabweichung keinen Ersatz.

Hinsichtlich der Tabellen sind zwei Punkte noch erwähnenswert:

Zunächst sind die Kriterien voneinander abhängig, deshalb kann es sein, daß man seine Ziele nicht alle gleichzeitig erreicht. Man kann daher entweder einen einzelnen Parameter als Kompromiß wählen oder mehrere, so daß jeder etwas andere Bedingungen erfüllt.

Zum zweiten sind die in den Tabellen enthaltenen Hinweise nicht absolut gemeint, es gibt Umstände, unter denen ein bestimmter Parameter in bezug auf die vorliegenden Daten zu einem falschen Bild führt. Betrachten wir zum Beispiel eine zweipunktige Anteilsverteilung: Die Anteile 10 % und 90 % mögen mit den Häufigkeiten 50 und 40 auftreten. Der Median ist 10 %, das ist völlig irreführend. Diese Verteilung ist natürlich eine Ausnahme. Aber es ist ebenso wichtig, Ausnahmen zu erkennen wie Regeln aufzustellen. Bei zu vielen unüblichen Ausnahmen riskiert man allerdings, vor lauter Bäumen den Wald nicht mehr zu sehen. Stellt man die genannten Kriterien durch typische Beispiele dar, so kann man sich ein Gespür für die Rohdaten beim Berechnen der Parameter erhalten.

Diskutiert man die Kriterien mit Studenten, so muß man wenigstens ansatzweise über Skalentypen und Verlustfunktionen

sprechen (siehe dazu COOMBS, DAWES und TVERSKY, 1970; NOVICK und JACKSON, 1974). Da diese Begriffe nicht "einfach" sind, erscheint es mir wichtig, für das Verstehen und korrekte Verwenden jedweder Statistik, sie von Anfang an mit den Studenten zu besprechen. Wir befassen uns, wie ich meine, zu oft mit Zahlen ohne Gefühl für ihre empirische Bedeutung (HUFF, 1954). Das Darstellen von Skalentypen und ihre Einschätzung für die Wahl geeigneter Maße sollte uns dazu führen, die empirische Seite der Daten nicht aus den Augen zu verlieren.

Bemerkung

Dank gebührt Don Mac Gregor und Baruch Fischhoff für nützliche Hinweise.

Literatur

- COOMBS, C. H.; DAWES, R. M.; TVERSKY, A.: Mathematical psychology: An elementary introduction. - Englewood Cliffs, N. J.: Prentice Hall, 1970
- FALK, R.: Minimize your losses. - Teaching Statistics, Vol. 2 (1980), Nr. 3 (80 - 83)
- HOPKINS, K. D.; GLASS, G. V.: Basic statistics for behavioral sciences. - Englewood Cliffs N. J.: Prentice Hall, 1978
- HUFF, D.: How to lie with statistics. - New York: W. W. Norton, 1954
- MCCALL, R. B.: Fundamental statistics for psychology. - New York: Harcourt, Brace, Jovannovich, 1980
- NOVICK, M. R.; JACKSON, P. H.: Statistical methods for educational und psychological research. - New York: McGraw Hill, 1974
- SPENCE, J.T.; COTTON, J.W.; UNDERWOOD, B.J.; DUNCAN, C.P.: Elementary statistics. - Englewood Cliffs N. J.: Prentice Hall, 1976